Applied Psycholinguistics (2021), 1–30
doi:10.1017/S014271642100014X

CAMBRIDGE
UNIVERSITY PRESS

ORIGINAL ARTICLE

Production-based training benefits the comprehension and production of grammatical gender in L2 German

Valérie Keppenne[1]*, Elise W. M. Hopman[2] and Carrie N. Jackson[1]

[1]Department of Germanic and Slavic Languages and Literatures, The Pennsylvania State University, University Park, PA 16802, USA and [2]Department of Psychology, University of Wisconsin-Madison, Madison, WI 53706, USA
*Corresponding author. Email: vxk57@psu.edu

(Received 17 January 2020; revised 7 January 2021; accepted 15 March 2021)

Abstract

Ongoing debate exists regarding the role of production-based versus comprehension-based training for L2 learning. However, recent research suggests an advantage for production training due to benefits stemming from the opportunity to compare generated output with feedback and from the memory mechanisms associated with language production. Based on recent findings with an artificial language paradigm, we investigated the effects of production-based and comprehension-based training for learning grammatical gender among beginning L2 German learners. Participants received production-based or comprehension-based training on grammatical gender assignment and gender agreement between determiners, adjectives, and 15 German nouns, followed by four tasks targeting the comprehension and production of the target nouns and their corresponding gender marking on determiners and adjectives. Both groups were equally accurate in comprehending and producing the nouns. For tasks requiring knowledge of grammatical gender, the production-based group outperformed the comprehension-based group on both comprehension and production tests. These findings demonstrate the importance of language production for creating robust linguistic representations and have important implications for classroom instruction.

Keywords: production; comprehension; L2 learning; grammatical gender

An important question in instructed second language (L2) acquisition is which type of instruction is most beneficial for acquiring an L2. Many studies have compared the effectiveness of comprehension-based instruction (CBI) and production-based instruction (PBI) for learning L2 grammatical forms, especially morphosyntactic structures (see Shintani, 2015; Shintani et al., 2013 for two meta-analyses). While these analyses show an overall immediate advantage of CBI for receptive knowledge and an overall long-term advantage of PBI for productive knowledge, questions remain regarding the underlying cognitive mechanisms associated with language production, and how they may be particularly beneficial to the acquisition of L2 grammatical forms. A recent study (Hopman & MacDonald, 2018) compared the effectiveness of comprehension-based tasks versus production-based tasks for the learning of simple morphosyntactic agreement in an artificial language paradigm and showed clear advantages for production-based over comprehension-based training on posttest measures testing the comprehension of the target morphosyntactic agreement features. The authors attribute this finding to language production drawing on a different type of memory processing than language comprehension, thereby strengthening the relevant agreement features in memory. While these are intriguing results, previous research suggests that findings from artificial language studies do not always generalize to natural language learning (e.g., Paul & Grüter, 2016), underscoring the need for additional research. Furthermore, replicating these findings with a natural language may also shed light on why production-based activities may provide a learning advantage in the first place.

Based on Hopman and MacDonald's (2018) experimental design, the present study investigates whether production-based training is more beneficial than comprehension-based training for *comprehending* grammatical forms when learning a more complex morphosyntactic agreement paradigm in a natural language, namely grammatical gender agreement in L2 German. Additionally, since Hopman and MacDonald only tested comprehension performance at posttest, this study investigates whether PBI is more effective than CBI for *producing* grammatical gender agreement in L2 German, as well as comprehending it.

Defining CBI and PBI

As the terms CBI and PBI suggest, the main difference between the two lies in the types of learning activities used, that is, whether the learner is required to produce target L2 structures during training. Underlying these two contrasting methods are different assumptions regarding how to encourage learners to attend to and process new grammatical forms in a manner that facilitates acquisition. CBI, for instance, does not require learners to produce target forms, assuming that L2 acquisition is driven largely by input and how learners interact with L2 input during comprehension (e.g., Krashen, 1982; Truscott & Sharwood Smith, 2004; VanPatten, 2004, 2013). Critically, activities in CBI structure L2 input in a manner in which the learner must successfully process the target L2 form to comprehend its meaning (Ellis, 2012; VanPatten, 1996, 2002). By limiting the extent to which learners can rely on lexical items to correctly interpret L2 input, CBI attempts to focus the learner's attention on target forms and the meaning encoded by these forms.

PBI, on the other hand, encourages learners to produce target forms. As Swain proposes in her Output Hypothesis (Swain, 1995, 2005), production of L2 output is necessary for

> learners to move away from the semantic, open-ended, nondeterministic, strategic processing prevalent in comprehension to the complete grammatical

> processing needed for accurate production. Output, thus, would seem to have a potentially significant role in the development of syntax and morphology. (Swain, 1995, p. 128)

Additionally, learners can use their L2 output to test hypotheses about the target language: the learner produces an utterance, receives either positive or negative feedback from her interlocutor, and has an opportunity to modify her output to be more target-like, thereby updating the state of her language knowledge. As opposed to theories that see comprehension as crucial for L2 acquisition, Swain identifies language production as the locus of language acquisition. Nevertheless, PBI does not preclude opportunities for learners to comprehend the target forms (e.g., Hopman & MacDonald, 2018; Morgan-Short & Bowden, 2006; Soruç et al., 2017).

Many studies have compared the effectiveness of CBI and PBI for learning L2 grammatical forms. Alongside questions regarding how production-based versus comprehension-based activities contribute to L2 learners' developing linguistic systems, another key aspect in this debate is which methods lead to better comprehension and production of the target L2 forms. In some studies, CBI is more effective than PBI for comprehending the target linguistic features and leads to similar performance as PBI on production posttests (e.g., Soruç et al., 2017; VanPatten & Cadierno, 1993; VanPatten & Wong, 2004). Other studies show PBI to be more effective than CBI for producing the target linguistic features and show similar gains to CBI on comprehension posttests (e.g., Allen, 2000; Farley & Aslan, 2012; Morgan-Short & Bowden, 2006; Yamashita & Iizuka, 2017). Comparing the results of 35 experiments in a meta-analysis, Shintani et al. (2013) concluded that CBI is more beneficial than PBI for comprehension on immediate posttests, but that this advantage disappears on delayed posttests. Conversely, there are no differences between PBI and CBI on immediate posttests for production, but PBI is more effective than CBI on delayed posttests. Together, these findings indicate that training benefits may be modality specific, that is, comprehension or production, when measured immediately after training, but that PBI is more effective than CBI when measured over an extended period of time.

While many individual studies find an advantage of PBI over CBI, many studies find the opposite. How can these conflicting findings be reconciled? In a review of studies comparing CBI and PBI, DeKeyser and Botana (2015) point out that many studies that find an overall advantage of PBI included activities in which language production was communicative and meaningful to the same degree as the comprehension activities used in CBI. When CBI was found more beneficial overall, however, studies often only included production-based activities that were mechanical grammar drills or repetitive in nature. As outlined by the Output Hypothesis (Swain, 1995, 2005), however, only meaningful output can push the learner to acquire more advanced forms of the target language.

Benefits of production for language learning

To address differences in task demands between production and comprehension tasks in previous research on CBI and PBI, Hopman and MacDonald (2018) compared the effectiveness of comprehension- and production-based activities for the learning of two semantic and two number agreement features in a lab-based experiment in an artificial language, as shown in (1). The semantic suffixes *-us* and *-ok* described the appearance of monsters, akin to grammatical markers used to classify nouns according to grammatical gender in natural languages, and the suffixes *-usu* and *-oko* marked nouns as plural. The artificial language required that all determiners, adjectives, nouns, and verbs were marked with the same suffixes to agree in both semantic/grammatical gender and number with the subject noun phrase.

(1)	Vusu	Fumusu	Teepusu	Traw	Ot	Divusu	Kredel
	Determiner	Adjective	Noun	Markings	Preposition	Verb	Location
	The two kind yellow teeps with curved lines grow bigger at the location with the mountain.						

Participants were exposed to training blocks containing phrases and sentences in the artificial language together with corresponding pictures. After each block of passive exposure, participants completed an active task block, with either forced-choice comprehension activities or free production activities. After each trial in the active task, participants in both groups saw and heard the correct pairing of the target picture and the target phrase describing it, providing the opportunity to learn the correct pairing regardless of accuracy during that active trial. Crucially, the use of comprehension-based versus production-based activities was the only difference between the two groups; the amount and type of input received during the passive exposure phase, as well as the feedback participants received, was identical across both groups.

Immediately following training, participants completed forced-choice comprehension tasks and an error-monitoring task targeting their comprehension accuracy and speed. In the forced-choice task targeting their understanding of the agreement suffixes, participants saw two pictures on the screen which differed either in monster number or in monster type. At the same time, they heard a phrase and were instructed to identify the picture matching the phrase as quickly as possible. The error-monitoring task targeted participants' sensitivity to agreement violations. In this task, participants heard a sentence without seeing a picture and were instructed to judge whether the sentence was correct or contained an error. In erroneous sentences, one suffix did not match the other three in the sentence — for instance, a noun marked with a singular suffix in a sentence where all other lexical items requiring agreement were marked with a plural suffix. The results showed a significant advantage for the production-based group over the comprehension-based group both in terms of comprehension accuracy and response speed for all tasks targeting the agreement features.

Hopman and MacDonald (2018) attribute this finding to the idea that language comprehension and language production typically draw on different memory processes. Whereas free language production involves recalling material from memory, language comprehension only involves recognition. When learning foreign language vocabulary, training involving recall leads to higher accuracy in comprehension and production for the vocabulary than training relying on recognition (Kang et al., 2013; Karpicke & Bauernschmidt, 2011; Karpicke & Roediger, 2008), known as the “testing effect” in the memory literature. These findings suggest that training via the production of foreign vocabulary items improves retrieval and recognition mechanisms. Hopman and MacDonald hypothesized that free production, involving recall rather than recognition, might have benefits beyond the single word level and help learners acquire morphosyntactic features of a new language. Specifically, they note that producing a longer phrase requires utterance planning, and during utterance planning the to-be-produced sentence, as well as the message, is held in working memory, providing opportunity for the memory traces of the different elements of the utterance to bind (MacDonald, 2016). This should lead to better learning of not just the novel words but also their grammatical features and the grammatical dependencies between words.

Rather than invoking different memory processes in comprehension and production, recent models of language processing suggest that production and comprehension are tightly interwoven and that similar mechanisms are active during both production and comprehension processes (e.g., Dell & Chang, 2014; Pickering & Garrod, 2013). Dell and Chang’s (2014) P-chain model states that individuals make predictions about upcoming language input using a top-down process, which Dell and Chang label a production process. Prediction thus links comprehension and production processes within the individual. In this model, production processes are integral to learning and language adaptation, as produced output constitutes a prediction of what is possible in the language, while upcoming input serves as feedback related to that prediction. Similarly, Pickering and Garrod’s (2013) model suggests that individuals make use of forward prediction models to facilitate both language comprehension and language production, all the while drawing on representations that are separate between comprehension and production. Critically, both of these forward prediction models rely on processes related to production and therefore ascribe a critical role to production processes in both language comprehension and language production. For language learning, this would imply that training production-based prediction processes in L2 learners would positively impact both comprehension and production skills.

In fact, recent studies investigating the learning of new word meanings, either those of infrequent lexical items in English (Potts & Shanks, 2014) or of foreign language vocabulary (Potts et al., 2019), show a learning advantage for guessing translations of these items, that is, predicting the language form, compared to simply reading the items. This advantage of generating and guessing language forms was found on both subsequent recognition tests (Potts et al., 2019) and on subsequent production tests (Kang et al., 2013). These studies emphasize the role of feedback, as feedback provides learners an opportunity to evaluate their translation guesses against the actual, correct response. Rather than just benefiting from retrieval mechanisms during language production, Potts et al. (2019) argue that producing (incorrect) translation guesses to vocabulary words prior to seeing the correct answer creates a sense of curiosity in the learner, who then wants to fill the gap in her current state of language knowledge (for similar discussions, see also Schmidt 1990, 2001; Swain, 1995, 2005). Under this account, production-based training leads to better encoding of relevant language input as a consequence of initially making erroneous predictions, rather than necessarily by improving retrieval mechanisms *per se*. While generating and guessing language forms, in conjunction with feedback, has thus been shown to benefit word learning, there are reasons to believe that L2 production similarly benefits the learner's emerging L2 grammar. Based on evidence from priming studies with L2 learners, Hartsuiker and Bernolet (2017) suggest that learners' L2 grammar initially consists of explicit memory and item-specific knowledge of specific language forms. Only over time do learners develop increasingly abstract mental representations in which grammatical features are generalized across lexical items. These accounts offer an explanation for why the meaningful production-based activities in Hopman and McDonald (2018) showed a clear advantage over comprehension-based activities for the acquisition of morphosyntactic dependencies, as they identify ways in which language production plays an important role in the learning process.

However, three key issues remain. First, the error-monitoring task Hopman and MacDonald (2018) used to assess learning did not require learners to process the meaning encoded in the form, since it was a task based purely on judging the suffix patterns in auditory sentences without a depicted referent. The creation of appropriate and accurate form-meaning connections, however, is a necessary prerequisite for successful second language acquisition (e.g., VanPatten et al., 2004). Second, natural language learners tested in a lab or classroom setting typically already have some prior experience with learning the target language, and this might mitigate the effects of training. For example, an order-of-learning effect initially shown in an artificial language study (Arnon & Ramscar, 2012) replicated for classifier-noun associations only for learners without any prior experience with Chinese, but did not replicate for learners with several weeks of classroom exposure to the language (Paul & Grüter, 2016; but see Ettlinger et al., 2016, for counterevidence). Third, the agreement paradigm created for the artificial language in Hopman and MacDonald was rather simple when compared to agreement paradigms found in many natural languages, for instance, compared to grammatical gender agreement in German. Thus, it is critical to test whether the advantage for production training they found still holds when learning more complex agreement paradigms in a natural language.

Grammatical gender in natural language

In languages with grammatical gender, nouns are assigned to one of several grammatical gender classes, and other linguistic elements in a sentence, such as determiners, adjectives, or verbs, must agree in gender with the noun they modify. While some languages, like Spanish, have a rather transparent system of gender assignment, where a noun's grammatical gender is reliably identified based on morphophonological cues on the noun itself, gender assignment in other languages is largely arbitrary in nature. In German, for instance, all nouns belong to one of three gender classes, namely masculine (*der/ein Becher,* "the$_{\text{MASC}}$/a$_{\text{MASC}}$ cup"), feminine (*die/eine Tasche,* "the$_{\text{FEM}}$/a$_{\text{FEM}}$ bag"), or neuter (*das/ein Geschenk,* "the$_{\text{NEUT}}$/a$_{\text{NEUT}}$ gift"), but the morphophonological cues governing gender assignment are complex and probabilistic in nature (Köpcke & Zubin, 1983, 1984). Further, German requires that determiners and attributive adjectives are marked to agree with the noun they modify, and these agreement markers take different forms depending on the definiteness of the determiner and the gender of the noun they modify (e.g., *der blaue Becher* "the blue cup" but *ein blauer Becher* "a blue cup"; see also Table 1).

Table 1. Determiners and adjectives for singular nominative nouns in German

	Determiner	Adjective	Noun	
Masculine	der	blau*e*	Becher	"cup"
	ein	blau*er*		
Feminine	die	blau*e*	Tasche	"purse"
	eine	blau*e*		
Neuter	das	blau*e*	Geschenk	"present"
	ein	blau*es*		

Research shows that late-learning L2 learners can eventually acquire gender agreement paradigms and learn to correctly mark determiners and adjectives when the grammatical gender of a given noun is known. However, correct gender assignment – correctly identifying a specific noun's grammatical gender – remains difficult, even amongst highly proficient L2 speakers (e.g., Bordag et al., 2017; Grüter et al., 2012; Hopp, 2013, 2016, among others). The difficulty of grammatical gender assignment is exacerbated in languages like German, where there are few reliable morphophonological cues that L2 learners can use to identify the grammatical gender of a given noun. Thus, the complexity of grammatical gender assignment and agreement in German, as compared to the agreement paradigm used in Hopman and MacDonald (2018) and previous studies comparing the effectiveness of production-based versus comprehension-based practice in natural languages (e.g., Benati & Lee, 2008; De Jong, 2005), raises the question of whether the advantages found for production-based training in terms of comprehension accuracy (Hopman & MacDonald, 2018) translate to a natural language learning context in which the agreement paradigm is more complex, especially among beginning L2 learners with some previous exposure to the target language.

The present study

The present study investigated whether the findings from Hopman and MacDonald (2018) could be replicated in the context of natural language learning by targeting the learning of grammatical gender among beginning classroom L2 learners of German, and therefore adopted the training methods and comprehension tests used in Hopman and MacDonald. Additionally, the present study examined whether the advantages for production-based training over comprehension-based training for comprehension extend to language production. The present study thereby differs from previous investigations of CBI and PBI in both the design of the training phase and the tests used to measure comprehension accuracy. We posed the following research questions:

Q1. Are there differences between production-based training and comprehension-based training for *comprehending* grammatical gender in a natural language learning context, specifically in L2 German, where gender assignment is opaque and the gender agreement paradigm is complex?

Q2. Are there differences between production-based training and comprehension-based training for *producing* accurate grammatical gender marking in L2 German?

Based on Hopman and MacDonald's (2018) findings comparing the benefits of production-based and comprehension-based tasks for the learning of morphological markers, we expect the following:

H1. If production-based training is more beneficial than comprehension-based training in natural language learning, similar to what Hopman and MacDonald found for artificial language learning, then participants receiving production-based training will exhibit greater benefits for the *comprehension* of grammatical gender in L2 German than participants receiving comprehension-based training on all comprehension measures. These results would contrast with the findings of Shintani et al.'s (2013) meta-analysis, which found CBI to be more effective than PBI for receptive knowledge at immediate posttest.

H2. If production-based training is more effective than comprehension-based training in helping learners produce sentences with accurate grammatical gender assignment and agreement in L2 German, then participants receiving production-based training will exhibit greater benefits for the *production* of grammatical gender than participants receiving comprehension-based training. Again, these results would contrast with Shintani et al.'s meta-analysis, which found no difference between PBI and CBI for productive knowledge at immediate posttest.

Methodology

Participants

Fifty-one undergraduate students were recruited from eight sections of a first-semester German language class at a large public university in the United States. The 51 participants were randomly assigned to one of two experimental groups: 25 participants in the comprehension-based group (COMP) and 26 participants in the production-based group (PROD).[1] All recruitment and testing took place during weeks 4–5 of the semester. At that point in the semester, students had learned that German uses definite and indefinite gender-marked articles with nouns, but had not been introduced to adjectives and the relevant gender suffixes yet, nor had they encountered the 15 non-cognate target nouns included in the study. Participants received monetary compensation or course credit for their participation. Most participants were native speakers of English (32) or Chinese (8), but other L1s included Vietnamese, Korean, Spanish, Russian, and Marathi.[2] Three participants were excluded because their data were lost due to a program malfunction and three other participants were excluded for not finishing the testing session. Of the remaining 45 participants, four more participants were excluded because they had taken German in middle school or high school or had spent time

Table 2. Age and self-rated proficiency by group

	COMP		PROD	
	M	*SD*	*M*	*SD*
Current age (years)	20.0	3.6	20.4	2.4
L2 proficiency ratings (out of 10)				
Reading	3.5	2.4	4.3	2.3
Spelling	2.7	1.9	3.5	2.4
Writing	2.5	1.8	3.5	2.3
Speaking	3.0	2.0	3.1	2.0
Listening	3.3	2.1	3.5	2.0
Overall	3.0	1.8	3.6	2.0

abroad in a German-speaking country, and thus had a different level of experience with German than the other participants. Data from the remaining 41 participants, all with maximally 4–5 weeks of German experience, are included in the analyses and results. Of these 41 participants, 24 were in COMP (13 males; nine females; two no answer) and 17 were in PROD (12 males; three females; two no answer).[3] Participants in both groups self-reported low overall proficiency in German on a 10-point Likert scale (see Table 2) with no significant difference between groups (all ps $> .15$).[4]

Materials and training

The target noun phrases included 15 singular German nouns, none of which were cognates with English and none of which had been previously introduced in the classroom at the time of the study. The decision to only include unknown lexical items was made to ensure that participants did not know the grammatical gender of the target nouns prior to training, given that language learners sometimes learn the grammatical gender hand in hand with new lexical items. This also allowed us to stay true to Hopman and MacDonald's (2018) training design. There were five masculine, five neuter, and five feminine nouns. All nouns were two syllables long and were concrete and imageable (see Appendix A for a full word list). A female native speaker of German recorded all training materials in a sound proof booth. Additionally, we created simple black and white, or colored, line drawings that illustrated all phrases and sentences about the imageable objects used in the experiment. All tasks and testing measures were piloted with L2 German learners from the same population. Based on their feedback, the training phase was shortened by one block.

Training included 10 blocks of *passive exposure* to the target materials. No explicit information about a noun's grammatical gender or gender agreement was provided at any point during the experiment. During the first passive exposure block, participants saw a picture paired with auditory and written input in the form of a noun phrase with a definite article that matched the picture (*der Becher* "the$_{\text{MASC}}$ cup"). Participants were told to pay attention to the input, but that no action was required (Figure 1a).

Each passive exposure block alternated with blocks of either comprehension-based (COMP) or production-based (PROD) active learning. In the first *comprehension-based active learning* block, participants saw a picture paired with auditory and written input in the form of a noun phrase with a definite article and had to indicate whether the auditory and written input matched the picture through a keypress, with J for "Yes" and F for "No" (Figure 1b). All pictures were identical to ones participants had just been exposed to during the immediately preceding passive exposure block. Half of the trials in each block were mismatches, meaning the correct response was "No." Mismatches targeted content words and not gender marking. Within each block, mismatches occurred with a balanced number of masculine, feminine, and neuter nouns. Immediately after each item, participants received feedback on their response accuracy (*Correct, Incorrect*). Regardless of their response accuracy, this feedback was followed by a repetition of the picture with its matching phrase in auditory and written mode.

In the first *production-based active learning* block, participants were prompted to orally describe a picture displayed onscreen in German using the vocabulary and structures they had been introduced to during the immediately preceding passive exposure phase (Figure 1c). A "d___ . . . " underneath the picture prompted them to use a definite article, an "e___ . . . " prompted the use of an indefinite article in later blocks. As in the active comprehension blocks, all pictures were identical to ones they had just been exposed to during the immediately preceding passive exposure block. Responses were recorded with tripod-mounted USB microphones. Participants pressed a key after describing the picture out loud and then saw the same picture accompanied with its matching phrase in auditory and written mode. This repetition phase occurred regardless of whether they had accurately described the target picture.

The second passive exposure and active learning blocks were similar to the first block, but this time introduced the same 15 pictures and noun phrases with an indefinite article (*ein Becher* "a$_{\text{MASC}}$ cup"). Within the first two training blocks, participants were thus introduced to 15 new nouns along with their associated grammatical gender marking on definite and indefinite articles.

With each additional block, more lexical material was introduced, gradually building up to complete sentences. In the third training block, participants learned four cognate color adjectives (e.g., *blau* "blue"), which were then combined with the target noun phrases in block four (e.g., *ein blauer Becher* "a$_{\text{MASC}}$ blue$_{\text{MASC}}$ cup"). Similarly, in the fifth training block the participants learned four non-cognate pattern adjectives (e.g., *gepunktet* "dotted"), which were then combined with the target noun phrases in block six (e.g., *der blaue gepunktete Becher* "the$_{\text{MASC}}$ blue$_{\text{MASC}}$ dotted$_{\text{MASC}}$ cup"). Starting with the fourth block, definite and indefinite articles were balanced across gender categories within each block. This also means that mismatches in the comprehension-based active learning blocks starting with the fourth block were balanced across both gender categories and definite versus indefinite articles. Blocks four through six thus introduced participants to the grammatical gender marking paradigm for adjectives.

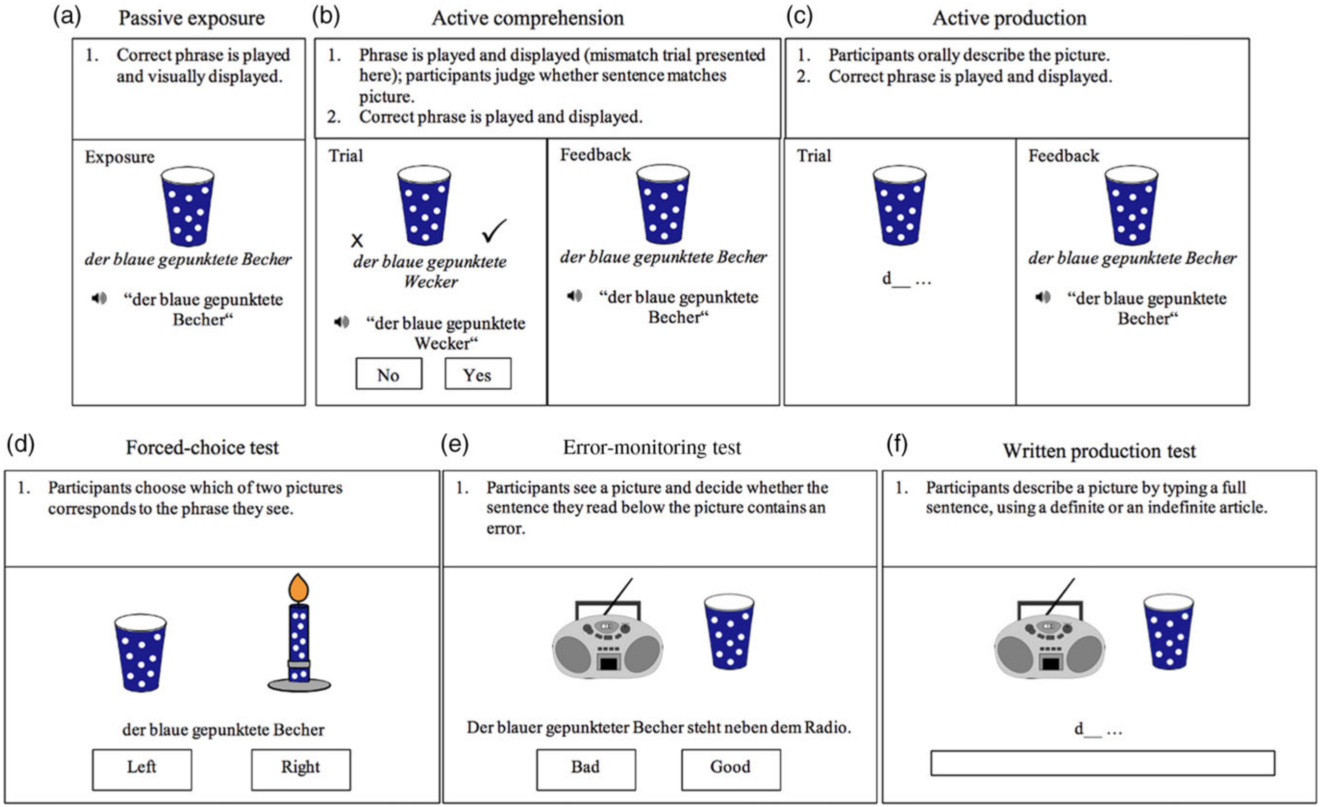

(a) Passive exposure
1. Correct phrase is played and visually displayed.
Exposure
der blaue gepunktete Becher
"der blaue gepunktete Becher"
(b) Active comprehension
1. Phrase is played and displayed (mismatch trial presented here); participants judge whether sentence matches picture.
2. Correct phrase is played and displayed.
Trial
x
✓
der blaue gepunktete Wecker
"der blaue gepunktete Wecker"
No
Yes
Feedback
der blaue gepunktete Becher
"der blaue gepunktete Becher"
(c) Active production
1. Participants orally describe the picture.
2. Correct phrase is played and displayed.
Trial
d__ …
Feedback
der blaue gepunktete Becher
"der blaue gepunktete Becher"
(d) Forced-choice test
1. Participants choose which of two pictures corresponds to the phrase they see.
der blaue gepunktete Becher
Left
Right
(e) Error-monitoring test
1. Participants see a picture and decide whether the sentence they read below the picture contains an error.
Der blauer gepunkteter Becher steht neben dem Radio.
Bad
Good
(f) Written production test
1. Participants describe a picture by typing a full sentence, using a definite or an indefinite article.
d__ …

Figure 1. Visualization of the Training Tasks (a–c) and Testing Measures (d–f).

In block seven, participants encountered noun phrases with two attributive adjectives, similar to those in block six (e.g., *ein blauer gepunkteter*… "a$_{\text{MASC}}$ blue$_{\text{MASC}}$ dotted$_{\text{MASC}}$ one"), but the target nouns were left out of the noun phrase, creating elliptical nominal phrases. Such phrases are grammatical in German (Günther, 2013), although infrequent. Doing so focused participants' attention on the gender marking on the article and the adjectives, as this was the only information identifying the correct target noun. Doing so emphasized the usefulness of gender marking for identifying real-world objects in German. Importantly, the phrase was still accompanied by a picture that depicted the relevant target noun, so that participants could make the form-meaning connection between the gender-marked items and the referent in the picture. In the comprehension-based active learning block following passive exposure where the target noun was omitted, mismatch trials always used incorrect gender markings on the noun phrase's article and the two attributive adjectives (e.g., *eine blaue gepunktete* "a$_{\text{FEM}}$ blue$_{\text{FEM}}$ dotted$_{\text{FEM}}$ …" to describe the masculine noun *Becher* "cup"). In this manner, knowledge of the grammatical gender associated with the noun was the only way to correctly judge whether the written and auditory phrase matched or mismatched the target picture. In the production-based active learning block following passive exposure where the target noun had been omitted, participants were encouraged to omit the target noun in their own productions.

In the eighth training block, participants encountered sentence frames that included one verb and alternated between three different locations described by a cognate (e.g., …*steht neben dem Radio/Bett/Sofa* "stands next to the radio/bed/sofa"). In block nine, the target noun phrases with two attributive adjectives were then embedded in the sentence frame, as shown in (2). In the final training block, participants were introduced to another set of full sentences, but this time the target nouns were omitted from the first noun phrase again, similar to block seven, in order to focus participants' attention on the gender markings (e.g., *Ein blauer gepunkteter … steht neben dem Radio* "A$_{\text{MASC}}$ blue$_{\text{MASC}}$ dotted$_{\text{MASC}}$ (one) stands next to the radio"). In total, participants completed 116 passive learning trials and 116 active learning trials across the 10 training blocks (see Appendix B). See the Supplementary Material for more details on the materials.

(2)	Ein	blauer	gepunkteter	Becher	steht	neben	dem	Radio.
	A$_{\text{MASC}}$	blue$_{\text{MASC}}$	dotted$_{\text{MASC}}$	cup$_{\text{MASC}}$	stands	next to	the	radio.
	A	blue	dotted	cup	stands	next to	the	radio.

Testing

Immediately after training, participants completed four testing measures. The first two were forced-choice tests (Figure 1d), one that targeted the comprehension of grammatical gender marking, and one that targeted knowledge of the lexical items themselves. The third test, an error-monitoring test (Figure 1e), targeted meta-linguistic knowledge of the grammatical gender agreement system. The fourth test targeted production skills through a written production test (Figure 1f). None of the test measures included any auditory input, and none of the color-pattern-noun pairings (and their corresponding pictures) in any of the test measures had been introduced in that particular combination during training, such that all color-pattern-noun pairings were new to participants. This was done to make sure participants could not successfully answer test questions simply by having memorized entire phrases during training. In so doing, we ensured that participants had to process the target vocabulary and grammatical forms to accurately answer test questions.

Within each test, the order of trials was randomized across participants. We measured accuracy and reaction times (RT) for the forced-choice and error-monitoring tests, and accuracy only for the production test.

Forced-choice tests

The forced-choice comprehension tests were similar in format to the active comprehension tasks in the training phase. Participants saw two pictures on the screen, read a phrase displayed underneath the pictures, and identified which picture matched the phrase by pressing F for the left picture and J for the right picture (Figure 1d). Participants completed 30 trials in which the target noun was omitted. Then they completed 30 trials that included the target noun. In blocks without the target noun, the foil item was always of a different gender than the target item but had the same color and pattern, thus ensuring that participants had to process the grammatical gender marking on the articles and adjectives to identify the correct target picture. In blocks that included the target noun, the foil item was always of the same gender as the target item and had the same color and pattern. Thus, participants had to know the target vocabulary words to identify the correct target picture. Within each block of 30 trials, half of the items contained a definite article and the other half contained an indefinite article. Target items and foil items were balanced in terms of location on the screen.

Error-monitoring test

In the error-monitoring test, participants saw a picture of a target noun in a specific location, read a sentence displayed underneath the picture, and identified whether the sentence contained an error by pressing J for a correct sentence and F for a sentence with an error (Figure 1e) . This test contained 105 items (see Table 3), with 35 grammatically correct sentences and 70 sentences that contained an error. Word order errors served as distractor items. None of the correct sentences or the incorrect sentences' correct alternative had been introduced during training, such that all color-pattern-noun pairings were new to participants.

Written production test

In the written production test, participants saw a picture of a target noun in a specific location on the screen and typed the corresponding picture description into a response box on the screen, with "d___ . . . " requiring the use of a definite article and "e___ . . . " of an indefinite article (Figure 1f). This test was similar to the active production task during the training phase, except that it was written rather than spoken. Fifteen items required the use of a definite article, and the remaining 15

Table 3. Error-monitoring trials

		Error location			
		Article			
Type	Number of trials	Definite	Indefinite	Adjectives	Word order
Correct	18	✓	–	✓	✓
	17	–	✓	✓	✓
Article error	15	x	–	✓	✓
	15	–	x	✓	✓
Adjective error	15	✓	–	x	✓
	15	–	✓	x	✓
Distractor	5	✓	–	✓	x
	5	–	✓	✓	x

Note. Check marks indicate correct articles, adjectives, or word order, x-marks indicate incorrect articles, adjectives, or word order, and dashes indicate an inapplicable category.

required an indefinite article. Participants thus had to produce each noun twice, one time with a definite article and one time with an indefinite article.

Procedure

All participants completed the training followed by all test measures on a computer in a computer lab during one single session of about 90 min. While participants completed each training task and test individually, using headphones to hear the auditory recordings, multiple participants assigned to the same experimental group were in the computer lab at the same time. After finishing the test measures, participants completed a language background questionnaire.

Data processing

Data from 41 participants are included in the analyses and results for the comprehension tests. For all accuracy analyses, we removed one trial from the forced-choice test with nouns, one trial from the error-monitoring test, and one trial from the production test due to a coding error. Data were analyzed using mixed-effects logistic regression analyses with the lme4 package version 1.1-21 (Bates et al., 2015) in R version 3.5.1 (R Development Team, 2018). For the error-monitoring task, we calculated aggregate *d'* scores on correct sentences and sentences with incorrect gender marking on articles and adjectives for each participant. Sentences with word order errors were not included, as these only served as distractors.[5] We then compared the *d'* score between PROD and COMP using a simple linear regression.

Due to a programming error, data for the written production test were not collected from seven participants, leaving 34 participants, ($n = 21$ for COMP; $n = 13$ for PROD). For the analysis of gender agreement accuracy in the written production data, we only included participants who produced at least one correct noun. Of all attempted nouns, only 29 participants produced at least one correct noun ($M = 15.1$; range: 5–26) and were included in the agreement analysis ($n = 18$ for COMP; $n = 11$ for PROD). A native speaker of German and a research assistant familiar with German coded the nouns for accuracy. Nouns with no spelling errors or maximally one spelling error were coded as correct (55.0% of all produced nouns). If there was more than one spelling error but the noun was still recognizable as the target noun, the primary researcher and the research assistant independently coded the nouns as correct or incorrect, with an interrater reliability of 92.3%. Cases in which the coding differed between the primary researcher and the research assistant were discussed until they reached agreement. This led to the inclusion of another 3.2% of the total nouns in the analysis, for an overall 58.2%. For gender agreement accuracy, we then coded productions as correct when both the article and the two attributive adjectives agreed in grammatical gender with the correctly produced noun. Productions in which either the article, the adjectives, or both the article and the adjectives mismatched the target noun's grammatical gender were first coded for which element(s) were incorrect, and were then subsequently coded as overall incorrect for the accuracy analysis of gender agreement.

The initial models for accuracy included experimental group as a fixed-effect (PROD vs. COMP), sum-coded −0.5 and 0.5. The final random effect structure was determined by starting with the maximum structure justified by the experimental design (Barr et al., 2013), which included random intercepts for participants and items and correlated random by-item slopes for group. For the agreement analysis of the production data, random slopes were subsequently removed due to non-convergence to fit the maximum model justified by the data.

While RT data were collected, there was no significant difference across groups for any measures. The results and analyses are therefore not further reported here (see Appendices C and D).[6]

Results

Accuracy

Descriptive results for all accuracy tests are shown in Figures 2–5, and Tables 4–6 present the results of all statistical analyses.

Forced-choice tests

As seen in Table 4 and Figure 2, on the forced-choice test without nouns, there was a significant effect of Group. Participants in PROD ($M = 0.77$, $SD = 0.41$) were more accurate than the participants in COMP ($M = 0.69$, $SD = 0.46$), $\beta = -0.58$, $p < .05$. On the forced-choice test with nouns, there was no effect of Group, with participants in PROD and COMP performing close to ceiling (PROD: $M = 0.98$, $SD = 0.14$; COMP: $M = 0.99$, $SD = 0.12$), $\beta = 1.04$, $p = .19$.

Error monitoring

As seen in Figure 3, on the error-monitoring test, participants in PROD were descriptively better at accurately judging correct sentences than participants in

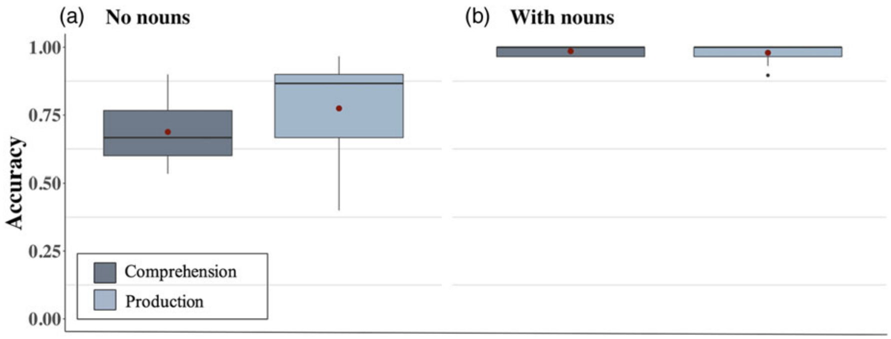


Figure 2. Accuracy for Forced-Choice Tests. Box Plots Show 1st and 3rd Quartiles as Well as Median (Horizontal Black Line) and Mean Accuracy (Red Dot) by Group. Whiskers on Each Box Plot Extend No Further than 1.5 Times the Interquartile Range.

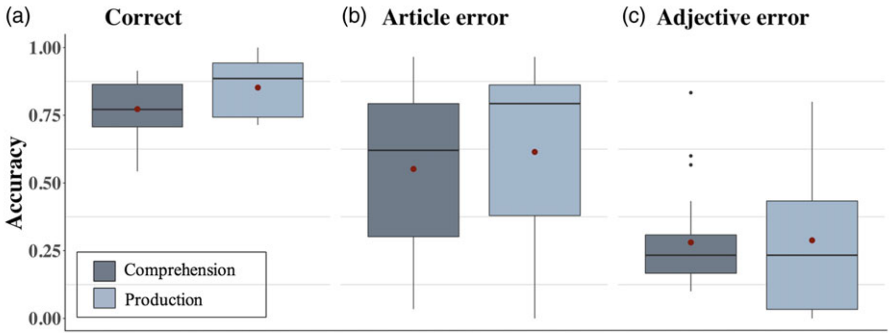


Figure 3. Accuracy for Error-Monitoring Test. Box Plots Show 1st and 3rd Quartiles as Well as Median (Horizontal Black Line) and Mean Accuracy (Red Dot) by Group. Whiskers on Each Box Plot Extend No Further than 1.5 Times the Interquartile Range.

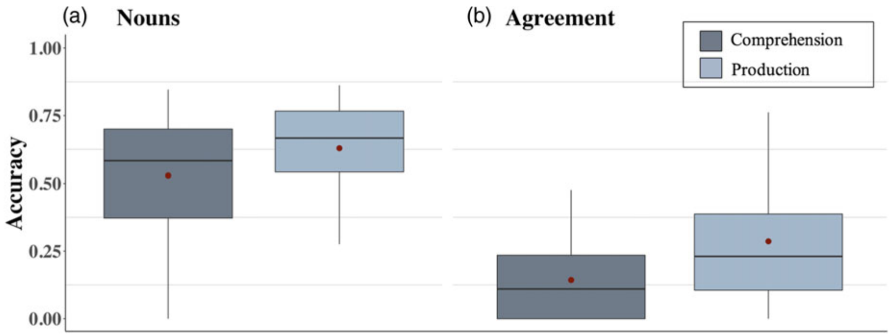


Figure 4. Accuracy for Written Production Test. Box Plots Show 1st and 3rd Quartiles as Well as Median (Horizontal Black Line) and Mean Accuracy (Red Dot) by Group. Whiskers on Each Box Plot Extend No Further than 1.5 Times the Interquartile Range.

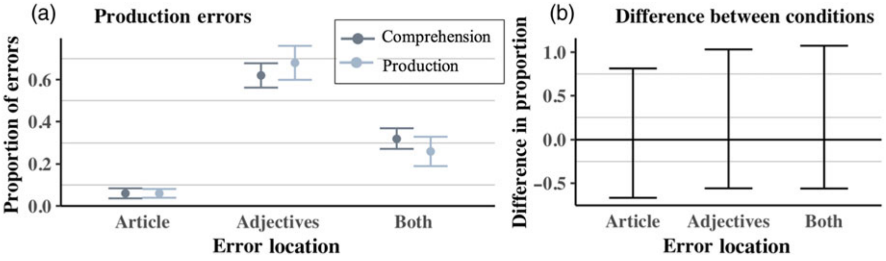


Figure 5. (a) Proportion of Errors by Location for Each Condition (*Mean* ± *SE*) (b) Simulated 95% CI of Difference Between Conditions in Proportion of Errors by Location.

Table 4. Summary of mixed logit models on accuracy for the forced-choice (FC) comprehension and written production tests

Predictor	Parameter estimates		Wald's test	
Fixed effects	Estimate	Std. error	*z*-value	*Pr* (>|*z*|)
FC suffix no nouns – full model				
(Intercept)	1.1673	0.1570	7.433	1.06e-13
Group	−0.5841	0.2603	−2.244	0.0248 *
FC suffix with nouns – full model				
(Intercept)	4.5808	0.4794	9.556	<2e-16
Group	1.0386	0.8063	1.288	0.198
Written production (nouns) – full model				
(Intercept)	0.4244	0.3692	1.150	0.250
Group	−0.6074	0.5085	−1.195	0.232
Written production (agreement) – full model				
(Intercept)	−1.9330	0.3979	−4.858	1.19e-06
Group	−1.2951	0.6338	−2.043	0.041 *

Note. Signif. codes: * $p \leq .05$.

COMP (PROD: $M = 0.85$, $SD = 0.36$; COMP: $M = 0.77$, $SD = 0.42$). For sentences with an error on the article, participants in PROD were also slightly more accurate than participants in COMP (PROD: $M = 0.61$, $SD = 0.48$; COMP: $M = 0.55$, $SD = 0.50$). However, note the rather large amount of within-group variation for sentences with article errors, as represented by the large boxes and whiskers in the box plots. For sentences with an error on the adjective, on the other hand, both groups performed equally low (PROD: $M = 0.29$, $SD = 0.45$; COMP: $M = 0.28$, $SD = 0.45$).

The *d'* score analysis, which included correct sentences as well as sentences with errors on articles or adjectives but not word order, indicated that participants in PROD were overall more accurate in identifying correct and rejecting incorrect sentences than participants in COMP (PROD: $M = 0.89$, $SD = 0.84$; COMP: $M = 0.54$,

Table 5. Summary of the simple linear regression model on the d' score in the error-monitoring test

	Unstandardized coefficients		Standardized coefficients		
Predictor	B	Std. error	*β*	*t*	Sig.
(Constant)	0.525	0.132	0.000	4.048	2.37e-4
Group	−0.350	0.205	0.2633	−1.705	0.096†

Note. Signif. codes: † ≤ .10.

Table 6. Proportions of agreement production errors by group

	COMP		PROD		
Error location	*M*	*SE*	*M*	*SE*	95% CI (low, high)
Article	0.06	0.024	0.06	0.021	[−0.668, 0.813]
Adjectives	0.62	0.058	0.68	0.081	[−0.559, 1.032]
Both	0.32	0.048	0.26	0.070	[−0.561, 1.074]

$SD = 0.46$). However, the effect of Group was only marginally significant ($\beta = -0.35$, $p = .096$; see Table 5).

Written production

As seen in Figure 4, participants in PROD descriptively produced more correct nouns than participants in COMP (PROD: $M = 0.64$, $SD = 0.48$; COMP: $M = 0.55$, $SD = 0.49$), but there was no significant effect of Group, $\beta = -0.61$ $p = .232$. In terms of accurately producing articles and adjectives that agreed in grammatical gender on those items where participants produced the correct noun, there was a significant effect of Group. As seen in Table 4 and Figure 4, participants in PROD were more accurate than participants in COMP (PROD: $M = 0.34$, $SD = 0.47$; COMP: $M = 0.17$, $SD = 0.38$), $\beta = -1.30$, $p < .05$. However, note the rather large amount of within-group variation, as represented by the large boxes and whiskers in the box plot in Figure 4.

On the production test, we were also interested in the errors produced by each group. As there was considerable variation within and across groups as to whether participants produced incorrect articles, adjectives, or both, we compared between-group effect sizes and confidence intervals (CIs) instead of using mixed-effects logistic regression models (Cumming, 2014). To do so, we calculated the proportion of erroneous articles, adjectives, and both to the total number of errors for each participant and then generated bootstrapped 95% CIs and effect sizes with a standardized scale and Hedge's *g* to compare the proportion of each error type between COMP and PROD using the BootES package in R (Kirby & Gerlanc, 2013).

As seen in Table 6 and Figure 5, there was no effect of Group when examining whether the experimental groups differed in terms of their erroneous productions, that is, whether one group produced more incorrect articles, adjectives, or combinations of both, as shown by CIs that span zero in Figure 5b. Thus, although the PROD participants were more accurate than the COMP participants overall, participants in both groups produced comparable proportions of errors within each category, with the greatest number of errors occurring on adjectives.[7,8]

Discussion

The present study asked whether there are differences in the effectiveness of comprehension-based versus production-based training for comprehending and producing grammatical gender marking among beginning L2 learners of German. We hypothesized that production-based training is more beneficial, similar to Hopman and MacDonald (2018). In our study, participants in both the comprehension-based and the production-based group were at ceiling for comprehending the nouns from the training session, as measured by a forced-choice comprehension task. In contrast, participants in the production-based group were more accurate than the comprehension-based group on a corresponding forced-choice task, in which comprehension required the accurate processing of grammatical gender information (e.g., *ein blauer gepunkteter* … “the$_{\text{MASC}}$ blue$_{\text{MASC}}$ dotted$_{\text{MASC}}$ one”). Similarly, participants in both experimental groups were equally accurate in their production of nouns in a written production task, but participants in the production-based group were more accurate at producing the correct grammatical gender marking on articles and adjectives than participants in the comprehension-based group. In terms of monitoring agreement patterns in the error-monitoring task, participants in the production-based group were marginally more accurate than participants in the comprehension-based group. Participants in both experimental groups were thus similarly accurate in comprehending and producing the trained nouns, but in all tasks that required knowledge of grammatical gender and gender agreement marking, participants in the production-based group were significantly more accurate. These results show a clear advantage of production-based training for the learning of grammatical gender in beginning L2 learners of German in both production and comprehension tests targeting grammatical gender, confirming both of our hypotheses.

These results largely replicate Hopman and MacDonald (2018), who found that participants with production-based training were faster and more accurate at comprehending grammatical dependencies in an artificial language than participants with comprehension-based training. In addition, our results show that the advantages of production-based training extend to producing accurate grammatical gender marking in a natural language. By replicating Hopman and MacDonald, with a more opaque agreement paradigm than the one implemented in their study, we demonstrate that this particular finding generalizes from an artificial language study to the context of natural language learning. Furthermore, advantages for production-based training were found despite the fact that participants had had 4–5 weeks of classroom exposure to L2 German prior to the experiment, suggesting that the generalization of effects from artificial to natural language learning settings may not necessarily be attenuated by prior basic L2 knowledge (see Paul & Grüter, 2016, for discussion).

The present results contrast with previous studies comparing the effectiveness of CBI and PBI for the learning of L2 grammatical forms, as no studies have found an advantage of PBI over CBI for the comprehension of the target grammatical forms immediately after training (e.g., Allen, 2000; Farley & Aslan, 2012; Morgan-Short & Bowden, 2006). Similarly, the meta-analysis conducted by Shintani et al. (2013) indicated that only on delayed posttest measures did an advantage for PBI over CBI emerge, with PBI only outperforming CBI on production-based but not comprehension-based tasks. Admittedly, the present study did not include any delayed posttest measures, limiting our ability to draw direct comparisons with the full set of outcomes of the meta-analysis of Shintani et al. However, a follow-up study is currently underway to investigate whether the advantages seen here for the production-based group are maintained over time.

In the present study, better performance by the production-based group cannot be due to teaching to the test; if that were the case, the comprehension-based group should have been more accurate on the forced-choice measures of comprehension, as their training was similar. Additionally, the production-based group showed higher accuracy on the production test, despite training requiring spoken production, whereas the production posttest required participants to produce written descriptions. There are various reasons for why the present findings differ from previous studies. As DeKeyser and Botana (2015) and Toth (2006) point out, the limited effectiveness of PBI in many previous studies may be due to the nature of the output-based activities in those studies, which often involve mechanical drills in the initial stages of training rather than meaning-oriented output activities. In the present study, however, both experimental groups completed meaning-based activities throughout the entire training. Further, both groups received the same amount and type of feedback during their respective active learning blocks, which provided the learners in both training groups with informative model utterances. Meaningful language production in conjunction with informative feedback may thus be particularly beneficial for learning to comprehend and produce L2 grammatical features.

What underlying mechanisms might contribute to the advantage of production-based training, not only on the post-tests targeting production but also on the post-tests targeting comprehension? We propose two complementary mechanisms that can account for these findings, namely memory retrieval and errorful generation. The first account suggests that different memory retrieval processes are involved in comprehension versus production. Whereas comprehension only requires that learners recognize target forms and map those forms to their intended meaning, language production, as implemented in the production-based group, involves recalling material from memory. Previous research has shown that recall practice leads to better learning than simple recognition in the domain of L2 vocabulary learning (e.g., Kang et al., 2013; Karpicke & Bauernschmidt, 2011; Karpicke & Roediger, 2008) and the learning of agreement relationships in an artificial language learning paradigm (Hopman & MacDonald, 2018).

In addition to recall practice, language production requires learners to retain both the to-be-produced message and the language forms associated with that message in working memory during utterance planning. This process allows for memory traces of the different elements of the utterance to bind, creating stronger representations in memory (MacDonald, 2016). For the present study, this would imply that during meaningful language production, a noun and its corresponding grammatical gender were more strongly associated than during meaningful language comprehension. Language production training thus not only offers opportunities for recall and retrieval but also generates beneficial opportunities for item-based learning by creating robust associations between a noun and its grammatical gender information, that is, gender assignment. This knowledge of a noun's gender assignment is then available not only for producing the target form but also generalizes to the domain of comprehension, as shown by the higher accuracy of the production-based group in the forced-choice comprehension test without nouns compared to the comprehension-based group. These findings are in agreement with the Output Hypothesis (Swain, 1995, 2005), which emphasizes the importance of language output for the development of deterministic processing in an L2.

The second mechanism that can account for our findings suggests that the errorful generation of language forms, especially in the absence of prior knowledge, is particularly beneficial for learning, as has previously been shown for vocabulary learning (Potts & Shanks, 2014; Potts et al., 2019). Under this account, learners in the production-based group generated predictions about the appropriate utterances to describe the images in the active learning task. The uncertainty related to the learner's lack of previous knowledge regarding whether the produced utterance was correct or incorrect increased her curiosity and the attention paid to the subsequent feedback. This then enabled the learner to better detect discrepancies between her production and the feedback, which in turn enhanced the encoding of language information provided in the feedback, leading to adjustments of the linguistic system (Dell & Chang, 2014; see also Swain, 1995, 2005), in this case for vocabulary and related grammatical features. Although we did not collect measures of awareness or attention, including such measures in future research would provide evidence for or against the account sketched here. However, the fact that in this study the production-based group outperformed the comprehension-based group even though both training groups received equivalent feedback after their responses during the active learning blocks is in line with previous research showing that errorful comprehension in the absence of errorful generation is less beneficial (Potts & Shanks, 2014).

In the present study, both groups exhibited similar learning of the nouns. However, the production-based group was significantly better at tasks that specifically required knowledge of the nouns' grammatical gender, showing that language production enhances the learning of item-specific grammatical gender features. Additionally, the present study also offers evidence that learning in the production-based group went beyond just the learning of individual lexical items and their gender assignment, as evidenced by their higher accuracy on agreement between article, adjective, and nouns on the production test, where all color-pattern-noun pairings were new and participants could therefore not rely on explicit memory of these phrases (cf. Hartsuiker & Bernolet, 2017). This suggests that production-based training may create more favorable contexts for the development of stable, abstract grammatical gender features compared to comprehension-based training. This applies in particular in the context of German, where few reliable cues exist to indicate a noun's membership in a particular gender category and rote learning of a noun's grammatical gender feature – as marked by the determiner – may be especially important (e.g., Hopp, 2016). Future research should investigate whether similar advantages for production-based activities exist for grammatical features that are less directly tied to individual lexical items.

The complementary accounts presented here both underline the importance of language production in enabling learners to acquire grammatical features and to move towards more generalized, abstract mental representations of these grammatical features. Future research should attempt to investigate how these accounts might interact and contribute individually or in tandem to the L2 acquisition of other grammatical dependencies.

A further factor contributing to the advantage we find for the production-based group might be that free language production is arguably more cognitively effortful than comprehension (e.g., Boiteau et al., 2014). Previous studies investigating L2 vocabulary learning have shown that difficult training conditions, which induce so-called desirable difficulties (Bjork, 1994), yield greater benefits than easier training conditions (e.g., Bjork et al., 2013; Bjork & Kroll, 2015; Karpicke & Roediger, 2008). Free language production could thus be one way of inducing desirable difficulties that can lead to improved learning of grammatical gender assignment.

Despite advantages for the production-based group in both comprehension and production tests that require knowledge of an item's grammatical gender, the results also show large within-group variation. While production-based training is thus more beneficial than comprehension-based training for learning grammatical gender in L2 German overall, it may not be equally beneficial for all participants in the production-based group. Future research that seeks to account for this variation and the factors that contribute to it may shed light onto why production-based training affords learning advantages in the first place.

Conclusion

The present study provides evidence that production-based training is more beneficial for the learning of grammatical gender in beginning L2 learners of German than comprehension-based training on both production- and comprehension-based posttest measures. We argue that these results are best explained by effortful meaning-based language production providing learners with more beneficial opportunities to recall forms from memory and to compare their output with the feedback they receive. These findings have important implications for the classroom: Production-based training appears to give beginning L2 learners an advantage at creating robust linguistic representations, thus highlighting the need for production-based exercises in the foreign language learning classroom, perhaps especially for lexically based structures like grammatical gender that initially require L2 learners to learn the appropriate target forms individually for each word. At the same time, many open questions remain. Future research should investigate (1) how memory retrieval and errorful generation might interact and contribute individually or in tandem to the L2 acquisition of grammatical dependencies, (2) the processes involved in noticing discrepancies between output and feedback, (3) whether similar advantages for production-based activities exist for grammatical features that are less directly tied to individual lexical items, and (4) whether advantages of production-based training along the lines of the training method implemented here are maintained over a longer period of time. With the present study as an important first step, investigating these questions will shed further light on why production matters while simultaneously enabling us to create more effective L2 learning materials.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/S014271642100014X

Acknowledgements. We thank audiences at the 32nd CUNY Conference on Human Sentence Processing 2019, the 12th International Symposium on Bilingualism 2019, and the PSMLA Fall Conference 2019 for their feedback on this work. We also want to thank Maryellen MacDonald for sharing her thoughts on this study, as well as Shane Cummings, Liz DeFelice, and Rosa Padt for their help with stimulus design, data collection, and analysis.

Financial support. This study was funded by resources from the College of Liberal Arts at Penn State. EH was supported by NSF Grant number 1849236 and by a Menzies Graduate Research Award, Department of Psychology, University of Wisconsin-Madison.

Conflict of interest. We have no known conflicts of interest to disclose.

Notes

1. A power analysis was conducted using GPower 3.1 (Faul et al., 2007) based on the effect size for comprehension tasks from Shintani et al. (2013) (Cohen's $d = 1.09$), a significance level of $\alpha = 0.05$ and a desired power level of 0.8. This revealed that 15 participants per group were necessary to detect significant between-group differences in comprehension tasks for the present study. No corresponding power analysis was conducted for production tasks, because the reported between-group effect size for production tasks in Shintani et al. was negligible.

2. Some participants had been exposed to grammatical gender and gender agreement through their L1, for example, in Russian or Spanish. We reran the data analyses for all tests excluding participants who had a gendered L1 ($N = 3$; all in the comprehension-based group). The group results were no different than before, showing an advantage for the production-based group. We therefore retained all participants in our analyses to increase statistical power.

3. Of all participants who were included in the final data analysis, three did not complete the language background questionnaire and one chose to not provide any gender information. They are indicated as "no answer." Proficiency information in Table 2 is therefore based on the remaining 38 participants.

4. Participants were first-semester German learners without prior German experience, who had not yet encountered the targeted nouns in their classroom vocabulary and had not been introduced to gender agreement between adjectives and nouns. Therefore, no pre-test to account for potential between-group differences in grammatical gender knowledge for these nouns was necessary, as performance on any testing measure prior to training would have been at chance level.

5. Sentences with word order errors still showed a descriptively higher accuracy in PROD than in COMP.

6. For the RT analyses, only correct trials were included. Of these, trials in which participants responded faster than 200 ms or slower than 10,000 ms, and in which participants responded more than 3 *SDs* above or below a participant's own mean, were also removed. These measures resulted in the exclusion of 29.9% of FC Suffix NN, 5.0% of FC Suffix WN, and 46.8% of EM data. The remaining data were analyzed using linear mixed-effects analyses in R. None of the models showed significant effects of Group.

7. We compared performance accuracy on definite and indefinite articles in error-monitoring trials and written production trials. Participants in both groups were more accurate at identifying incorrect indefinite than definite articles in the error-monitoring test, and more accurate in producing correct indefinite than definite articles in the production test, but there were no significant differences between groups. Furthermore, the differences in accurately identifying incorrect articles and producing accurate articles are in the expected direction: Due to the overlap in form between indefinite masculine and neuter articles, only two forms are available, increasing the probability of making a correct error-monitoring decision and producing a correct indefinite article compared to definite articles, all of which take different forms. As this difference in accuracy for definite versus indefinite articles is not central to our primary research questions, these differences will not be discussed further.

8. We attribute the lower accuracy in trials targeting adjectives in the error-monitoring test and the higher proportion of errors on adjective gender agreement in the written production test compared to articles to the relative perceptual saliency of gender marking on determiners compared to adjectives. While determiners constitute separate lexical items, grammatical gender is marked via suffixation on adjectives, leading to lower perceptual saliency (see Ellis, 2006).

References

Allen, L. Q. (2000). Form-meaning connections and the French causative: An experiment in processing instruction. *Studies in Second Language Acquisition*, **22**(1), 69–84.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, **122**, 292–305.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.

Benati, A., & Lee, J. F. (2008). From processing instruction on the acquisition of Italian noun-adjective agreement to secondary transfer-of-training effects on Italian future tense verb morphology. In A. Benati & J. F. Lee (Eds.), *Grammar acquisition and processing instruction: Secondary and cumulative effects* (pp. 54–87). Multilingual Matters.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, **64,** 417–444.

Bjork, R. A. & Kroll, J. F. (2015). Desirable difficulties in vocabulary learning. *The American Journal of Psychology*, **128**(2), 241–252.

Boiteau, T. W., Malone, P. S., Peters, S. A., & Almor, A. (2014). Interference between conversation and a concurrent visuomotor task. *Journal of Experimental Psychology: General*, **143**(1), 295–311.

Bordag, D., Kirschenbaum, A., Rogahn, M., Opitz, A., & Tschirner, E. (2017). Semantic representation of newly learned L2 words and their integration in the L2 lexicon. *Studies in Second Language Acquisition*, **39**(1), 197–212.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, **25**, 7–29.

De Jong, N. (2005). Can second language grammar be learned through listening? An experimental study. *Studies in Second Language Acquisition*, **27**, 205–234.

DeKeyser, R., & Botana, G. P. (2015). The effectiveness of processing instruction in L2 grammar acquisition: A narrative review. *Applied Linguistics*, **36**(3), 290–305.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **369**(1634), 20120394.

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, R. (2012). *Language teaching research and pedagogy*. Wiley-Blackwell.

Ettlinger, M., Morgan-Short, K, Faretta-Stutenberg, M., & Wong, P. C. M. (2016). The relationship between artificial and second language learning. *Cognitive Science*, **40,** 822–847.

Farley, A. & Aslan, E. (2012). The relative effects of processing instruction and meaning-based output instruction on L2 acquisition of the English subjunctive. *ELT Research Journal*, **1**(2), 120–141.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3. *Behavior Research Methods*, **39**, 175–191.

Günther, C. (2013). *The elliptical noun phrase in English. Structure and use.* Routledge.

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research*, **28**(2), 191–215.

Hartsuiker, R. J. & Bernolet, S. (2017). The development of shared syntax in second language learning. *Bilingualism: Language and Cognition*, **20**(2), 219–234.

Hopman, E. W. M. & MacDonald, M. C. (2018). Production Practice During Language Learning Improves Comprehension. *Psychological Science*, **29**(6), 961–971.

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research*, **29**(1), 33–56.

Hopp, H. (2016). Learning (not) to predict: Grammatical gender processing in second language acquisition. *Second Language Research*, **32**(2), 277–307.

Kang, S. H., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin & Review*, **20**(6), 1259–1265.

Karpicke, J. D. & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, **37**(5), 1250–1257.

Karpicke, J. D. & Roediger H. L. (2008). The critical importance of retrieval for learning. *Science*, **331**, 772–775.

Kirby, K. N., & Gerlanc, D. (2013). BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods*, **45**, 905–927.

Köpcke, K.-M. & Zubin, D. A. (1983). The cognitive organization of gender assignment of monosyllabic nouns in contemporary German. *Zeitschrift für Germanistische Linguistik*, **11,** 166–182.

Köpcke, K.-M. & Zubin, D. A. (1984). Sechs Prinzipien für die Genuszuweisung im Deutschen: Ein Beitrag zur natürlichen Klassifikation. *Linguistische Berichte*, **93,** 26–50.

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon Press.

Lee, J. F., & Benati, A. (2007). Comparing modes of delivering processing instruction and meaning-based output instruction on Italian and French subjunctive. In J. F. Lee & A. Benati (Eds.), *Delivering processing instruction in classrooms and in virtual contexts: Research and practice* (pp. 99–136). Equinox.

McDonald, M. (2016). Speak, act, remember: The language-production basis of serial order and maintenance in verbal memory. *Current Directions in Psychological Science*, **25**(1), 47–53.

Morgan-Short, K. & Bowden, H. (2006). Processing instruction and meaningful output-based instruction: Effects on second language development. *Studies in Second Language Acquisition*, **28,** 31–65.

Paul, J. Z., & Grüter, T. (2016), Blocking effects in the learning of Chinese classifiers. *Language Learning*, **66**, 972–999.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, **36**(4), 329–347

Potts, R., Davies, G., & Shanks, D. R. (2019). The benefit of generating errors during learning: What is the locus of the effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **45**(6), 10231041.

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, **143**(2), 644–667.

R Development Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved Feb 9, 2018 from http://www.R-project.org

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, **11**(2), 129–158.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press.

Soruç, A., Qin, J., & Kim, Y. (2017). Comparing the effectiveness of Processing Instruction and production-based instruction on L2 grammar learning: The role of explicit information. *TESL Canada Journal*, **34**(2), 49–70.

Shintani, N., Li, S., & Ellis, R. (2013). Comprehension-based versus production-based grammar instruction: A meta-analysis of comparative studies. *Language Learning*, **63**(2), 296–329.

Shintani, N. (2015). The effectiveness of processing instruction and production-based instruction on L2 grammar acquisition: A meta-analysis. *Applied Linguistics*, **36**(3). 306–325.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics* (pp. 125–144). Oxford University Press.

Swain, M. (2005). The Output Hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook on research in second language learning and teaching* (pp. 471–483). Lawrence Erlbaum.

Tanaka, T. (2001). Comprehension and production practice in grammar instruction: Does their combined use facilitate second language acquisition? *JALT Journal*, **23,** 6–30.

Toth, P. D. (2006). Processing Instruction and a role for output in second language acquisition. *Language Learning*, **56**(2), 319–385.

Truscott, J., & Sharwood Smith, M. (2004). Acquisition by processing: A modular perspective on language development. *Bilingualism: Language and Cognition*, **7**, 1–20.

VanPatten, B. (1996). *Input processing and grammar instruction in second language acquisition: Theory and research.* Ablex.

VanPatten, B. (2002). Processing instruction: An update. *Language Learning*, **52**, 755–803.

VanPatten, B., Williams, J., & Rott, S. (2004) Form-meaning connections in second language acquisition. In B. VanPatten, J. Williams, S. Rott, & M. Overstreet (Eds.), *Form-meaning connections in second language acquisition* (pp. 1–30). Lawrence Erlbaum.

VanPatten, B. (2004). Input and output in establishing form-meaning connections. In B. VanPatten, J. Williams, S. Rott, & M. Overstreet (Eds.), *Form-meaning connections in second language acquisition* (pp. 31–50). Lawrence Erlbaum.

VanPatten, B. (2013). Input processing. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 268–281). Routledge.

VanPatten, B. & Cadierno, T. (1993). Explicit instruction and input processing. *Studies in Second Language Acquisition*, **15**(2), 225–243.

VanPatten, B. & Wong, W. (2004). Processing instruction and the French causative: Another replication. In B. VanPatten (Ed.), *Processing Instruction: Theory, research, and commentary* (pp. 99–120). Lawrence Erlbaum.

Yamashita, T. & Iizuka, T. (2017). The effectiveness of structured input and structures output on the acquisition of Japanese comparative sentences. *Foreign Language Annals*, **50**(2), 387–397.

Appendix A. Training material

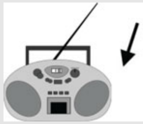
Type	German	English
Nouns	der/ein Becher	the/a cup
	der/ein Ordner	the/a binder
	der/ein Wecker	the/an alarm clock
	der/ein Stiefel	the/a boot
	der/ein Eimer	the/a bucket
	die/eine Dose	the/a can
	die/eine Kerze	the/a candle
	die/eine Schüssel	the/a bowl
	die/eine Flasche	the/a bottle
	die/eine Tasche	the/a purse
	das/ein Fahrrad	the/a bike
	das/ein Sparschwein	the/a piggybank
	das/ein Flugzeug	the/an airplane
	das/ein Geschenk	the/a gift
	das/ein Plüschtier	the/a stuffed animal
Adjectives	rot	red
	blau	blue
	grün	green
	braun	brown
	gepunktet	dotted
	gestreift	striped
	kariert	checkered
	gemustert	patterned
Verbs	steht	stands
Prepositional phrases	neben dem Radio	next to the radio
	neben dem Bett	next to the bed
	neben dem Sofa	next to the sofa

Appendix B. Training method

Block	Trials	Type	Block type	Example utterance	Example picture	Rationale for number of trials
1	15	Definite article + noun	PE	*der Becher*		There are 15 different nouns
	15	Definite article + noun	AL			
2	15	Indefinite article + noun	PE	*ein Becher*		
	15	Indefinite article + noun	AL			
3	4	Cognate color adjective	PE	*blau*		There are four different colors
	4	Cognate color adjective	AL			
4	15	Article + color adjective + noun	PE	*ein blauer Becher*		The four different colors are balanced as well as possible across the 15 nouns
	15	Article + color adjective + noun	AL			
5	4	Non-cognate pattern adjective	PE	*gepunktet*		There are four different patterns
	4	Non-cognate pattern adjective	AL			
6	15	Article + color and pattern adjectives + noun	PE	*ein blauer gepunkteter Becher*		The four different colors and patterns are balanced as well as possible across the 15 nouns
	15	Article + color and pattern adjectives + noun	AL			
7	15	Article + color and pattern adjectives	PE	*ein blauer gepunkteter . . .*		The target nouns are left out to focus attention on gender marking
	15	Article + color and pattern adjectives	AL			

(*Continued*)

Appendix B. (*Continued*)

Block	Trials	Type	Block type	Example utterance	Example picture	Rationale for number of trials
8	3	Verb and location	PE	*steht neben dem Radio*		There are three different phrases capturing different locations
	3	Verb and location	AL			
9	15	Article + color and pattern adjectives + noun + verb and location	PE	*Ein blauer gepunkteter Becher steht neben dem Radio.*		There are 15 sentences with the 15 different nouns, counterbalanced for colors, patterns, and locations
	15	Article + color and pattern adjectives + noun + verb and location	AL			
10	15	Article + color and pattern adjectives + verb and location	PE	*Ein blauer gepunkteter . . . steht neben dem Radio.*		The target nouns are left out to focus attention on gender marking
	15	Article + color and pattern adjectives + verb and location	AL			

Note. PE = passive exposure; AL = active learning.

Appendix C. Descriptive statistics of RTs in ms in the forced-choice comprehension and error-monitoring tests after data exclusion and trimming

	RTs			
	COMP		PROD	
Test	*M*	*SD*	*M*	*SD*
FC suffix no nouns	3,394	1,982	3,326	1,763
FC suffix with nouns	1,898	1,387	1,376	880
Error monitoring	3,885	2,092	4,396	2,088

Appendix D. Summary of linear mixed-effects models on RTs for the forced-choice (FC) comprehension and error-monitoring (EM) tests

Predictor	Parameter estimates			*F* test	
Fixed effects	Estimate	Std. error	*F*	*df* error	*Pr*(> *F*)
FC suffix no nouns – full model					
(Intercept)	3,375.8	158.3	454.486	38.81	<2e-16
Group	114.8	316.7	0.131	38.81	0.719
FC suffix with nouns – full model					
(Intercept)	1,642.9	134.1	150.172	39	6e-15
Group	527.1	268.1	3.865	39	0.057
EM – full model					
(Intercept)	4,079.4	175.5	539.996	42.65	<2e-16
Group	−473.1	342.7	1.906	38.94	0.175

Note. *p*-values based on Kenward–Roger approximation.

Cite this article: Keppenne, V., Hopman, EWM., and Jackson, CN. (2021). Production-based training benefits the comprehension and production of grammatical gender in L2 German. *Applied Psycholinguistics*. https://doi.org/10.1017/S014271642100014X