JESLA | Journal of the European Second Language Association

Baten, K., and Cornillie, F. (2019). Elicited imitation as a window into developmental stages. *Journal of the European Second Language Association,* 3(1), 23–34. DOI: https://doi.org/10.22599/jesla.56

RESEARCH

Elicited imitation as a window into developmental stages

Kristof Baten* and Frederik Cornillie†,‡

In the second language acquisition literature, data of naturally occurring language use are considered the most ideal data to make statements about second-language (L2) development. This study examines to what extent experimentally elicited data can provide an equally valid basis for determining L2 development, by testing predictions derived from Processability Theory regarding the L2 acquisition of the German case system. Using naturally occurring language data, previous research on L2 German case acquisition has uncovered three developmental stages. The present cross-sectional study investigates whether the same stages occur in data obtained from an experimental task (i.e., a computer oral elicited imitation task (OEIT). Thirty-six university L2 learners of German participated in the study. The results show that the elicited data prove comparable to the naturally occurring data. As such, this study corroborates a previous validation study on developmental stages in L2 English, which demonstrated the comparability of naturally occurring and experimentally elicited data. In addition, concerning methodological advancement of the OEIT design, the present study proposes to include a direct measure of comprehension.

Keywords: data elicitation; elicited imitation task; implicit knowledge; interlanguage development; L2 German

1. Introduction

In second language acquisition (SLA), it is a long-established (but not indisputable) finding that second-language (L2) learners follow predictable stages in the acquisition of grammatical structures (Abrahamsson, 2013). This finding is largely based on evidence from naturally occurring L2 use. Indeed, for reasons of external validity, many SLA researchers prefer data of naturally occurring L2 use for making statements about what learners have learned (Ellis, 2008). However, the problem with this kind of data is that it often does not contain enough examples of the more difficult linguistic structures. When the data fail to include sufficient examples of these structures, the issue arises concerning to what extent the data provide valid evidence of developmental stages. A possible solution for this problem is to employ elicitation techniques in more experimental settings. A good candidate of such a technique, which, as this study will show, can tap into both receptive and productive L2 knowledge, is the oral elicited imitation task (OEIT) (Erlam, 2006). In its most basic form, the OEIT requires learners to repeat oral stimuli. Because of its closed-response design, the obvious methodological advantage of the OEIT is that researchers can control relatively well which linguistic structures appear in the learners' speech. As such, more evidence will be available with regard to specific linguistic structures, especially the more difficult ones.

However, in order to validly use OEITs for determining developmental stages, it needs to be established whether or not they tap into the type of knowledge that is considered to be underlying spontaneous L2 use (i.e., implicit knowledge). According to Ellis (2005), implicit knowledge is part of procedural knowledge. It is not consciously held and processed automatically. In contrast, explicit knowledge is conscious knowledge that can be consciously searched and expressed in a verbal statement (Ellis, 2005; Dörnyei, 2009). In recent years, SLA research has focused on examining whether tests can be developed that provide separate measures of implicit and explicit L2 knowledge (for a review of such measures, see Rebuschat, 2013). A handful of such studies have lent empirical support for OEITs as a measure of implicit linguistic knowledge. Studies using exploratory factor analyses revealed that the tests intended to measure implicit knowledge (among which the OEIT) loaded on one factor, while the tests intended to measure explicit knowledge loaded on another factor (Ellis, 2005; Bowles, 2011; Spada et al., 2015). Further, in a correlation study, Erlam (2006) found a significant positive correlation between scores on the OEIT and scores on an oral narrative task. Similarly, Tracy-Ventura et al. (2014) showed strong correlations among scores on a French OEIT, lexical diversity in an oral interview task and speech rate on an oral retelling of a picture-based narrative. These studies suggest that OEITs tap into the same type of implicit knowledge that is employed in more spontaneous narrative tasks.

* Ghent University, BE

† KU Leuven, Faculty of Arts, Research Unit Linguistics, BE

‡ KU Leuven, Imec research group ITEC, BE

Corresponding author: Kristof Baten (kristof.baten@ugent.be)

Further support for the validity of the OEIT as a measure of implicit knowledge came from Ellis (2008). This study investigated whether OEIT data yield the same developmental stages that have been previously observed in data collected from naturally occurring language use. To identify such previously observed developmental stages, Ellis (2008) appealed to research within the framework of Processability Theory (PT) (Pienemann, 1998). This theory assumes a set of so-called processing procedures which determines how L2 development evolves (i.e., in stages). Importantly, predictions based on PT relate to implicit knowledge, and therefore, the only kind of data that is commonly considered appropriate for testing PT are data of naturally occurring L2 use. However, Ellis (2008) precisely found that his OEIT data of four English grammatical features (possessive –s, since/for, 3rd person –s, question tags) revealed the same developmental stages that were observed in earlier PT research using data of naturally occurring L2 use, both in terms of group means and individual scores. Recently, in a similar study, Baten (2019) replicated previous findings of developmental stages in L2 German. Specifically, in a group of migrant learners, the study uncovered the same three stages that were previously found with regard to L2 German case marking (Diehl et al., 2000; Baten, 2013).

The available research seems to suggest that OEITs can be considered a valid and reliable test to measure L2 production. However, the role of comprehension is rarely taken into account. Indeed, the previous OEIT studies did not always verify whether the participants comprehended the experimental items. This is remarkable, because as Jessop et al. (2007 p. 217) contend, "participants have to first comprehend the stimuli before they can reproduce what they hear." Therefore, the current study will both build on the previous OEIT validation studies and add comprehension as a feature to the task design.

2. Elicited imitation task

OEITs are not new in SLA, but form part of a long research tradition which began in the 1960s and peaked in the 1970s and early 1980s. A number of reviews on the topic indicated that there was a consensus among researchers on the usefulness of applying OEITs. Nevertheless, after its auspicious start the use of OEITs declined because a number of aspects related to its construct validity were unclear and because the field also began to embrace more communicative approaches (see Jessop et al., 2007; Vinther, 2002; Yan et al., 2016). One of the main critiques was that OEITs were believed to involve learners merely in rote repetition (see, McDade et al. 1982). Erlam (2006) demonstrated, however, that specific design features reduce the likelihood of rote repetition, namely the inclusion of a time delay and ungrammatical items. In OEITs with these design features, participants are presented with both grammatical and ungrammatical stimuli, which they either have to reproduce (in case of the grammatical items) or reconstruct (in case of the ungrammatical items) after a short time delay.[1] To realize such a delay between stimulus and response, OEIT studies have employed several techniques: Counting to a given number between 5 and 12 (Baten, 2016), making a true/false judgment about the stimulus sentences (Ellis, 2005, 2008; Erlam, 2006; Bowles, 2011; Spada et al., 2015) and matching the stimulus sentence with the right picture (Baten, 2019). The assumption underlying OEITs of this type is that a particular linguistic rule can be considered acquired if, after the time delay, a learner is able to reproduce target-like structures and reconstruct deviant structures into target-like structures.

This assumption is based on the role of the working memory and its limited capacity for processing information (McLaughlin et al., 1983). In normal L2 processing, meaningful items (i.e., lexical items) are processed before less meaningful or non-meaningful items (e.g., grammatical morphemes) (VanPatten, 2004). The limited capacity of the working memory forces the L2 learner to strategically allocate cognitive resources. Applied to the OEIT, upon hearing the oral stimuli, learners will first process for meaning and only later for form. Furthermore, memory-span research has demonstrated that the memory of the form (with respect to syntax, morphology and lexicon) quickly disappears after a sentence has been understood; the memory of meaning is retained longer (see McDade et al., 1982). The implication is that, after the time delay, the meaning but not the linguistic form will be retrieved from working memory. Instead, the linguistic form (i.e., the syntactic structure and the morphemes used) needs to be put together again, both for grammatical and ungrammatical structures. For this reproduction/reconstruction, the learners will draw on their L2 knowledge. As such, the reproduction/reconstruction is assumed to grant insight in the learner's interlanguage grammar, in that a linguistic rule is probably not acquired when no successful reproductions/reconstructions occur. On the other hand, the linguistic rule must be part of the learners' interlanguage when there are able to successfully reproduce target-like structures as well as reconstruct deviant structures into target-like structures.

Evidence for the reconstructive nature of OEITs was provided in Erlam (2006). This study established a strong, positive correlation between the reproduction of grammatical items and the correction of ungrammatical items. This finding counters earlier claims that OEITs involve rote repetition: If rote repetition had been at work, then an inverse correlation between reproductions and reconstructions would have been found. In the original design of the OEITs used in the 1960s to 1980s, the OEITs most of the time only included grammatical items and involved immediate imitation. In such design, it is right to question the reconstructive nature of OEITs. As Erlam (2006) demonstrated, however, this consideration is no longer relevant when ungrammatical items and a time delay are also included. Therefore, the present study adopts these design features.

In addition, the current study uses the time delay to control for comprehension. This comprehension check is necessary, because it is unclear how a response should be interpreted if there is no information about whether or not the learner has comprehended the stimulus sentence. A number of previous studies have used true/false belief statements (Ellis, 2005, 2008; Erlam, 2006; Bowles, 2011;

Spada et al., 2015) or a picture-matching task (Baten, 2019) during the time delay. Naturally, these tasks aimed to focus participants' attention on the meaning of the stimulus sentences, but they do not really verify if the particular items were comprehended or not. Therefore, by directly measuring comprehension, the current study addressed the following question: Is there construct validity for the OEIT as a measure of developmental stages?

3. Method

3.1. Participants

Data were collected from 36 undergraduates/postgraduates, aged between 18 and 23 (first language Dutch). They were enrolled in a language programme at a Belgian university and all took L2 German courses (linguistics and literature). Eleven students were in their second year of the bachelor's programme, 10 in their third year of the bachelor's, and 15 in the master's programme. Thirteen of the master's students had studied one semester abroad in a German-speaking country. The students showed mixed language proficiency. **Table 1** summarizes the mean scores of a self-rating questionnaire, ranging from 1 (minimal proficiency) to 5 (near-native proficiency).

The mean scores suggest an increasing proficiency from the second year to the final master year. Furthermore, the self-perceived proficiency in the two receptive skills is rated higher than in the two productive skills. These participants did not rate their own speaking skills highly. Judging by the means between 2.9 and 3.5, it shows that the participants rated their own speaking skills in German as average. At the time of data collection, on average, the participants had been learning German as a foreign language for 3.1 (Bachelor 2), 5.0 (Bachelor 3) and 5.6 (Master) years (which includes the time spent learning German prior to the beginning of the respective degree programmes).

3.2. OEIT content

The grammatical content in the present study's OEIT is known to be one of the most problematic areas for L2 learners of German, namely case marking (see Krumm et al., 2010, pp. 518–736). To identify developmental stages with regard to L2 German case marking, the present study will, analogous to Ellis (2008), build on PT (Pienemann, 1998), because it affords a basis for describing such stages. Recent work in PT has shown a surge of interest in understanding the L2 development of case systems: Apart from L2 German (Baten, 2013), L2 Russian (Artoni & Magnani, 2013) and L2 Serbian (DiBiase et al., 2015) have been studied.

Table 1: Mean scores of self-rated proficiency on the four skills.

Group	Listening	Reading	Speaking	Writing
Bachelor 2	3.9	3.8	2.9	3.3
Bachelor 3	4.2	3.9	3.2	3.5
Master	4.2	4.1	3.5	3.4

The findings of these studies are quite similar cross-linguistically. Broadly speaking, they all distinguish three stages. In the first stage, learners rely on basic linguistic means (e.g., canonical word order, prepositions) to indicate grammatical functions (i.e., subject, direct object, indirect object). Regarding case marking, three sub-steps exist at this stage: At first, learners only use nominative case markers on all arguments, then direct mapping occur, and finally positional marking. While direct mapping involves a binary case differentiation between initial nominative arguments and non-initial or post-verbal accusative arguments (i.e., there is no differentiation between accusative and dative), positional marking means that cases are linked to the canonical position of the arguments. In other words, with transitive verbs, the first argument is marked in nominative, the second in accusative; with ditransitive verbs, the first argument is again marked in nominative, the second in dative and the third in accusative (as in (1)).

(1) Der Lehrer gibt dem Jungen den Apfel
The-NOM teacher gives the-DAT boy the -ACC apple
'The teacher gives the boy the apple.'

In the next stage, the distinction between accusative and dative case markers emerges in prepositional phrases. This means that learners are then able to associate certain prepositions with accusative case marking (e.g., *gegen* 'against' in (2)) and certain others with dative case marking (e.g., *mit* 'with' in (2)).

(2) Der Mann fährt mit dem Auto gegen einen Baum
The-NOM mann drives with the-DAT car against a-ACC tree
'The man drives with the car against a tree.'

In the final stage, learners do not necessarily maintain the canonical word order, which entails that case markers are now needed to indicate the grammatical functions. In this stage, learners show target-like use of case markers in utterances with canonical word order (as in (1)), as well as in utterances with non-canonical word order (as in (3)–(4)).

(3) Den Apfel gibt der Lehrer dem Jungen
The-ACC apple gives the-NOM teacher the-DAT boy
'The teacher gives the boy the apple.'

(4) Dem Jungen gibt der Lehrer den Apfel
The-DAT boy gives the-NOM teacher the-ACC apple
'The teacher gives the boy the apple.'

These three stages (i.e., positional > prepositional > functional) are the basis for the content of the OEIT task that will be described next.

3.3. OEIT procedure

For the present study, a web-based OEIT was administered in a computer lab. The OEIT consisted of 48 stimuli sentences: 16 transitive sentences (e.g., *Der Hund verfolgt den Mann*, 'the dog chases the man'), 16 ditransitives (e.g.,

Die Lehrerin schenkt dem Direktor die Blumen, 'the teacher gives the headmaster flowers') and 16 sentences with prepositional phrases (e.g., *Der Mann spaziert durch den Tunnel*, 'the man walks through the tunnel'). Half of the sentences were grammatical, the other half ungrammatical with respect to case morphology. Furthermore, half of the transitive and ditransitive sentences comprised a canonical word order, the other half a non-canonical one that topicalized the objects. For each combination of these parameters, there were four stimuli sentences.[2] **Table 2** summarizes the different types of stimulus sentences.

The stimuli sentences consisted of simple syntactic constructions with the most minimal sentence length possible: The transitive sentences contained five words (determiner – noun – verb – determiner – noun); the ditransitive sentences seven words (determiner – noun – verb – determiner – noun – determiner – noun); and the prepositional phrases six words (determiner – noun – verb – preposition – determiner – noun).[3] As far as the lexicon is concerned, all nouns and verbs belonged to the basic vocabulary knowledge (Level A1–A2 of the Common European Framework of Reference) according to the German vocabulary list *Profile Deutsch* (Glaboniat et al., 2005). These items should be known to university learners of German.

The stimuli sentences were recorded by a female native speaker of German in a sound-proof recording studio. They were presented to the participants in a random order. The presentation of each stimulus sentence consisted of three steps. First, the stimulus was presented aurally through headphones (students wore headsets throughout the entire experiment). In the second step, immediately after the oral stimulus, two pictures were shown, visualizing competing semantic interpretations of the stimulus (see **Figure 1**), one of which was incorrect. The participants were required to choose only one picture. In the third step, while the chosen picture remained on screen, the participants were instructed to repeat the sentence in proper German.[4]

The purpose of the picture-selection task was to achieve a real reproduction/reconstruction, whereby the participants would be delayed in repeating the sentences and draw on their internalized lexicon and grammar while speaking. Additionally, and in contrast to the task design of Erlam (2006) and other OEIT studies, the picture-selection phase was used to measure the comprehension of the stimuli, thereby expanding the utility of the OEIT. The software logged the students' interpretations of the sentences, allowing researchers to determine whether they selected the correct picture. Their speech was recorded with Audacity. Before the experiment started, the participants were familiarized with the procedure by a trial that included four sentences.

3.4. Analysis

After the experiment, the speech response data were transcribed manually and scored in terms of correct (one point) or incorrect (zero points) case reproductions/reconstructions. Regarding the transcription process, it should be noted the speech data were automatically transcribed. A comparison between the manually and automatically transcribed responses by means of Levenshtein distance calculations revealed a remarkable close fit, thus indicating high reliability of the transcriptions (Cornillie et al., 2017). Concerning the scoring, it should further be noted that sentences that were reconstructed in terms of word order (e.g., when participants produced a canonical sentence after hearing a non-canonical one) were removed from further analysis. The data of the picture-matching task were automatically annotated for either correct or incorrect matches. In the analysis of the speech-response data only the responses that corresponded to correct matches are included, because incorrect matches point to comprehension difficulties. Three analyses are reported on this response data. The first is based on mean correctness scores of the group as a whole. The second used mean correctness scores of the individual

Table 2: The different types of stimulus sentences in the OEIT.

Structure	Grammaticality	Canonicity	Targeted case	#
Transitive	Grammatical	Canonical	Accusative	4
		Non-canonical/Topic	Accusative	4
	Ungrammatical	Canonical	Accusative	4
		Non-canonical/Topic	Accusative	4
Ditransitive	Grammatical	Canonical	Dative	4
		Non-canonical/Topic	Dative	4
	Ungrammatical	Canonical	Dative	4
		Non-canonical/Topic	Dative	4
Prepositional	Grammatical		Accusative	4
			Dative	4
	Ungrammatical		Accusative	4
			Dative	4
				48

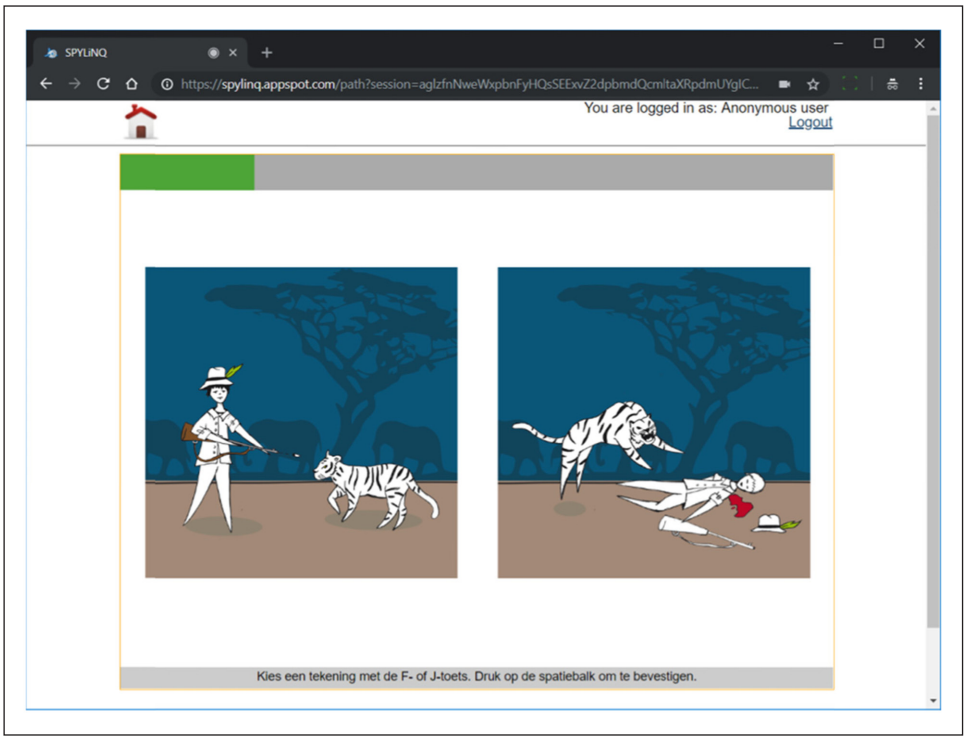

Figure 1: Competing pictures of the ungrammatical stimulus *Der Tiger tötet der Jäger (the-NOM tiger kills the-NOM hunter, 'The tiger kills the hunter').

participants. The third applied the emergence criterion on the data of the individual participants.

The emergence criterion is the criterion for operationalizing acquisition in PT studies. It can be defined as the "point in time corresponding to the first systematic and productive use of a structure" (Pallotti, 2007, p. 366). However, first use does not actually involve an isolated case but is embedded in a number of different contexts. The present study elicited a minimum of four contexts for each case context. Importantly, because individual cases owe their existence to other cases (Jakobson, 1936), a case cannot be acquired independently, but only in opposition to one or more other cases. In concrete terms, this means that evidence of emergence of a stage can only be assumed when the proportion accusative:dative is 1:1 or higher. This analysis is presented in the form of an implicational scale, which presents the binary result whether or not the developmental stage has emerged (i.e., reached this 1:1 proportion, marked by plus or minus). An implicational scale describes the systematic relationship between stages, such that higher stages imply the presence of lower stages, but not vice versa (Håkansson, 2013a).

Before moving on to the results, it is necessary to address our theoretical and methodological stance toward the analyses used in the present study, because accuracy scores and emergence patterns do not measure the same construct. The emergence criterion describes the beginning of the acquisition process. It shows the cut-off point that remains constant: The underlying processing skills to produce certain linguistic structures are either available or not. Conversely, accuracy scores represent the level of mastery of certain linguistic structures. Naturally, the path from emergence to full mastery is dynamic and characterized by fluctuations. In other words, accuracy scores do not evolve linearly, but are highly variable. Several variables, such as intrinsic feature properties, task properties, learning conditions and individual learner characteristics contribute to the variation in accuracy scores. As such, the accuracy scores of an individual learner on a particular test at a given moment reveal the level of difficulty (Housen & Simoens, 2016). In accordance with this distinction between accuracy and emergence, the present study will consistently differentiate between levels of difficulty when referring to accuracy scores and developmental stages when relating to the emergence criterion.

4. Results

4.1. Picture-matching task

Table 3 presents the results of the picture-matching task: The first column lists the different types of the OEIT's stimulus sentences; the two middle columns give the number of correct and incorrect matches; the last column indicates the total number of matches.[5]

The results of the picture-matching task show that the participants have little difficulty in selecting the corresponding picture. In other words, the participants' receptive knowledge is quite high. However, the results clearly indicate that comprehension is still somewhat difficult when case markers are used purely functionally, namely in non-canonical sentences. Topicalized accusative and topicalized dative arguments yield, respectively, 58 and 53 mismatches. This finding is in line with previous research, which has shown that, in comprehension, both beginning and proficient learners mainly rely on linear word order instead of case information to determine the subject and the object of a sentence (Hopp, 2010; Jackson, 2007; VanPatten & Borst, 2012).

In the present study the mismatches only occurred in the grammatical items of the OVS sentences (i.e., TOP_A and TOP_D, (5)). There is no problem with comprehension

in their ungrammatical counterparts (i.e., *TOP_A and *TOP_D, (6)). This disparity of findings can be logically explained by the design of the experiment and more specifically by the competing pictures involved.

(5) dem Mann gibt die Frau den Apfel
the-DAT man gives the-NOM woman the-ACC apple
'The woman gives the man the apple.'

(6) *der Mann gibt die Frau den Apfel
*the-NOM man gives the-NOM woman the-ACC apple
'The woman gives the man the apple.'

In the case of topicalized arguments with correct morphological marking (5), the two pictures visualized competing interpretations, in that both animate arguments can perform the action (i.e., 'to give the apple'; (a) and (b) in **Figure 2**). Conversely, in the case of topicalized arguments with incorrect morphological marking (6), the two pictures did not present a competition, because the object in the picture (i.e., the apple in (a) and the flowers in (c) in **Figure 2**) made the choice quite self-evident. In the experiment, this was necessary because only then would the participants be guided towards reconstructing a non-canonical sentence. More specifically, it aimed for a reconstruction of the morphological incorrect sentence in (6) into the morphological correct sentence in (5). In a scenario with two competing pictures, the sentence in (6) could also be reconstructed into a canonical sentence (*der Mann gibt der Frau den Apfel*, 'the-NOM man gives the-DAT woman the-ACC apple'), but this would have deviated from the aim of this sentence type.

Table 3: Results of the picture-matching task.

Context	Correct matches	Incorrect matches	Total
CAN_A	138	6	144
CAN_D	129	15	144
PP_A	141	3	144
PP_D	141	3	144
TOP_A	86	58	144
TOP_D	91	53	144
*CAN_A	124	3	127
*CAN_D	122	14	136
*PP_A	139	5	144
*PP_D	141	3	144
*TOP_A	141	3	144
*TOP_D	144	0	144
Total	1537	166	1703

Note: CAN: Canonical, PP = Prepositional Phrase, TOP = Argument in Topic position, A = Accusative, D = Dative, * = ungrammatical.

4.2. Production data

For the analysis of the production data only the responses to the stimuli of which the picture-matching was correct were considered (k = 1537). **Table 4** presents the group mean results for positional, prepositional and functional marking, in both grammatical and ungrammatical contexts (the latter indicated by an asterisk).

The results show that the participants performed highly on the grammatical items, reproducing 97% to 99% of the items correctly, on average. This performance indicates that the OEIT successfully forces learners to reproduce sentences. However, the question is whether the OEIT is also a valid reconstructive test. In this regard, the learners corrected 49% of the ungrammatical topicalized items and more than 80% of the ungrammatical canonical and prepositional items. This ability to correct ungrammatical items suggests that the OEIT is reconstructive. According to Erlam (2006, p. 472) additional evidence of the reconstructive nature of the OEIT is provided when there is a significant positive relationship between participants' ability to reproduce grammatical items correctly and their ability to correct ungrammatical items. The present study shows a significant positive correlation (r = 0.62, p < .001). Consequently, it can be assumed that the OEIT in the present study is reconstructive and, as such, a valid measure of learner interlanguage.

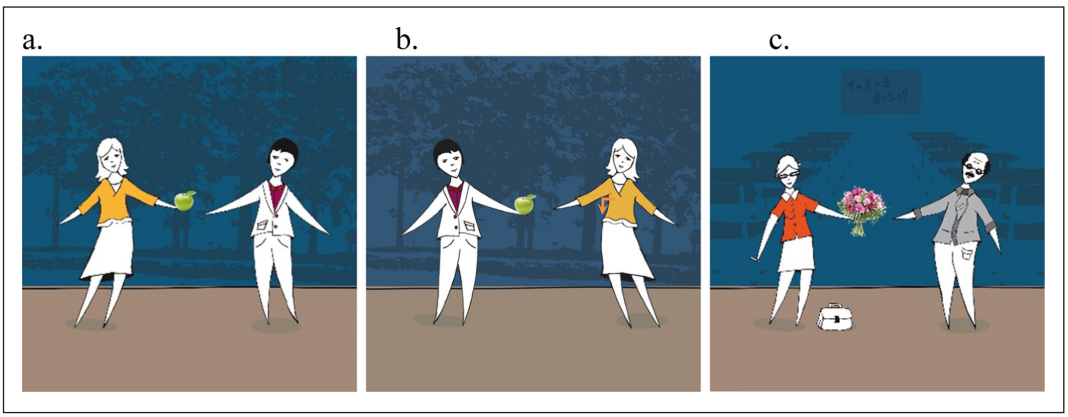

Figure 2: Competing pictures of the stimulus sentence 'dem Man gibt die Frau den Apfel' (pictures a and b)/'*der Mann gibt die Frau den Apfel' (pictures a and c).

Table 4: Descriptive statistics for reproducing grammatical items and correcting ungrammatical items ($k = 36$).

Context	Minimum	Maximum	Mean	SD
Positional (canonical)	.86	1.00	.9777	.05076
Prepositional	.88	1.00	.9931	.02904
Functional (topicalized)	.00	1.00	.9663	.17455
*Positional (canonical)	.00	1.00	.8169	.25668
*Prepositional	.38	1.00	.8403	.17285
*Functional (topicalized)	.00	1.00	.4881	.31493

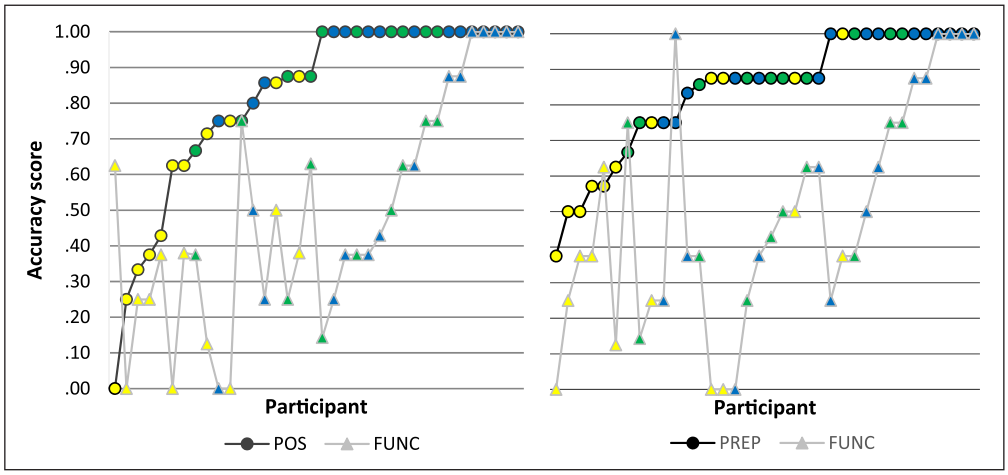

Figure 3: Individual accuracy scores of successful reconstructions (colour-code: yellow = Bachelor 2, green = Bachelor 3, blue = Master).

The present study also examines whether the OEIT can reveal the same developmental stages that were observed in previous research using data of naturally occurring language use. With respect to the three developmental stages of case acquisition in L2 German, a one-way ANOVA reveals that the mean scores of the successful reconstructions are significantly different according to stage (F (2,105) = 21.48, $p < .001$). Post-hoc comparisons show a significant difference between positional and functional marking ($p < .001$), as well as between prepositional marking and functional marking ($p < .001$), but not however, between prepositional and positional marking ($p > .05$). These findings suggest a distinction between two levels of difficulty. However, mean accuracy scores can be misleading, because individual learners do not necessarily follow the same order of difficulty as the sample as a whole. Therefore, the next analysis compares individual accuracy scores.

Figure 3 shows the individual scores of successful reconstructions in the three types of case marking. The left graph presents the scores for positional marking (from lowest to highest) and the corresponding scores for functional marking. The right graph the scores for prepositional marking (from lowest to highest) and the corresponding scores for functional marking.

In most individuals the scores for functional marking are lower than the scores for positional and prepositional marking. This finding indicates that the distinction of two difficulty levels also exists on the level of the individual learner. However, sometimes the scores are equal and in a few cases the scores for functional marking are higher. Equal scores are not problematic, because this means that those students have full mastery in both stages. Reverse scores may seem more problematic. However, given that these reverse scores only occurred in four out of 72 cases, the impression arises that other variables not controlled for in the current investigation may have caused these results. For example, the first learner in the graph on the left was rather unsuccessful in the picture-matching task, and as a result, only one ungrammatical stimulus for positional marking remained for production. In other words, regarding this learner, there are not enough data to interpret the score for positional marking. Regarding the three learners in the graph on the right, it is unclear what may have caused the reverse scores (although it should be noted that the scores are close to similar in two learners). Nevertheless, the main picture that emerges from these results is that positional and prepositional marking can be considered as a difficulty level that is separate from functional marking.

Zooming in on positional and prepositional marking, the group analysis above did not reveal a distinction between the two. **Figure 4** shows the individual scores of successful reconstructions in these two types of case marking, with scores from positional marking presented from lowest to highest.

The scores yielded a complicated pattern: 11 learners scored higher on positional marking than on

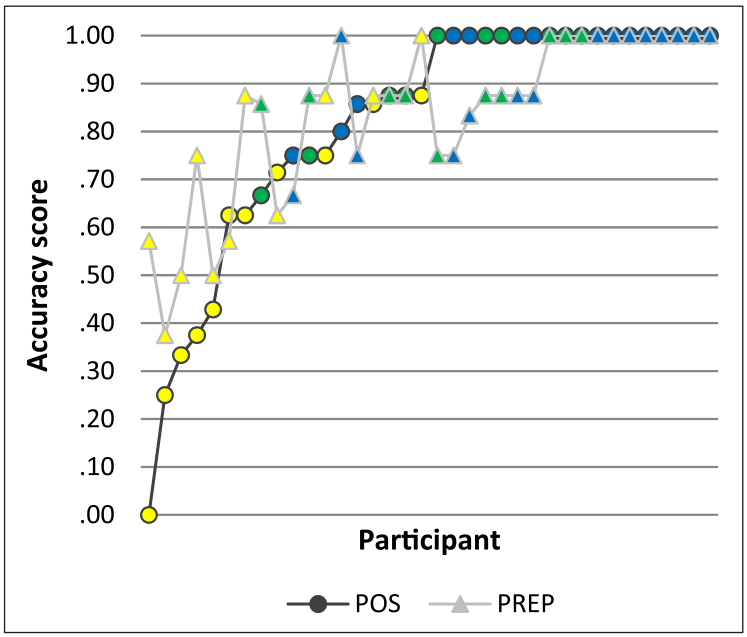

Figure 4: Individual accuracy scores of successful reconstructions (colour-code: yellow = Bachelor 2, green = Bachelor 3, blue = Master).

Table 5: Implicational scaling of the 36 learners' L2 German case production.

Stage	Status of emergence					*n*
*positional	/	(+)	+	+	+	35
*prepositional	+	+	+	+	+	36
*functional	+	+	+	(+)	–	30
n	1	2	23	4	6	

prepositional marking, 12 learners showed the opposite, and 13 learners showed equal scores. Seeing this pattern, it is not surprising that there was no statistical difference in the group findings. In sum, both accuracy analyses (i.e., the analysis of the group as a whole and that of the individual learner) did not reveal whether there was a distinct level of difficulty between positional marking and prepositional marking. In itself, this result is not surprising, as accuracy scores can fluctuate because of various context- or learner-related factors. It is for this reason that research based on PT makes a distinction between variation and development as two separate dimensions of L2 acquisition (Meisel et al., 1981). In this regard, PT researchers argue that structures that are produced quite accurately at a given moment are not necessarily the structures that were acquired early (and vice versa) (Håkansson, 2013b, p. 118). While accuracy scores are a valid measure to reveal variation in L2 proficiency, PT uses another measure (i.e., the emergence criterion) to operationalize L2 development in terms of developmental stages.

Table 5 presents the abridged implicational scale of the results in terms of emergence (for the full version, see the Appendix). A + means that the case opposition has emerged (i.e., a minimum 1:1 proportion of accusative:dative was reached); - means that it has not emerged. A special case is denoted by (+), indicating that successful reconstructions only occurred in one case. The/means that there were not enough data to determine emergence or not. The bottom row shows the number of participants who represent a certain developmental profile, and the column on the right totals the number of learners who have reached the stage at hand.

The results show that the majority of the learners have acquired all stages ($n = 23$). This means that these learners show the ability in all contexts to reconstruct sentences with incorrect nominative marking into sentences with correct accusative and dative marking. Six additional learners can be added, but their ability to reconstruct incorrect marking into correct marking does not always show in both cases (i.e., accusative and dative) in all contexts. In the context of positional marking, for example, two learners reconstruct incorrect nominatives into correct datives, but never into accusatives. In the context of functional marking, two learners are able to reconstruct incorrect nominatives into correct accusatives, but never into datives, and vice versa for two other learners. Finally, in the data of one learner there is evidence regarding the emergence of the prepositional and functional marking stage. However, there is no evidence regarding the emergence of the positional marking stage. This is the learner who was unsuccessful in the picture-matching task, and only one item of positional marking remained (see above, **Figure 3**, left graph). Nevertheless, in all likelihood, this particular learner has also acquired the positional marking stage. Taken together, 30 learners in the present study have acquired the properties of German case marking.

However, in six learners the stage of functional case marking has not emerged. In other words, these learners were not able to reconstruct the incorrect use of the nominative on topicalized objects into the correct use of either accusative or dative. This pattern of 30 learners reaching three stages and six learners reaching two stages

suggests an implicational relationship: Learners who have acquired functional marking have also acquired the other two types of case marking. The reverse is clearly not true. An additional implicational relationship between the stages of positional and prepositional marking could not be uncovered, because both stages have emerged among all learners. In other words, the implicational scaling lends partial support to the developmental stages derived from previous PT research. Thus, the OEIT data are comparable to the data of naturally occurring language use that have been collected in PT studies (e.g., Baten, 2013). However, further data collection, particularly with beginning learners, will be needed to provide more robust evidence of the validity of the OEIT.

Finally, it is interesting to relate the results to the participants' programme level. In **Figures 3** and **4** the individual scores of the successful reconstructions were colour-coded: Yellow represents the participants from Bachelor 2, green from Bachelor 3 and blue from the Master year. It can be seen that Bachelor-2 students mainly appear on the left side (i.e., lower scores) and Master students mainly on the right side (i.e., higher scores); in the middle, students from all programme levels can be found. Regarding the developmental stages, the six learners who did not reach the highest stage (i.e., functional case marking) involve four Bachelor-2 students and one student from each of the other two programme levels. While these observations seem to suggest some kind of relationship between level and score or stage, the distribution in the middle and the fact that participants from all levels show the ability to reach the highest stage of case marking indicate that development in terms of individual mean scores and in terms of developmental stages cannot be predicted by the programme level. This is, of course, not a surprise, as levels are based on several requirements within the programme that are not directly related to the language development of specific structures.

5. Discussion

The results suggest that the web-based OEIT described in this study can be a valid measure of L2 development, both in terms of order of difficulty and in terms of developmental stages. This validity, first of all, pertains to the reconstructive nature of the test. The fact that the L2 learners corrected ungrammatical items is evidence that the test is reconstructive. Furthermore, the positive correlation between their ability to correct ungrammatical items and their ability to reproduce grammatical items adds to this evidence. These findings are similar to Erlam (2006). It should be noted, though, that the OEIT in Erlam (2006) included multiple structures, while the OEIT in the present study involved a single target structure (just as in Spada et al., 2015, which focussed on the passive). Despite this difference in the number of structures, the findings of the present study corroborate the earlier evidence that it is justified to use ungrammatical items in OEITs. Indeed, the ability or the failure to correct ungrammatical sentences can clearly be seen as evidence of the extent to which structures have been internalized.

However, different from the previous OEIT validation studies, the present study integrated comprehension into the task design. Methodologically, this seems to be an essential step, because in order to actually measure L2 production abilities, comprehension difficulties need to be reckoned with first. In the present study, 166 responses were excluded from further analysis, because they corresponded to incorrect comprehension matches. If these responses had been included, it would have been impossible to disentangle whether production (in)abilities or (in)adequate functioning of comprehension abilities were assessed. Clearly, this issue awaits further investigation, for example, by comparative analyses of responses that are either controlled for comprehension or not.

Returning to the analyses of the production data, the results on the ungrammatical items in the test produced some interesting insights with respect to L2 difficulty levels. The individual accuracy scores established that functional case marking was consistently more difficult than positional and prepositional marking. This indicates that something intrinsic to the feature determines its L2 difficulty. In all likelihood this is the non-linearity between the functional structure and the constituent structure. Obviously, this non-linearity does not apply to positional and prepositional marking, which probably explains why the study could not observe any consistent contrast in terms of L2 difficulty between these two types of case marking. Actually, for some learners positional marking proved more difficult than prepositional marking, while for other learners it was the other way around. This finding squares well with Housen and Simoens' (2016) taxonomic framework of L2 difficulty, which in addition to intrinsic properties, also includes context-related and learner-related characteristics as determinants of L2 difficulty. It is likely that individual learner differences and perhaps contextual differences have contributed in different ways to the mixed pattern of the accuracy scores of positional and prepositional marking. Interestingly, this opens perspectives for the OEIT as a tool to measure L2 difficulty as a dependent variable. For example, the OEIT could be employed to examine the possible effects of, for example, different types of instruction (context-related) or various levels of cognitive abilities (learner-related) on L2 difficulty.

Nevertheless, in this study we were mainly interested in L2 development, in terms of developmental stages. The results on the ungrammatical items also produced interesting insights in this respect. The implicational analysis revealed that functional case marking is a developmental stage that is separate from positional and prepositional case marking. The late emergence of functional case marking is a recurrent finding in PT (Artoni & Magnani, 2013; Baten, 2013; DiBiase et al., 2015). As such, the present study corroborates Ellis (2008), which demonstrated that the OEIT is capable of determining the stage of development that learners have reached regarding specific linguistic features. Interestingly, Ellis (2008) investigated four grammatical features that are independent from each other and located at different PT stages, while the present study involved a single grammatical feature that appears at

different stages. In other words, the OEIT has the ability to describe a learner's developmental stage of both multiple and single target structures.

However, despite the promising results, the OEIT in the present study failed to uncover a stage distinction between positional and prepositional case marking. This absence of evidence should not be considered as an invalidation of the OEIT. We rather believe that it is an indication of the advanced level of L2 proficiency of the participants in the study. By way of comparison, the OEIT data of recently arrived migrant learners in Germany, who are at the beginning of their acquisition process, clearly showed that positional and prepositional case marking are two distinct developmental stages (Baten, 2019). In other words, if we were to replicate this study with beginning-level learners, it is expected that positional and prepositional marking would surface as separate stages.

The advantages of the OEIT are considerable. Perhaps the most important advantage is that it can elicit any structure. The learners in this study were pushed to produce higher-staged structures, which they would otherwise not produce, or only minimally, in naturally occurring L2 speech. Recall that every learner had to reconstruct eight non-canonical sentences. Taking all learners and ungrammatical items together, and excluding the mismatches on the comprehension task, the non-canonical sentences which had to be reconstructed, represented 35% of the corpus. This number is significantly higher than the number of occurrences that is usually obtained from tasks that elicit free speech. Another advantage of the online OEIT is that it is practical. It can be administered in a relatively short timeframe and with a group of learners in a computer classroom at the same time. This is similar to Spada et al. (2015), which utilized PowerPoint to automatize the OEIT procedure, but different from the other OEIT studies, which administered the task manually, in a one-on-one design (i.e., the test administrator and the test taker) and with the stimuli sentences presented either orally (e.g., Ellis, 2005, 2008; Bowles, 2011; Baten, 2019) or through audio cassette (Erlam, 2006). The online OEIT also has the advantage that it is much less labour-intensive to transcribe and annotate the responses of an OEIT than transcribing and annotating free speech. For example, it is possible to automatize the transcription process (Cornillie et al., 2017).

Finally, another potential advantage of the online OEIT that could be exploited further, particularly for research purposes, relates to timing. It has been pointed out that timing is a feature in the test design that may impact learner performance (Erlam, 2006; Spada et al., 2015; Kim & Nam, 2017). When learners are not time-pressured, they may take their time to produce their response, which could result in accessing their explicit knowledge to monitor their response in terms of accuracy of form. Offline OEITs do not usually strictly control for time latencies (such as the OEITs in Erlam, 2006; Ellis, 2008; Bowles, 2011), but online OEITs can. Spada et al. (2015), for example, controlled for time, in that participants had eight seconds to repeat the stimulus sentence. In the present study, the continuation of the task was self-paced. However, the time log information reveals that the participants, on average, needed 15 seconds from the start of the stimulus sentence to the end of the response. Assuming that half of that time was spent on the comprehension part and half on the production part of the OEIT, the participants in the present study produced their responses in a comparable number of seconds as in Spada et al. (2015).[6] Future research could make use of the logged time information or set different time restrictions in order to examine whether and to what extent performance patterns vary according to the time spent or the time given. Research of this kind would greatly contribute to our understanding of what type of knowledge is measured by what type of OEIT designs.

6. Conclusion

The present study aimed to investigate whether an online OEIT has validity as a measure of L2 development. In particular, the study examined whether the OEIT would yield the same developmental stages that had been found in previous research using naturally occurring L2 data. The results revealed three main findings. First, the OEIT is reconstructive; second, it is capable of uncovering levels of L2 difficulty; and third, it shows stages of development. With these findings, the current investigation confirms earlier claims that it is justified to use ungrammatical items (Erlam, 2006) and it corroborates previous evidence on the validity of experimental data in showing both levels of difficulty and developmental stages (Ellis, 2008). Moreover, the present study indicates that the OEIT is sensitive enough to measure the development of a single grammatical structure. Importantly, the present study also shows that it proves useful to integrate a comprehension check in the task design of OEITs. Naturally, further research is needed, not only to investigate to what extent OEITs measure implicit knowledge (Erlam, 2006; Ellis, 2008; Bowles, 2011; Spada et al., 2015), but also to examine the scope of what can be tested by OEITs and how it can be tested and to explore potential added advantages.

7. Appendix

Appendix 1. Implicational scaling of the 36 learners' L2 German case production.

		Participants																																					
		15	**2**	**3**	**4**	**6**	**9**	**12**	**13**	**16**	**18**	**19**	**21**	**22**	**23**	**24**	**25**	**27**	**28**	**29**	**30**	**32**	**34**	**35**	**36**	**7**	**20**	**10**	**11**	**14**	**33**	**1**	**5**	**8**	**17**	**26**	**31**
Stages	*POS	/	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	(+)	(+)	+	+	+	+	+	+	+	+	+	+
	*PREP	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	*FUNC	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	(+)	(+)	(+)	(+)	-	-	-	-	-	-

Notes

[1] To avoid (terminological) misunderstandings, the present article consistently differentiates between the terms 'reproduction' and 'reconstruction'. The former denotes the participants' responses to grammatical structures, the latter their responses to ungrammatical structures.

[2] The material will be made available through the IRIS database (www.iris-database.org).

[3] Previous studies found that varying sentence lengths determine participant performance (Gaillard & Tremblay, 2016; Kim & Nam, 2017). In the present study, this variable was not considered, because the OEIT only used the lowest length possible.

[4] We are well aware that the instructions that go with the test may influence the test takers' performance (see, Erlam, 2006). We used the phrasing "repeat in proper German" instead of "repeat in correct German" in an attempt to avoid the possible focus on form. Nevertheless, the issue remains moot and more research on the impact of instructional phrasing is necessary.

[5] The total of matches should equal 144 (four sentences multiplied by 36 participants). However, two sentence types (*CAN_A and *CAN_D) have a lower number of matches. The explanation for this is that a small number of participants re-interpreted stimuli that were intended as canonical sentences with morphologically incorrect marking (*der Hund verfolgt *der Mann*, 'The-NOM dog chases the-NOM man') as ungrammatical non-canonical sentences with topicalized arguments. Therefore, these specific stimulus sentences were removed from further analysis.

[6] Unfortunately, time data for each item were only logged of the onset of the stimulus sentence and the completion of the response. Thus, it is possible that less than half of the time was spent on the comprehension section. Clearly, it would be useful for future research with OEITs to include detailed time data.

References

Abrahamsson, N. (2013). Developmental sequences. In P. Robinson (Ed.), *Routledge Encyclopedia of second language acquisition* (pp. 173–177). London: Routledge.

Artoni, D., & **Magnani, M.** (2013). The development of case in L2 Russian. In M. Butt, & T. Holloway King (Eds.), *Proceedings of the LFG13 Conference* (pp. 69–89). Stanford, CA: CSLI Publications.

Baten, K. (2013). *The acquisition of the German case system by foreign language learners*. Amsterdam: Benjamins. DOI: https://doi.org/10.1075/palart.2

Baten, K. (2016). A study on explicit instruction and its relation to knowing/using linguistic forms and individual learner readiness. In S. Liszka, P. Leclercq, M. Tellier, & D. Véronique (Eds.), *EUROSLA Yearbook 16* (pp. 116–143). Amsterdam: Benjamins. DOI: https://doi.org/10.1075/eurosla.16.05bat

Baten, K. (2019). Teaching the German case system: A comparison of two approaches to the study of learner readiness. In A. Lenzing, H. Nicholas, & J. Roos (Eds.), *Widening Contexts for Processability Theory: Theories and issues* (pp. 301–326). Amsterdam: Benjamins.

Bowles, M. (2011). Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute? *Studies in Second Language Acquisition*, *33*, 247–271. DOI: https://doi.org/10.1017/S0272263110000756

Cornillie, F., Baten, K., & **De Hertog, D.** (2017). The potential of elicited imitation for oral output practice in German L2. In K. Borthwick, L. Bradley, & S. Thouësny (Eds.), CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017 (pp. 86–91). Research-publishing.net. DOI: https://doi.org/10.14705/rpnet.2017.eurocall2017.694

DiBiase, B., Bettoni, C., & **Medojevic, L.** (2015). The development of case in a bilingual context: Serbian in Australia. In C. Bettoni, & B. Di Biase (Eds.), *Grammatical development in second languages: Exploring the boundaries of Processability Theory* (pp. 195–212). Amsterdam: The European Second Language Association.

Diehl, E., Christen, H., Leuenberger, S., Pelvat, I., & **Studer, T.** (2000). *Grammatikunterricht: alles für der Katz? Untersuchungen zum Zweitsprachenerwerb Deutsch*. Tübingen: Niemeyer.

Dörnyei, Z. (2009). *The Psychology of SLA*. Oxford: Oxford University Press.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172. DOI: https://doi.org/10.1017/S0272263105050096

Ellis, R. (2008). Investigating grammatical difficulty in second language learning: Implications for second language acquisition research and language testing. *International Journal of Applied Linguistics*, *18*, 4–22. DOI: https://doi.org/10.1111/j.1473-4192.2008.00184.x

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*, 464–491. DOI: https://doi.org/10.1093/applin/aml001

Gaillard, S., & **Tremblay, A.** (2016). Linguistic proficiency assessment in second language acquisition research: the elicited imitation task. *Language Learning*, *66*, 419–447. DOI: https://doi.org/10.1111/lang.12157

Glaboniat, M., Müller, M., Rusch, P., Schmitz, H., & **Wertenschlag, L.** (2005). *Profile Deutsch: Gemeinsamer europäischer Referenzrahmen*. Berlin: Langenscheid.

Håkansson, G. (2013a). Implicational scaling. In P. Robinson (Ed.), *The Routledge Encyclopedia of SLA* (pp. 293–294). London: Routledge.

Håkansson, G. (2013b). Processability theory. Explaining developmental sequences. In M. del P. García Mayo, M. Junkal Gutierrez Mangado, & M. Martínez Adrián (Eds.), *Contemporary approaches to SLA* (pp. 111–127). Amsterdam: Benjamins.

Hopp, H. (2010). Ultimate attainment in L2 inflection: Performance similarities between non-native and native speakers. *Lingua*, *120*, 901–931. DOI: https://doi.org/10.1016/j.lingua.2009.06.004

Housen, A., & **Simoens, H.** (2016). Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition*, *38*, 163–175. DOI: https://doi.org/10.1017/S0272263116000176

Jackson, C. (2007). The use and non-use of semantic information, word order, and case markings during comprehension by L2 learners of German. *Modern Language Journal*, *91*, 418–432. DOI: https://doi.org/10.1111/j.1540-4781.2007.00588.x

Jakobson, R. (1936 [1971]). Beitrag zur allgemeinen Kasuslehre. In R. Jakobson (Ed.), *Selected Writings II* (pp. 23–71). The Hague: Mouton.

Jessop, L., Suzuki, W., & **Tomita, Y.** (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review*, *64*, 215–238. DOI: https://doi.org/10.3138/cmlr.64.1.215

Kim, J., & **Nam, H.** (2017). Measures of implicit knowledge revisited: Processing modes, time pressure, and modality. *Studies in Second Language Acquisition*, *39*, 431–457. DOI: https://doi.org/10.1017/S0272263115000510

Krumm, H. J., Fandrych, Ch., Hufeisen, B., & **Riemer, C.** (2010). Handbuch Deutsch als Fremd- und Zweitsprache. Berlin: De Gruyer. DOI: https://doi.org/10.1515/9783110240245

McDade, H., Simpson, M., & **Lamb, D.** (1982). The use of elicited imitation as a measure of expressive grammar: A question of validity. *Journal of Speech and Hearing Disorders*, *47*, 19–24. DOI: https://doi.org/10.1044/jshd.4701.19

McLaughlin, B., Rossman, T., & **McLeod, B.** (1983). Second language learning: An information-processing perspective. *Language Learning*, *33*, 135–158. DOI: https://doi.org/10.1111/j.1467-1770.1983.tb00532.x

Meisel, J., Clahsen, H., & **Pienemann, M.** (1981). On determining developmental stages in natural second language acquisition. *Studies in Studies in Second Language Acquisition*, *3*, 109–135. DOI: https://doi.org/10.1017/S0272263100004137

Pallotti, G. (2007). An operational definition of the emergence criterion. *Applied Linguistics*, *28*, 361–382. DOI: https://doi.org/10.1093/applin/amm018

Pienemann, M. (1998). *Language processing and second language development: Processability theory*. Amsterdam: Benjamins. DOI: https://doi.org/10.1075/sibil.15

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*, 595–626. DOI: https://doi.org/10.1111/lang.12010

Spada, N., Shiu, J. L. J., & **Tomita, Y.** (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning*, *65*, 723–751. DOI: https://doi.org/10.1111/lang.12129

Tracy-Ventura, N., McManus, K., Norris, J. M., & **Ortega, L.** (2014). 'Repeat as much as you can': Elicited imitation as a measure of oral proficiency in L2. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 Proficiency: Perspectives from SLA* (pp. 143–166). Bristol: Multilingual Matters. DOI: https://doi.org/10.21832/9781783092291-011

VanPatten, B. (2004). Input processing in second language acquisition. In B. VanPatten (Ed.), *Processing instruction. Theory, research, and commentary* (pp. 5–31). Mahwah: Erlbaum. DOI: https://doi.org/10.4324/9781410610195

VanPatten, B., & **Borst, B.** (2012). The role of explicit information and grammatical sensitivity in processing instruction: nominative-accusative case marking and word order in German L2. *Foreign Language Annals*, *45*, 92–109. DOI: https://doi.org/10.1111/j.1944-9720.2012.01169.x

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*, 54–73. DOI: https://doi.org/10.1111/1473-4192.00024

Yan, X., Maeda, Y., Lv, J., & **Ginther, A.** (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*, 497–528. DOI: https://doi.org/10.1177/0265532215594643

How to cite this article: Baten, K., and Cornillie, F. (2019). Elicited imitation as a window into developmental stages. *Journal of the European Second Language Association*, 3(1), 23–34. DOI: https://doi.org/10.22599/jesla.56

Submitted: 21 December 2018 **Accepted:** 03 July 2019 **Published:** 18 July 2019