LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

Working Memory and the Observed Effectiveness of Recasts on Different L2 Outcome Measures

Andrea Révész

Lancaster University

This study examined whether the observed effectiveness of recasts is influenced by the type of outcome measure used and whether different aspects of working memory are differentially associated with learners' performance on the various outcome measures. The participants were 90 learners of English as a foreign language, who were randomly assigned to a recast, a nonrecast, and a control group. A pretest–posttest–delayed posttest design was employed to detect any improvement in the learners' knowledge of one usage of the English past progressive construction. Many-facet Rasch measurement and correlational analyses yielded two main findings. First, recasts generated the greatest gains on an oral production test, lesser gains on a written production test, and the least gains on a written grammaticality judgment test. Second, in the recast group, participants with higher reading spans achieved more development on the written tests, while those with higher digit and nonword spans showed greater improvement on the oral test. For the nonrecast group, no association was found between the working memory and developmental measures.

Keywords recasts, working memory, negative feedback, phonological short-term memory, executive control, procedural knowledge, declarative knowledge

Introduction

The role of recasts has been the object of much interest among second language acquisition (SLA) researchers in recent years. In a variety of contexts

I would like to thank ZhaoHong Han and James Purpura for their insightful comments on the research on which this article is partly based. I am also grateful to Judit Kormos, Rebecca Sachs, and the anonymous reviewers for their helpful suggestions on this article. Any errors, of course, are my own. This research was supported in part by the international research foundation for English language education (TIRF) and the Spencer foundation.

Correspondence concerning this article should be sent to Andrea Révész, Department of Linguistics and English Language, County South C70, Lancaster University, LA1 4YT, UK. Internet: a.revesz@lancaster.ac.uk

DOI: 10.1111/j.1467-9922.2011.00690.x

employing diverse research methods, a large number of studies have set out to determine the utility and/or efficacy of recasts in promoting second language (L2) development. As generally described, recasts involve a teacher's or other interlocutor's reformulation of a learner's utterance by altering one or more errors in it while retaining its semantic content. The example below, taken from data collected for the current study, illustrates a recast preceded by a nontargetlike learner utterance:

(1) Learner: And I saw a boy next to the bar. I think he was with his girlfriend. *They talking to each other.*

Recast: *They were talking to each other.*

Although their efficacy in the past has been questioned (e.g., Lyster & Ranta, 1997), a beneficial role for recasts in L2 learning is by now well established. According to recent meta-analyses of research on interaction (Mackey & Goo, 2007), corrective feedback (Li, 2010), and classroom oral feedback (Lyster & Saito, 2010), empirical studies conducted in different settings and with different learner populations collectively indicate that recasts can facilitate L2 learning outcomes.

The confirmation of a link between recasts and L2 development has led to a change in the focus of cognitive-interactionist research in this domain. Instead of addressing the question of whether recasts contribute to SLA, researchers have turned increased attention to the issue of how they promote L2 learning (Mackey, 2007). One of the key areas of current interest is the nature of the relationships among recasts, learner-internal and -external factors, and linguistic development. A number of factors have been identified as mediating the influence of recasts on L2 learning outcomes, including developmental readiness or proficiency (e.g., Ammar & Spada, 2006; Long, Inagaki, & Ortega, 1998; Mackey & Philp, 1998; Philp, 2003); context of learning (e.g., R. Ellis, Basturkmen, & Loewen, 2001; Lyster & Mori, 2006; Sheen, 2004); type of linguistic target (e.g., R. Ellis, 2007; Jeon, 2007; Long, 2007); characteristics of recasts (Kim & Han, 2007; Loewen & Philp, 2006; Lyster, 1998; Sheen, 2006); task complexity (Nuevo, 2006; Révész & Han, 2006; Révész, 2009; Révész, Sachs, & Mackey, 2011); and individual differences in cognitive variables, such as working memory (e.g., Mackey, Adams, Stafford, & Winke, 2010; Mackey, Philp, Egi, Fujii, & Tatsumi, 2002; Sagarra, 2007a; Trofimovich, Ammar, & Gatbonton, 2007).

Some empirical studies have also shown that the observed effectiveness of recasts may be affected by the type of outcome measure used, that is, the

assessment tasks employed as dependent variables, suggesting that recasts may promote different sorts of L2 knowledge to a lesser or stronger degree (e.g., R. Ellis, Loewen, & Erlam, 2006; R. Ellis, 2007; Loewen & Nabei, 2007; Révész & Han, 2006). The present study intends to contribute to the existing literature by further exploring how the type of outcome measure employed might impact on the results of recast studies. The novelty of this research lies in the fact that it also examines whether working memory, as measured by digit span (DS), nonword span (NWS), and reading span (RS) tests, mediates the extent to which any effects of recasts are demonstrated on various outcome measures.

Theoretical Background

Much of the keen interest in recasts has stemmed from the observation that they are by far the most frequently used feedback technique in L2 classrooms (Lyster & Ranta, 1997; Mackey, Gass, & McDonough, 2000; Sheen, 2004). Perhaps one reason for this is that recasts are unlikely to disrupt the flow of communication between the teacher and a learner (Long, 2007). Recasts seek to draw learners' attention to form-meaning mappings unobtrusively, without overt reference to rules and forms, and they thereby minimize the interruption of the pedagogic intervention on processing language for meaning.

Besides their high frequency, researchers' attention has been drawn to recasts as a result of their unique psycholinguistic characteristics. For instance, Long (2007) argues that, because recasts are reformulations of the learners' own utterances, learners are apt to comprehend their content, thus allowing more processing resources to be allocated to form and to form–meaning connections. In addition, given that recasts and the erroneous learner utterances are juxtaposed, they may prompt learners to notice the gap (Schmidt, 1990), and subsequently make cognitive comparisons (Doughty, 2001; R. Ellis, 1995), between their own incorrect form and the targetlike form. Another potential benefit is that recasts may push learners to modify their output and thereby contribute to automatizing and/or expanding their linguistic knowledge (McDonough, 2005; Swain, 1995). As Kormos (2006, p. 135) explains, if an error is detected and an error-free solution is rehearsed in short-term memory (as is the case when learners immediately repair their output), the resulting long-term memory (LTM) trace may assist in the proceduralization of declarative knowledge and in generating memorized solutions.

A third reason why L2 researchers have been interested in recasts has to do with theoretical issues surrounding the role of negative feedback in SLA. A number of scholars maintain that negative feedback, defined as information

indicating that a learner utterance is ungrammatical in the L2 (e.g., by the means of recasts), plays a limited role in L2 learning. According to some Universal Grammar researchers (Schwartz, 1993), for instance, exposure to negative feedback can merely impact on performance via learned linguistic behavior and cannot foster a change in underlying competence. In a similar vein, Krashen (1985) argued that, although information obtained from negative feedback can lead to improved monitoring behavior, any positive effects of such feedback will not transfer to spontaneous, real-world language use. However, consistent with cognitive accounts of SLA such as connectionism (N. Ellis, 2005) and skill acquisition theory (DeKeyser, 2007a), there is growing empirical evidence suggesting that recasts can facilitate improved performance on a range of outcome measures, including not only tests tapping metalinguistic knowledge of the L2 grammar (e.g., R. Ellis et al., 2006; Loewen & Nabei, 2007), but also tasks involving the ability to use the L2 fluently and spontaneously in communication (e.g., Long et al., 1998; Mackey & Philp, 1998; Révész & Han, 2006). Nonetheless, it remains relatively unexplored whether recasts differentially affect changes on outcome measures that draw differentially on the application of various types of L2 knowledge—an issue the present study set out to investigate.

Recasts, L2 Knowledge, and Different Types of Outcome Measures

There are a number of conceptualizations of L2 knowledge (DeKeyser, 2009) that may inform the interpretations of results obtained on different types of outcome measures in recast studies. Of particular relevance to the present research is the distinction between declarative and procedural knowledge. The declarative–procedural dichotomy is related to SLA theories that regard adult language learning as similar to the acquisition of other complex cognitive skills (e.g., learning to drive or play the piano). In most models of skill acquisition, learning progresses in three consecutive stages (DeKeyser, 2007a, b). First, learners acquire factual information (e.g., L2 rules) through verbal explanation and/or by observing and analyzing the behavior of others engaged in the target skill. The resulting declarative knowledge is conscious and generalizable, but because the processing costs of retrieving information from declarative memory are relatively high, performance utilizing declarative knowledge tends to be slow. The second step involves the transformation of declarative knowledge (knowledge *that*) into procedural knowledge (knowledge *how*) by the process of proceduralization. Unlike declarative knowledge that provides a routine for piecing together bits of information in working memory, procedural knowledge

consists of ready-made chunks that can be accessed directly from procedural memory. As a result, procedural knowledge enables faster and more efficient performance, although with the disadvantage of being highly specific and hard to transfer. For example, there seems to be limited transfer between seemingly parallel production and comprehension skills such as writing and reading, and speaking and listening (DeKeyser, 1997). In the last stage, procedural knowledge is automatized via a large amount of practice, leading to the final outcome of automatic procedural knowledge, which allows for fluent, spontaneous, and effortless performance.

Feedback can play a useful role at all three stages of skill acquisition, from facilitating the learning of declarative knowledge to promoting proceduralization and the automatization of procedural knowledge (Leeman, 2007). Some L2 researchers, however, have argued that, depending on learners' stages of acquisition, various types of negative feedback might be differentially effective. Lyster (2004), for instance, explains that prompts, which include feedback techniques such as clarification requests, repetitions, metalinguistic clues, and elicitation, are more likely to "assist learners in the transition of declarative to procedural knowledge" (p. 406), given that they provide learners with opportunities to practice by encouraging modified output. Recasts, on the other hand, owing to the fact that they potentially afford both negative and positive evidence, may help learners to gain new declarative knowledge. Not all L2 researchers, however, hold the position that recasts are likely to promote the encoding of novel knowledge representations. Han (2002), for example, speculates that recasts, constituting an implicit and nonelaborate form of feedback, may "act more favourably on linguistic forms that are in the process of being proceduralized than on forms that are at the onset of developing knowledge" (p. 552). While a number of studies (e.g., Doughty & Varela, 1998; Han, 2002) demonstrate that recasts can enhance L2 development when learners have prior knowledge of a target construction (declarative and/or procedural), the little empirical research (e.g., Long et al., 1998) which has to date investigated the effects of recasts on the acquisition of novel linguistic features has yielded mixed findings. Evidently, further research is needed to clarify whether recasts have the potential to facilitate the creation of new knowledge representations.

A related and unresolved issue concerns the extent to which outcome measures may shape the results of recast research. As mentioned above, there is growing empirical evidence suggesting that recasts can promote improved performance on various types of outcome measures used as dependent variables. Nonetheless, only a small number of studies exist that directly investigate whether recasts appear to lead to more or less substantial change depending on

the outcome measure employed (e.g., R. Ellis, 2007; R. Ellis et al., 2006; Han, 2002; Loewen & Nabei, 2007; Révész & Han, 2006; Sheen, 2007). Rather than incorporating multiple outcome measures tapping different behaviors, most recast studies have utilized a single type of oral communication task to tap changes in learners' interlanguage (e.g., Leeman, 2007; Long et al., 1998; Mackey & Philp, 1998; Mackey & McDonough, 2006). However, as Norris and Ortega (2003) point out, failing to collect a multiplicity of behavioral observations may result in interpretations that are "based on a lack of evidence, as opposed to evidence for the lack of emergence" (p. 733), because a form or structure may emerge in different contexts for different learners. The need for a multiplicity of data sources is also emphasized by R. Ellis (2004; R. Ellis et al., 2006) and Doughty (2003), who argue that the potentially different effects of instruction on distinct types of linguistic knowledge may be overlooked unless studies include a variety of carefully selected measures. In general, there appears to be a growing recognition that various types of assessments rather than a single instrument need to be incorporated into effects-of-instruction studies so that the treatment effects, or the lack thereof, on different types of L2 knowledge are detected.

Although they are small in number, it is worth considering the design and results of recast studies that have utilized multiple developmental measures. Such studies have typically employed both measures that are likely to encourage the application of declarative knowledge, requiring metalinguistic judgments under no time pressure, and measures likely to draw more heavily on procedural knowledge, eliciting fast and/or fluent spontaneous performance.[1] The instruments prone to tap declarative knowledge have included untimed grammaticality judgment tasks (GJTs; Loewen & Nabei, 2007; R. Ellis, 2007; R. Ellis et al., 2006), error correction tasks (Sheen, 2007), and metalinguistic tests (R. Ellis, 2007), whereas among the instruments likely to require procedural knowledge have been elicited imitation (R. Ellis, 2007; R. Ellis et al., 2006), oral production (Loewen & Nabei, 2007; Révész & Han, 2006), speeded dictation (Sheen, 2007), and timed grammaticality judgment tests (Loewen & Nabei, 2007). The results, overall, indicate that, while recasts can positively affect performance on both types of measures, instruments that draw more substantially on procedural knowledge are more likely to show greater evidence of change. Note that similar results were obtained in recent meta-analyses of feedback studies (Li, 2010; Lyster & Saito, 2010; Mackey & Goo, 2007).

According to Mackey and Goo's (2007) analysis, while open-ended prompted production measures (e.g., oral communicative tasks) resulted in

medium-to-large effect sizes, prompted response tasks (e.g., untimed grammaticality judgments) indicated small effect sizes on immediate posttests. Likewise, Li (2010) and Lyster and Saito (2010) identified larger effect sizes for free constructed-response measures than for constructed-response measures and metalinguistic judgments in terms of Norris and Ortega's (2000) classification of outcomes measures, although the differences did not reach significance in Li's meta-analysis. As the authors of two of the meta-analyses point out (Mackey & Goo, Lyster & Saito) more empirical studies are necessary to make claims about the role of outcome measures in assessing the effects of interactional treatments. One aim of the present study is to contribute to filling this gap.

Working Memory and L2 Acquisition

Another goal of this study is to examine whether working memory mediates the relationship between recasts and learner performance on different types of outcome measures. Working memory is defined by Baddeley (2003) as "the temporary storage and manipulation of information that is assumed to be necessary for a wide range of complex cognitive activities" (p. 189). Miyake and Friedman (1998, p. 339) have proposed that working memory "may be one (if not) the central component of . . . language aptitude" because it allows for the kind of serial processing of information that is required by both comprehending and producing language. Sawyer and Ranta (2001) likewise relate working memory to language aptitude, arguing that it is implicated in attentional processes that are crucial in SLA (e.g., noticing). In their view, working memory also serves as a temporary cognitive arena where other components of aptitude, such as phonemic coding ability, grammatical sensitivity, and memory ability, are integrated. With specific reference to recasts, Robinson (2005), too, points out that individual differences in working memory, in combination with other cognitive abilities, are likely to be important determinants of how much is learned from recasts. In particular, he explains that phonological working memory capacity and speed help enable learners to maintain the incorrect learner utterance and the targetlike recast "in working memory long enough for the analytic 'cognitive comparison' . . . to be made" (p. 51).

The most widely accepted model of working memory was proposed by Baddeley and Hitch (1974; Baddeley, 1986). The model constitutes a multicomponent memory system composed of a central executive and two domain-specific slave systems: a phonological loop, responsible for the temporary storage and manipulation of verbal and acoustic information, and a visual-spatial

sketchpad, specialized for storing and processing visual and spatial information. Baddeley (2000) later added a fourth subsystem to the model called the episodic buffer. This component integrates visual, spatial, and verbal information from the two slave systems and from LTM into single multimodal units or *episodes* (e.g., a story or a scene in a film).

The two working memory components that have received the most attention from language acquisition researchers are the phonological loop and the central executive. The phonological loop comprises two subcomponents: a phonological store and a process of articulatory rehearsal. The phonological store can hold verbal information in phonological code for short periods of time before the stored information is lost due to decay or interference. The articulatory rehearsal process is equivalent to subvocal speech and is used to translate nonauditory material into phonological form and to reactivate the fading memory traces in the phonological store by retrieving and rearticulating them. Because articulation takes place in real time, the phonological loop is of limited capacity—the more items need to be reactivated by rehearsal, the more likely that the first item will decay before it can be rehearsed. Phonological loop capacity is typically assessed by immediate serial recall, using either a set of numbers or unrelated words or nonwords (Baddeley, 2003). Two common measures of phonological short-term memory (PSTM) capacity are the forward DS and NWS tasks, involving the repetition of increasing numbers of digits and nonwords of varying lengths, respectively. Although the codes on which these two tasks are based are different, they are both highly sensitive to phonological storage capacity, that is, the endurance of verbal information in memory, which enables further, deeper processing to occur.

There is a considerable body of empirical evidence suggesting that PSTM plays a key role in L2 acquisition. For example, PSTM has been shown to correlate with aspects of speech production (Kormos & Sáfár, 2008; O'Brien, Segalowitz, Freed, & Collentine, 2007) and L2 learners' ability to learn new vocabulary (Cheung, 1996; Masoura & Gathercole, 2005; Papagno & Vallar, 1992; Service & Kohonen, 1995; Speciale, R. Ellis, & Bywater, 2004). The relationship between PSTM and vocabulary learning is likely to be a result of a direct interaction between PSTM and LTM. The phonological representations temporarily stored in PSTM serve as a basis for constructing novel long-term phonological representations (Baddeley, 1986), and the LTM traces that result are used to support subsequent processing of verbal information in the phonological loop (Gathercole, Hitch, Service, & Martin, 1997).

Some L2 researchers have suggested that similar mechanisms underlie the acquisition of grammatical rules, and therefore PSTM might play a role

not only in the acquisition of lexis but also in that of morphosyntactic rules. Ellis and colleagues (N. Ellis, 1996; N. Ellis & Schmidt, 1997; N. Ellis & Sinclair, 1996; N. Ellis, 2005), for instance, explain that language utterances stored in PSTM enable the establishment of LTM representations for those same utterances, and "[LTM] for language utterances serves as the database for automatic implicit processes that abstract distributional frequency information, allowing the representation of word class and other grammatical information" (N. Ellis & Schmidt, 1997, p. 159). That is, in this view, PSTM is implicated in the learning of grammatical rules. Indeed, there is an increasing number of empirical studies demonstrating that phonological loop capacity is predictive of L2 learners' ability to acquire new grammatical forms (N. Ellis & Schmidt, 1997; N. Ellis & Sinclair, 1996; O'Brien, Segalowitz, Collentine, & Freed, 2006; Williams & Lovatt, 2003). Of particular relevance here is that PSTM has also been shown to mediate the relationship between recasts and L2 grammatical development. Learners with high phonological loop capacity, probably due to their superior capability to hold recasts and forms in memory (N. Ellis, 2005), appear more likely to benefit from recasts in the longer term (Mackey et al., 2002; Trofimovich et al., 2007). So far, however, no empirical research has been done to examine how PSTM might mediate the link between recasts and changes in learner performance on different types of outcome measures.

Besides the phonological loop, the central executive component of working memory has also been the object of considerable attention in language acquisition research. Baddeley (2003) states that the central executive or supervisory attentional system is "the most important but least understood component of working memory" (p. 835). It controls complex cognitive operations such as focusing, dividing, and switching attention; activating and inhibiting processing routines; and regulating the information flow from the short-term storage subsystems and from LTM. The central executive, like PSTM, is limited in capacity; the efficiency with which it can complete a task is largely dependent on the number and nature of other activities that it needs to complete. The most common measures used to assess central executive capacity are complex working memory tasks such as reading and listening span (Daneman & Carpenter, 1980; Waters & Caplan, 1996). In these tasks, the participant needs to read or listen to a series of sentences and subsequently recall the final word of each preceding sentence at the end of each set. Unlike PSTM tasks, which primarily tap storage capacity, these complex verbal working memory tasks require both storage and processing of verbal information and hence have been argued to provide information also about the functioning of the central executive (Gathercole, 1999).[2]

Individual differences in complex verbal working memory capacity have been found to predict abilities in a variety of complex cognitive activities, including L2 literacy skills and language learning and processing abilities. In particular, there is considerable evidence suggesting that complex working memory capacity is involved in L2 morphosyntactic processing (Juffs, 2004; Miyake & Friedman, 1998; Sagarra, 2007b), comprehension (Harrington & Sawyer, 1992; Kormos & Sáfár, 2008; Leeser, 2007; Walter, 2004), production (Kormos & Sáfár, 2008; Mackey et al., 2010), and the acquisition of grammar (Harrington & Sawyer, 1992; Kempe & Brooks, 2008; Leeser, 2007). Most importantly for the present study, complex verbal working memory capacity has also been proposed as a factor influencing the extent to which learners notice and retain recasts (e.g., N. Ellis, 2005).

To date, however, the results of empirical research regarding the relationship between complex working memory capacity and the efficacy of recasts are mixed. Mackey et al.'s (2002) study suggests that working memory capacity, as measured by a composite score of PSTM and listening span tasks, may moderate the link between recasts, noticing, and development in L2 question formation. The researchers reported that learners with higher working memory scores showed more noticing of recasts than learners with lower working memory scores, and those with higher working memory achieved greater development in the longer term. Interestingly, however, those with lower working memory demonstrated more substantial immediate gains. In a study investigating computer-delivered oral recasts, Sagarra (2007a) likewise found that higher RS learners outperformed lower-span participants in terms of grammatical accuracy. Notably, in this study high-span learners showed greater gains on both immediate and delayed posttest measures. In contrast, Trofimovich et al. (2007) detected no significant effects for complex working memory capacity in examining the effects of complex working memory capacity, PSTM, analytical ability, and attention control on the ability to notice and learn from computer-administered recasts.

One possible explanation for the distinct findings in the three studies (Mackey et al., 2002; Sagarra, 2007a; Trofimovich et al., 2007) might lie in differences in the methodology used. For instance, Trofimovich et al. speculate that they probably failed to detect a significant effect for complex working memory capacity because, in the computerised treatment task they employed, each learner utterance was followed by a native speaker response, regardless of whether an error occurred. This is likely to have made the native-speaker reactions predictable and salient to learners, and therefore the recasts might have succeeded in attracting the focal, conscious attention of even low complex

working memory learners, resulting in a ceiling effect. In Mackey et al.'s and Sagarra's studies, on the other hand, the recasts were much less predictable, being exclusively delivered in response to erroneous utterances. Noticing of feedback, therefore, was probably more dependent on learners' ability to switch and focus attention, mechanisms which are associated with the central executive component of working memory.

Clearly, more empirical research is needed to elucidate the nature of any relationship between recasts and working memory. The present study is among the first to investigate whether working memory, as measured by DS, NWS, and RS tests, might affect the extent to which any effects of recasts are observed on various L2 outcome measures.

Research Questions

This study addressed the following research questions:

1. Does the type of outcome measure employed influence the observed effects of recasts on L2 development?
2. Are PSTM and/or complex working memory capacity related to whether any effects of recasts are observed on different types of outcome measures?

In the present study, recasts were operationalized narrowly as recasts provided in response to learner errors in the use of the past progressive construction, and the effects of recasts were evaluated in terms of any improvement learners showed in the knowledge of this construction as a result of receiving recasts.

Method

Design

The data set for the present study, except for the working memory data, was collected as part of a larger study (Révész, 2007) investigating the observed effectiveness of recasts in relation to task complexity and various outcome measures. The original study employed a pretest–posttest–delayed posttest design, with 90 participants randomly assigned to one of four experimental groups and a control group, reflecting a 2 × 2 design defined by two independent variables: *recasts* and *task complexity*. The focus of the present study is on the link between recasts, working memory, and different types of outcome measures (see Révész, 2009, for a description of the effects of recasts and task complexity on oral development). Accordingly, it examined any difference among the performance of the control group ($n = 18$), the two experimental groups

who received recasts ($n = 36$, henceforth the "recast group"), and the two experimental groups who did not receive recasts throughout the study ($n = 36$, henceforth the "nonrecast group").[3] The control group participated only in the pretest and posttests, whereas the recast and nonrecast groups took part in three treatment sessions between the pretest and the posttest. At each testing session, three types of tasks were used to assess the extent of learning triggered by the respective treatments: a GJT, a written production task, and two oral production tasks. The delayed posttest was administered to half of the participants in each group. A subset of the learners (recast group: $n = 22$, nonrecast group: $n = 23$) was also administered tests of PSTM (DS and NWS) and a test of complex working memory capacity (RS). The design of the study is summarized in Figure 1.

Target Construction

Recasts in the present study targeted the past progressive construction in English. The rationale for choosing this particular linguistic feature was twofold. First, the past progressive is realized via a free morpheme (*was/were*) and a syllabic bound morpheme (*-ing*); thus, it is physically salient (Goldschneider & DeKeyser, 2001). Second, it denotes grammatical tense and aspect, meaning that it has some communicative value. Previous research suggests that recasts are more likely to draw learners' attention to linguistic forms that fit the criteria of being perceptually salient and meaning bearing (Long, 2007).

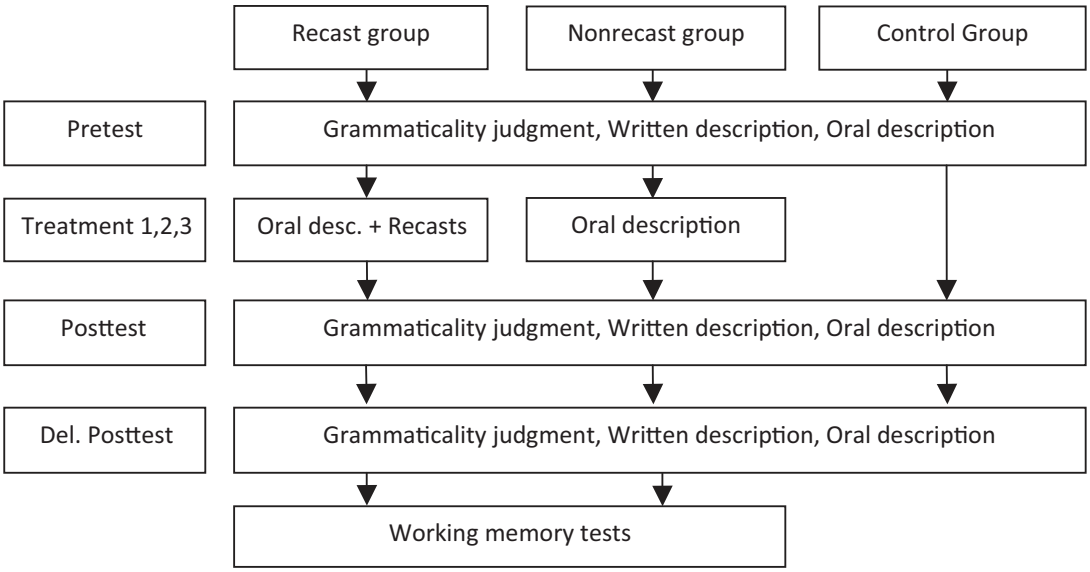

Figure 1 Design of the study.

The experimental tasks in the current study elicited only one context of the past progressive, generally considered to be the prototypical usage of this form: when it refers to something that was in progress at a particular time in the past, for example, *He was jogging at 7 o'clock last night.* As demonstrated by Bardovi-Harlig (2000), learners pass through relatively fixed developmental stages in learning this construction:

Stage 1: bare progressive (e.g., *walking*)
Stage 2: present progressive (e.g., *is walking*)
Stage 3: past progressive (e.g., *was walking*)

In light of these findings, participants in the present study were considered developmentally ready if they showed accurate use of the present progressive form (but not yet of the past progressive) on the pretest.

Participants

The 90 participants in this study were learners of English as a foreign language (EFL) enrolled in beginner-level language classes in three high schools in Hungary. The pedagogical approach adopted by the schools was a mix of focus-on-forms and communicative language instruction. The students were placed in their classes either based on the results of a placement test administered at the beginning of the school year or due to the successful completion of a prior-level course. The written parts of the pretest, the GJT and the written production task, were administered to 139 students from 11 intact classes. The classes were selected with the help of expert opinions of teachers from the three institutions. Only classes that had not received prior instruction targeting the past progressive construction were administered the pretest. Each participant had been in their current class for at least 6 months at the onset of the experiment. Of the initial 139 participants, 105 students, who showed no accurate use of the target construction on the written pretest, took the oral pretest. Based on the overall pretest results, 95 students proved developmentally ready for the linguistic focus and thus eligible for participating in the study. Two students declined participation at this point owing to scheduling conflicts. Of the remaining 93 students, 90 randomly selected students were invited to continue the study.

For all three groups combined, there were 47 female and 43 male students. The participants' ages ranged from 16 to 20 ($M = 16.87$, $SD = 1.42$), and they were all native speakers of Hungarian. They had received between 6 months and 8 years of English instruction prior to the study ($M = 2.68$, $SD = 2.37$). Most of the students had never visited an English-speaking country. Kruskal-Wallis tests run on the factors of age and length of previous English study

revealed no significant differences among the three groups with regard to these variables, $\chi^2(2, N = 90) = 1.12$, $p = .57$, and $\chi^2(2, N = 90) = 4.73$, $p = .09$, respectively.

Treatment

Tasks

The treatment tasks were contextualized in the hypothetical scenario that the participants were taking photos in a New York City neighborhood (e.g., Soho) exactly at a time when a crime (e.g., a bank robbery) happened in that area. During the experiment, the participants' task was to describe their photos to the researcher, who played the role of a police officer. The participants were told that they should describe each activity in the photos carefully, because the police wanted to know what everybody was doing at the time the crime occurred.

Three versions of the treatment task were developed, one for each treatment session. Each version of the task contained 10 photos, that is, each participant in the recast and nonrecast groups described 30 photos altogether over the course of the treatment. The tasks were prepared using the computer program Microsoft PowerPoint. The first slide of each presentation displayed the title (e.g., *Soho*) for 5 seconds, and was followed by a slide describing the task instructions, which were visible for 2 minutes. Then, after the researcher had orally checked that the participants understood the task, 10 photos followed. On each photo, a title appeared on the screen indicating the time the photo had been taken. This title disappeared after 10 seconds. Next, participants were asked to describe the photo in 40 seconds. Half of the participants in both the recast and nonrecast groups could see only a blank screen while speaking, whereas, for the rest of the participants, the photo remained available on the screen (the variable examined in Révész, 2009). In each photo, at least three people were engaged in clearly identifiable activities, such as sitting, painting, or walking. The time intervals allocated to view and describe the photos were determined through piloting the task with elementary-level EFL learners.

The three versions of the task were piloted with native speakers of English and beginning-level Hungarian learners of English. The native baseline data indicated that the tasks generated approximately the same number of obligatory contexts for the target form and that the use of the target construction was natural in each task. The three versions were also found comparable in terms of lexical variation, measured by Guiraud's index, and syntactic complexity, measured by clauses per Analysis of Speech unit (Foster, Tonkyn, & Wigglesworth, 2000), based on both the native baseline and EFL learner data.

Recasts

During the treatment sessions, the participants in the recast group consistently received recasts from the researcher whenever they produced errors in their use of the past progressive construction, as shown earlier in (1). The recasts were typically of the simple isolated declarative type (Kim & Han, 2007; Lyster, 1998), provided with falling intonation, without added emphasis on the targeted feature. A small number of recasts were also provided, albeit randomly, in response to other interlanguage forms. Participants were given no other type of feedback during the study. The nonrecast group did not receive feedback in any form.

Assessment Tasks and Scoring

Three different but comparable versions of each of the three assessment tasks were developed. These were administered in a split-block design.

Grammaticality Judgment task

The GJT consisted of 36 sentences: 16 past progressive items, 8 present progressive items, and 12 distractors. Half of the items for each item type were grammatical and the other half ungrammatical. Participants were required to indicate whether each sentence was grammatically correct, and to correct the relevant error if they judged a sentence to be ungrammatical. The ungrammatical items contained typical interlanguage error types as documented by previous SLA research. For the past progressive sentences, they included use of the bare progressive and that of the present progressive in past progressive contexts (Bardovi-Harlig, 2000). For the present progressive sentences, they contained instances of undersuppliance and use of the bare progressive in present progressive contexts. No time limit was given for the task. Participants, therefore, had sufficient time to deliberate about a judgment and, thus, to deploy their knowledge of metalinguistic rules (R. Ellis, 2005). That is, this instrument is likely to have allowed learners to draw on their declarative knowledge.

Learners' responses to the past progressive items were scored in terms of four categories, motivated by Bardovi-Harlig (2000). Three points were given for grammatical items judged as grammatical and for ungrammatical items supplied with appropriate corrections. Two points were awarded for changing grammatical sentences to present progressive, for changing ungrammatical bare progressive sentences to present progressive, and for judging ungrammatical present progressive sentences as grammatical. One point was given for correcting a grammatical item into a bare progressive, for changing an ungrammatical present progressive form into a bare progressive, and for judging a bare

progressive item as grammatical. Zero points were awarded for grammatical items judged as ungrammatical with any nontargetlike change to a form different from the ones listed above. Any sentence that was judged ungrammatical but for which no correction was supplied was excluded. Also excluded were items where the corrections indicated that the sentences were judged on the basis of linguistic forms other than the targeted form. In assessing development from pretest to posttests, only the 16 past progressive items were coded. The maximum score was 48 points. Participants were considered developmentally ready for the past progressive on the GJT pretest if they provided an appropriate correction for at least two ungrammatical present progressive items.

As part of piloting, the three versions of the GJT were administered to 15 Hungarian EFL learners from a population similar to that of the participants. A Kruskal-Wallis test found no significant differences among the versions, $\chi^2(2) = 3.43$, $p = .18$. In the main study, the internal consistency reliability coefficients of the past progressive items were high for each version of the test ($\alpha > .90$), and the means and standard deviations were in a similar range ($25 < M < 29$; $15 < SD < 18$). A second rater scored 20% of the data. Cohen's kappa was .96, indicating a high level of consistency between the two raters.

Written Picture Description Task

As part of this written task, the participants were asked to describe a picture showing eight people engaged in various activities in a park. The written instructions contextualized the picture as having been taken at a particular time of day, for example, *4 o'clock yesterday*, and asked the participants to describe what the people were doing at that time. Through piloting with native speakers and EFL learners, it was ensured that the three versions of the task generated the same number of obligatory contexts for the target construction. The participants were provided 10 minutes for the task. This test is likely to have allowed for the use of both declarative and procedural knowledge: While it was relatively unpressured in terms of time, allowing for the use of declarative knowledge, it required learners to produce the target construction, which involved the deployment of procedural knowledge.

The coding of the data consisted of four steps. The first step was to identify obligatory contexts for the past progressive. Then, it was determined whether any progressive marking had been produced in these contexts. Next, the data were analyzed in terms of four categories based on the developmental sequence for progressives (Bardovi-Harlig, 2000). In obligatory contexts, participants received three points for the use of the past progressive, two points for using the present progressive, one point for the bare progressive, and zero points for

any nonprogressive form. The participants' total score was calculated using the formula below.

$$T = \sum_{i=1}^{N} p_i / N \, x \, 3,$$

where T is the total score on the task, Σ is the sum, N is the total number of obligatory contexts, and p is the number of points received for each obligatory context. A second rater scored 20% of the data, randomly selected across the three groups. Cohen's kappa was .96, demonstrating strong inter-coder agreement. In scoring the pretests, the data were also checked for developmental readiness: Participants were considered developmentally ready if they used the present progressive form at least twice in a past progressive context, but displayed no correct use of the past progressive form.

Oral Description Tasks

Two oral production tasks served as the oral part of the assessment. The specifications of these instruments were aligned with those of the treatment task, except for being shorter in length. Instead of 10, they required participants to describe 5 photos. Six versions of the tasks were developed. These were subjected to the same piloting procedures as the treatment tasks. The sequence of the six versions was counterbalanced within the testing sessions. Arguably, these tasks required the use of procedural knowledge, because they involved spontaneous, meaning-based, productive use of language and were performed under time pressure.

The participants' oral production, altogether 87 hours of oral data, was tape-recorded, transcribed, and then coded following the same procedure described above for the written picture description task. Ten percent of the data, randomly selected, was transcribed by a second researcher. Inter-transcriber agreement was high (.97). A second rater also coded 20% of the data, randomly selected across the three groups. Cohen's kappa was .92, indicating strong inter-coder agreement.

Tests of PSTM and Complex Working Memory Capacity

PSTM, operationalized as the serial repetition of unknown sequences of auditory input, was assessed with a DS and a NWS test. DS was measured by a Hungarian version of this type of test (Racsmány, Lukács, Németh, & Pléh, 2005). Participants heard a series of random numbers in Hungarian at the rate of one digit per second and had to repeat the digits in the presented order. The

stimuli, recorded by a female native speaker of Hungarian, consisted of four lists of numbers for each consecutive sequence length. The sequences ranged from three to nine digits, and were presented in ascending series. Participants' DS was determined as the maximum list length at which they could repeat at least two of the four sequences correctly.

NWS was also assessed by a Hungarian version of this type of test (Racsmány et al., 2005). It was important to administer this test in Hungarian, given that knowledge of L2 phonological rules and phonotactic structure has been shown to affect the ability to recall novel strings in the L2 (Masoura & Gathercole, 1999). In other words, using the Hungarian version made it more likely that any correlation found between participants' development and the NWS test was not confounded with differences in the learners' L2 proficiency. The test, recorded by the same female native Hungarian speaker, contained 36 nonwords that adhered to Hungarian phonotactics. The nonwords ranged from one to nine syllables in length, with a total of four nonwords for each number of syllables. The participants were presented with and asked to repeat the nonwords one by one in a predetermined sequence (nonwords with equal syllable length were not presented together). Participants' NWS was indicated by the highest number of syllables that they were able to recall at least twice out of the four nonwords for that length.

Complex working memory capacity, defined as the mechanism for the temporary storage and processing of linguistic information, was measured with a RS test (Daneman & Carpenter, 1980). The test was adapted from a Hungarian version of the RS test (Racsmány et al., 2005) in order to avoid the potential confounding effect of variation in L2 proficiency. It consisted of 60 sentences which participants were asked to read aloud one by one. The sentences, typed on index cards, were organized into three large sets consisting of smaller subsets of two to six sentences and presented in an order of gradually increasing set length. At the end of each subset, participants were asked to recall the last word of each sentence in that set, and to answer a question designed to assess their understanding of one of the sentences in the set. The words to remember were mid-frequency words composed of two syllables. The sentences targeted by the comprehension questions were randomly selected but the same for all the participants. The rationale for including a comprehension task was to ensure that the participants would focus not only on remembering the words but also on processing the meaning of the sentences (Waters & Caplan, 1996). The learners were told that their performance on the recall and the comprehension task was equally important. RS was indicated by the maximum number of sentences participants could read while correctly repeating the final words averaged across

the three sets. The comprehension scores were not included in the test results, because all the participants were able to answer the comprehension questions correctly at least 98% of the time, suggesting that they were not exclusively concentrating on recalling the words but were also processing for meaning.

Data Collection Procedures

Each participant attended six to seven sessions over the course of 6 weeks. On the first day of the experiment, the written parts of the pretest, a version of the GJT and of the written description task, were administered during normally scheduled class times. Based on the results of the written pretest, the researcher invited learners who appeared developmentally ready to acquire the target construction to participate in the oral pretest. Learners were given the oral pretest individually on a different day within the same week. Students who also proved eligible for the study in terms of the oral pretest were invited to continue. The written pretest took 30 minutes, whereas the oral pretest lasted 10 minutes. The treatment started a week after the oral pretest and took place on 3 separate days over a 1-week period. The treatment sessions lasted 15 minutes each. The day after the last treatment session, the participants were given the written and oral posttests. Half of the participants, randomly selected from each group, also performed a delayed posttest 4 weeks later. Both the immediate and delayed posttests were administered to participants individually and lasted 40 minutes. In an attempt to control for exposure to the target construction outside the experiment, the teachers of the participating EFL classes agreed not to focus on the target feature during the period of data collection. The working memory tests were administered to the participants individually approximately 6 months after the posttests.

Statistical Analyses

To examine the impact of the treatment on the participants' knowledge of the target construction, many-facet Rasch measurement (MFRM) was employed (Linacre, 1989). MRFM concurrently computes estimates for the effects of various facets, that is, definable factors of an assessment setting that may contribute to test score variation (e.g., person ability, item difficulty, task type, group assignment in experimental designs, background variables such as gender and age). The procedure transforms raw data into true interval scores and produces measures for each facet on a true interval scale, known as the logit scale. In addition to calculating logit estimates for each facet, Rasch analysis also computes the significance of any differences that may exist among elements of a given facet, for example, differences in person abilities, item difficulty, or gains among experimental groups.

By using true rather than raw scores, Rasch measurement offers an important advantage over classical statistical analyses. In particular, it helps avoid "the problem of bias towards scores in the middle of the scale, and against persons who score at the extremes" (Bond & Fox, 2001, p. 17). As Bond and Fox explain, in analyses drawing on raw scores, a small move in scores near the average of test results (e.g., 48 to 55, where the average score is 50) is assumed to reflect the same ability gap needed for a leap toward the top of the test scores (e.g., from 88 to 95). In other words, the sizes of gaps between test scores are presumed to correspond to actual distances, while, in fact, we can only deduce the ordering of, but not the true distance between, person abilities from raw data. Rasch analysis, by converting raw scores into their natural logarithm or log-odds (logits), solves this problem.

Once the analysis has transformed raw scores into log-odds, it tests the resulting true scores against the stochastic expectations of the model. What is expected by the Rasch model is a probabilistic form of Guttman scaling, which, in simple terms, assumes that persons who are better able are more likely to correctly answer all the items on a test, and there is a greater likelihood that easier items will be answered correctly by all persons (Bond & Fox, 2001). Rasch analysis also provides what is known as fit statistics for each element of a facet, which indicate how well the data fit the stochastic expectation of the model. In other words, fit statistics identify observations that fall outside the expected range of variability. In the case of person abilities, for example, fit statistics isolate irregular response behavior, such as correctly answering a greater number of difficult than easy items on a test. For items and treatments, fit statistics show how consistently a particular item or treatment type has impacted participants' performance. In the current study, infit values were considered, which indicate the extent to which the observations fit the modeled expectations, weighted to provide increased value to on-target observations.

For the current study, a total of three MFRM analyses were performed using the computer program FACETS 3.61 (Linacre, 2006). The first Rasch analysis included the pretest–posttest and pretest–delayed posttest gain scores obtained from the GJT. The second MFRM analysis was performed on the pretest–posttest and pretest–delayed posttest gain scores on the written description task. The third Rasch analysis was run on the pretest–posttest and pretest–delayed posttest gain scores obtained from the oral description tasks. The rationale for submitting the data to three separate Rasch analyses lay in that the GJT, the written description task, and the oral description tasks were expected to tap various types of L2 knowledge to different degrees.

The model used for the analyses was the Rating Scale model, which assumes that the steps of a scale are equivalent across all elements of a given facet. The MFRM analyses for the GJT and oral description tasks were specified as having four facets: (1) participants' gains, (2) group (recast, nonrecast, control), (3) time (posttest versus delayed posttest), and (4) the difficulty of test items (GJT)/testing tasks (oral description). The mathematical model used for the analyses was:

$$\log\left(P_{njikx}/P_{njikx-1}\right) = B_n - C_j - D_i - E_k - F_x,$$

where P_{njikx} is the probability of participant *n* in group *j* achieving a gain score of *x* on testing task *i* at time *k*, $P_{njikx-1}$ is the probability of participant *n* in group *j* achieving a gain score of *x-1* on testing task *i* at time *k*, B_n is the pretest-posttest gain of participant *n*, C_j is group *j*, D_i is the difficulty of item/testing task *i*, E_k is time *k*, and F_x is the difficulty of achieving a gain score of *x* relative to a gain score of *x-1*. The specifications and the mathematical model for the MFRM analysis for the written description task only differed in that they did not include the facet *item/testing task.*

The data from the GJT were originally modeled on a 4-point rating scale in line with the possible raw score values (0–3) determined for the task. However, FACETS yielded disordered average measures and step calibrations for this scale, indicating problems with its functioning.[4] Therefore, the data were recoded into a 3-point rating scale by collapsing the partial scores 1 and 2 into one category. For the oral and written description tests, originally a 100-point rating scale was specified in correspondence with the range of percentage scores obtained. However, owing to disordered average measures and step calibrations obtained for the 100-point scale from FACETS, the percentage scores were collapsed into a 6-point rating scale by recoding the data (Linacre, personal communication, 2006). Although rating scales with different step structures were also tested, the 6-point scale appeared best for representing the current data sets, because they included the most categories while still yielding ordered average measures and step calibrations.

For all three analyses, the mean ability value of each group was anchored at zero logits. In this way, it was ensured that there was enough connectivity in the data set, which, due to the between-subjects design, would have otherwise contained disjoint subsets. Given that the participants were assigned to the experimental and control groups at random, we could assume that the mean abilities of the participants assigned to each group were randomly equivalent. The mean difficulty of each facet was set at 0 logits, except for the facet *time*

on the written description task and the facet *item/testing task* on the GJT and oral tasks.

To examine whether working memory modulated the effects of the recast and/or nonrecast treatment on L2 development, a series of correlations were computed. Specifically, in separate analyses for the two experimental groups (recast and nonrecast), Pearson correlations were calculated between the three working memory measures and the pretest–posttest gain scores obtained from the GJT, the written description task, and the oral tasks. All correlations were performed using the Statistical Package for the Social Sciences (SPSS) 16.

Results

Number of Errors and Recasts

Participants in the recast group on average committed 25.30 errors in the use of the target construction, with a standard deviation of 16.93. All of these errors were addressed with recasts. The mean number of recasts gradually decreased across the three treatment sessions (treatment 1: $M = 13.17$, $SD = 6.81$; treatment 2: $M = 7.42$; $SD = 6.34$; treatment 3: $M = 4.72$, $SD = 5.42$), indicating that learners tended to commit fewer errors in the use of past progressive over the course of the experiment.

Research Question 1: Recasts and Different Types of Outcome Measures

Grammaticality Judgment task

Table 1 shows the descriptive statistics for the scores obtained on the GJT. The recast group showed considerable improvement from the pretest to the

Table 1 Descriptive statistics for the grammaticality judgment task

Group	Test	*N*	Mean	Mean Gain	*SD*
Recast	Pretest	36	19.61	—	13.61
	Posttest	36	39.14	19.53	10.68
	Delayed Posttest	18	38.98	−.16	13.23
Nonrecast	Pretest	36	22.23	—	12.36
	Posttest	36	22.43	.19	13.37
	Delayed Posttest	18	24.10	1.69	12.29
Control	Pretest	18	19.39	—	10.25
	Posttest	18	20.55	1.16	14.53
	Delayed Posttest	9	19.14	−1.41	11.73

Note. The maximum total score was 48 points.

posttest (19.53), whereas the nonrecast group (.19) and the control group (1.16) remained relatively stable. The recast group maintained their gains on the delayed posttest.

Moving on to the results of the FACETS analysis, the summary map of the four facets for the GJT is presented graphically in Figure S1 in the Supporting Information online. The first column in the figure displays the logit scale. The logit scale is an equal-interval scale, which provides a single frame of reference for all the facets of the MFRM analysis, allowing for comparisons both within and between the facets. The second column presents the distribution of the GJT pretest–posttest gain scores in logits. Each star (*) represents one participant's gain score. Participants with higher gains appear at the top of the column and participants with lower gains at the bottom of the column. The third column demonstrates the variation among the GJT task items in terms of difficulty. Items appearing higher in the column were more difficult to achieve high pretest–posttest gain scores on than items appearing lower in the column. The fourth column compares group gains. Groups appearing higher in the column displayed lower gains, while the groups appearing lower in the column exhibited higher gains. The fifth column compares participants' pretest–posttest gain scores with their pretest–delayed posttest gain scores. The test appearing higher in the column was more difficult for participants to achieve gain scores on than the test appearing lower in the column. Finally, the sixth column graphically describes the rating scale used to score participants' performance on the GJT.

Next, a summary of the statistics for the two facets directly relevant to the research questions (namely, *group* and *time*) is presented. For the facet *group*, the gain score estimates ranged from −1.47 to 1.13 logits, yielding a logit spread of 2.60. The standard deviation generated by the analysis was 1.33 logits. The overall difference between the group estimates was significant, $\chi^2(2) = 747.5$, $p < .001$. The separation reliability, analogous to Cronbach's alpha, was .99. These indices demonstrate that the three groups' pretest–posttest gains reliably differed from each other. In particular, the results showed that the recast group achieved considerably higher gains (−1.47) than the control (.35) and the nonrecast group (1.13). Note that lower logit values are associated with higher gains; this reflects the fact that groups with higher gains experienced less difficulty on the posttest as compared to the pretest. Overall, the results for the facet *group* suggest that, at least in part, it was the recasts that had led to the observed changes in the recast group's pretest–posttest gains. As per the infit statistics, the infit mean-square was 2.09 for the facet, with a standard deviation of .32. Hence, following Pollitt and Hutchinson's (1987) criterion (i.e.,

M ± *2SD*), any value outside the range of 1.45 to 2.73 would have been considered misfitting. All three elements of the facet, however, had an infit value inside this range, in line with the modeled expectations.

Finally, the results for the *time* facet on the GJT are summarized. The difficulty estimate for the posttest was .07 logits, whereas the difficulty estimate for the delayed posttest was −.07 logits. Thus, the overall difficulty span between the posttest and delayed posttest measures was a small .14. The standard deviation was .07, and the separation reliability was moderate (.60). The fixed chi-square test was not significant $\chi^2(1) = 2.5, p = .11$. Overall, these indices suggest that the participants' gain scores were not significantly different on the delayed posttest and the posttest. Neither element of the facet was identified as misfitting: the fit values for both the posttest (2.14) and the delayed posttest (2.06) were within the acceptable range of 1.98 to 2.22.

Written Description Task

The descriptive statistics for the written description task appear in Table 2. The recast group improved considerably more (.87) than the nonrecast group (.22) from the pretest to the posttest. The control group also showed a small apparent increase compared to the pretest (.10). All groups, however, exhibited a small decrease from the posttest to the delayed posttest (−.14 to −.07).

The summary of the FACETS results for the written description task appear in Figure S2 in the Supporting Information online (the figure can be interpreted similarly to Figure S1). For the *group* facet, the gain estimates for the three groups ranged from −3.40 to 2.20 logits, with a spread of 5.60 logits. The

Table 2 Descriptive statistics for the written description task

Group	Test	*N*	Mean	Mean Gain	*SD*
Recast	Pretest	36	.08	—	.13
	Posttest	36	.95	.87	.13
	Delayed Posttest	18	.88	−.07	.24
Nonrecast	Pretest	36	.09	—	.13
	Posttest	36	.31	.22	.24
	Delayed Posttest	18	.17	−.14	.17
Control	Pretest	18	.07	—	.11
	Posttest	18	.17	.10	.18
	Delayed Posttest	9	.09	−.08	.14

Note. Scores ranged from 0 to 1.0 points, reflecting the proportion of developmentally advanced forms.

standard deviation was 2.44 logits. The gain estimates were found to be reliably separable (separation reliability = .99), as well as significantly different, $\chi^2(2) = 361.9, p < .001$. The recast group achieved substantially greater gain (−3.40) than the nonrecast group (1.21). The lowest gain was exhibited by the control group (2.20), suggesting that the treatment was responsible for the experimental groups' pretest–posttest gains on the written description task. All the infit statistics were between the acceptable limits of .77 and 4.09.

The Rasch analysis yielded the following results for the *time* facet on the written description task. The difficulty estimate for the delayed posttest was .77 logits, whereas the difficulty estimate for the posttest was .01 logits. Thus, the difference in difficulty between the two elements of the facet was .76 logits. The standard deviation generated by the analysis was .38 logits. While the reliability of the separation was moderate (.67), the fixed chi-square test was significant, $\chi^2(1) = 10.0, p < .01$. This means that, although the difference between the two elements of the facet was relatively small (.76), the participants' gains on the immediate posttest were significantly greater than those on the delayed posttest. The infit mean-square values for both the posttest (2.21) and the delayed posttest (2.91) were within the acceptable range of two standard deviations around the mean (1.86 to 3.26).

Oral Description Tasks

Table 3 shows the descriptive statistics for the mean scores on the oral description tasks. Similar to what was observed on the GJT and the written description task, the recast group exhibited a substantial gain from the pretest

Table 3 Descriptive statistics for the oral description tasks

Group	Test	*N*	Mean	Mean Gain	*SD*
Recast	Pretest	36	.13	—	.10
	Posttest	36	.87	.74	.13
	Delayed Posttest	18	.84	−.03	.12
Nonrecast	Pretest	36	.22	—	.11
	Posttest	36	.31	.08	.16
	Delayed Posttest	18	.36	.05	.19
Control	Pretest	18	.22	—	.11
	Posttest	18	.24	.02	.10
	Delayed Posttest	9	.18	−.06	.11

Note. Scores ranged from 0 to 1.0 points, reflecting the proportion of developmentally advanced forms.

to the posttest (.74). The nonrecast group also showed a small increase of .08. The control group displayed no change from the pretest to the posttests. On the delayed posttest, the recast group maintained their gain, and the nonrecast group showed a slight increase (.05) in their use of the past progressive. The control group showed a slight decrease from the posttest to the delayed posttest (−.06).

Figure S3 in the Supporting Information online shows graphically the results of the FACETS analysis for the oral production tasks. First, the summary statistics for the *group* facet are presented. The gain estimates for the groups spanned from −4.09 to 2.43 logits, yielding a logit spread of 6.52 logits. The standard deviation was 2.91 logits. The overall difference between the group estimates was significant, $\chi^2(2) = 1041.7, p < .001$, with a separation reliability of .99. These statistics indicate that the three groups' pretest–posttest gains reliably differed from each other. Again, the recast group showed considerably higher gain (−4.09) than the nonrecast group (1.66). The analysis yielded the lowest gain estimate for the control group (2.43), which suggests that the experimental groups' pretest–posttest gains were a result of their respective treatments. The infit mean-square values for the facet ranged from 1.19 to 1.83, that is, all infit statistics were in the acceptable range of .90 to 2.14.

Finally, the Rasch analysis results for the *time* facet on the oral description tasks are described. The difficulty estimate for the delayed posttest was .03 logits, whereas the difficulty estimate for the posttest was −.03 logits. The standard deviation was .04 logits. Thus, the overall difficulty span between the posttest and delayed posttest measures was small (.06). The reliability of the separation (.01) was low, and the fixed chi-square test was not significant, $\chi^2(1) = .1$, $p = .75$. In other words, there was no significant difference in participants' performance on the posttest versus the delayed posttest. Neither element of the *time* facet was identified as misfitting or overfitting; the infit values for both (posttest = 1.38; delayed posttest = 1.93) were within the acceptable range of 1.11 to 2.19.

Correlations of Gain Scores

There were large correlations between participants' gain scores on the three types of assessment. The sizes of the correlations, however, were considerably larger between the oral and written description gain scores (pretest–posttest: $r = .86$; $p < .01$; pretest–delayed posttest: $r = .85$; $p < .01$) than between the learners' GJT gains and their gains on the other two measures ($.53 < r < .57$; $p < .01$).

Table 4 Descriptive statistics for the working memory measures

WM measure	Test	*N*	Mean	*SD*
Digit span	Recast	22	6.26	0.96
	Nonrecast	23	6.09	0.87
	Total	45	6.18	0.91
Nonword span	Recast	22	5.74	0.86
	Nonrecast	23	5.64	1.00
	Total	45	5.69	0.92
Reading span	Recast	22	3.09	0.60
	Nonrecast	23	2.97	0.48
	Total	45	3.03	0.54

Note. WM = working memory.

Research Question 2: Recasts, Working Memory, and Different Types of Outcome Measures

The descriptive statistics for the mean scores on the three working memory measures—DS, NWS, and RS—are presented in Table 4. The participants' working memory scores appear close to the national averages (NA) for the participants' age groups, DS (NA): ages 12–16 = 6.32, ages 16–20 = 6.39; NWS (NA): ages 12–16 = 5.17, ages 16–20 = 6.34; RS (NA): ages 14–17 = 2.56, ages 18–30 = 3.86 (Racsmány et al., 2005). Simple correlational analyses revealed a significant correlation between students' performance on the DS and NWS tests ($r = .34$, $p < .05$), but the scores on neither of these tests were found to be correlated with the RS results.

The mean working memory scores for the recast and nonrecast groups were in a similar range, with the recast group (DS = 6.26; NWS = 5.74 syllables; RS = 3.09) slightly outperforming the nonrecast group (DS = 6.09; NWS = 5.64 syllables; RS = 2.97) on all three tests. Independent samples *t* tests did not find the observed differences between the recast and nonrecast groups to be significant for any of the working memory tests, however ($.37 < t(43) < .62$, $p > .47$).

Table 5 provides the correlations between the working memory measures and the pretest–posttest gain scores. As shown, in the recast group, the extent of participants' development on the GJT and written description tests showed moderate to strong correlations with their performance on the RS test (GJT: $r = .53$; $p < .05$; written description: $r = .47$; $p < .05$), but no significant correlations were detected between the GJT and the DS and NWS tests. Conversely, moderate to strong correlations were found between the gain scores achieved

Table 5 Correlations of working memory test scores and pretest-posttest gain scores

Group	Test	GJT	Written Description	Oral Description
Recast group ($N = 23$)	Digit span	.08	.39	.49*
	Nonword span	−.05	.38	.55**
	Reading span	.53*	.47*	.01
Nonrecast group ($N = 22$)	Digit span	−.15	−.15	−.04
	Nonword span	−.30	−.01	.07
	Reading span	−.10	−.16	−.15

*$p < .05$; **$p < .01$.

by participants on the oral description test and the tests of PSTM (DS: $r = .49$; $p < .05$; NWS: $r = .55$; $p < .01$), but the oral test gain scores showed no significant correlations with the RS test results. For the nonrecast group, none of the correlations computed between the working memory and developmental measures were found to be significant. In summary, working memory did not seem to be related to the gain scores of learners who did not receive recasts, but it was associated with the extent of development achieved by learners who received recasts. Specifically, participants with higher DS and NWS showed more substantial development on the oral tests, whereas participants with higher RS were observed to exhibit greater gains on the written tests.

Discussion

The first research question asked whether the type of outcome measure used influences the observed effectiveness of recasts. The results of the study provided an affirmative answer to this question. While the FACETS analyses conducted on the scores obtained from all three types of measures (i.e., the GJT, the written description task, and the oral production tasks) revealed a significant advantage for the recast treatment over the nonrecast treatment, the extent of development varied considerably from measure to measure. Recasts were observed to have the greatest impact on gain scores on the oral production tests (approximately 5.5 logits). Participants showed substantially less improvement on the written production test (about 4 logits), and the least development was observed on the GJT (approximately 2.5 logits). As pointed out above, the GJT is likely to have drawn primarily on the use of declarative knowledge, by permitting unlimited response time and by encouraging the use of metalinguistic rules. In contrast, the oral production tasks, which entailed more spontaneous

language use, a focus on meaning, and time pressure, required learners to deploy procedural knowledge. The written description task is likely to have allowed for deployment of both declarative and procedural knowledge, because it was unpressured but required a focus on meaning. In light of this, the results of the current study allow for the speculation that recasts proved more effective in engendering gains in participants' procedural than declarative knowledge of the target construction. These findings appear to be in line with previous empirical research (Loewen & Nabei, 2007; R. Ellis, 2007; R. Ellis et al., 2006; Sheen, 2007; Révész & Han, 2006), which overall suggests that, although recasts can foster gains in both declarative and procedural knowledge, they have a greater positive influence on the acquisition of procedural, than on that of declarative, knowledge.

One possible explanation for the superior impact of recasts on improving learners' performance on the oral and written description tests has to do with the notion of transfer-appropriate processing (TAP). The underlying principle of TAP is that we can better transfer and "remember what we have learned if the cognitive processes that are active during learning are similar to those that are active during retrieval" (Lightbown, 2007, p. 27). One implication of TAP, Lightbown explains, is that "learning to use language in a communicative context may improve the ability to retrieve it in such contexts" (p. 27). From the perspective of TAP, then, recasts may have facilitated participants' development on the oral and written description tasks to a greater extent because, as opposed to the GJT, these testing tasks required learners to use the past progressive construction in communicative contexts similar to those during which the recasts had been delivered. In particular, the oral and written description tasks were parallel in that they requested learners to describe people engaged in various activities at different places at a particular time.

TAP can also explain why participants displayed higher gains on the oral production tasks than on the written description task. According to skill-acquisition theory, procedural knowledge is skill specific, that is, proceduralized knowledge acquired via practice is hard to transfer from one skill, such as speaking, to another, such as writing (DeKeyser, 2007a, 2007c). Applying this tenet to the current study, it could be argued that it was more difficult for learners to retrieve the newly learned material in a written context, because they were exposed to an oral rather than a written treatment. However, it is also important to emphasize that, although transfer was limited from spoken to written production, it did take place in the current study, as evidenced in the learners' pretest–posttest gains on both the GJT and written description test. The existence of these transfer effects suggests that the treatment might not

have exclusively promoted learners' procedural knowledge but also declarative knowledge of the target construction. In terms of the skill acquisition approach (DeKeyser, 2007a), transfer between skills is likely to occur through declarative knowledge of rules.

Taking this line of thought further, a logical question is: If participants indeed acquired declarative knowledge associated with the past progressive construction as a result of the experimental treatment, why did their posttest performance prove considerably less successful on the GJT (designed to assess primarily declarative knowledge) compared to the written description test (likely to generate less reliance on declarative knowledge)? Besides the fact that the GJT and written description test required retrieval under distinct operating conditions (see above), an additional, related explanation for the participants' inferior performance on the GJT might lie in the nature of whatever declarative rule the learners might have acquired. The rule might have been too narrow in scope to tackle the GJT items, which required the use of the past progressive in slightly more varied contexts than the written description test. For example, the GJT items, unlike the written description test, included verb forms in the first person and did not always contain reference to a particular place. DeKeyser (2007c) argues that, to facilitate rule application with a wider scope, "what is needed is repeated rule retrieval under increasingly demanding task conditions after initial proceduralization" (p. 293), a condition not met in the current experiment.

Finally, it is worth pointing out in relation to research question 1 that the results of the study appear to support the view that recasts can facilitate the encoding of new declarative knowledge. Participants seemed to have no prior knowledge of the past progressive construction, declarative or procedural, in this study. Due to receiving recasts, however, they showed an increase in declarative knowledge, which, as mentioned above, was evident in the fact that at least part of the knowledge they gained from the oral treatment was usable on the written outcome measures. Hence, similarly to Long et al.'s (1998) study, the findings indicate that recasts can promote the creation of novel L2 representations. This might be so, following Long's (2007) and Lyster's (2004) reasoning, because recasts are potential sources of both positive and negative input.

The second research question examined the extent to which PSTM and complex working memory capacity are related to the observed effectiveness of recasts on different types of outcome measures. For the recast group, a number of significant correlations were found between the working memory and developmental measures: the GJT and written description gain scores were positively correlated with the RS results; and the gain scores on the

oral description test demonstrated positive correlations with the DS and NWS. That is, participants with high complex verbal working memory capacity (i.e., those who scored high on the RS test) were likely to show more substantial improvement on the written tests. On the other hand, participants with high PSTM (i.e., learners who achieved high DS and NWS) tended to exhibit greater growth on the oral test. For the nonrecast group, none of the correlations were found to be significant between the working memory and pretest–posttest measures.

One way of accounting for these findings is to hypothesize that, depending on their PSTM and complex verbal working memory capacity, participants engaged in different types of learning processes to various degrees. High PSTM capacity, for example, is likely to have enabled learners to maintain the information in recasts in short-term memory longer, resulting in greater likelihood that, based on the recasts, LTM traces were created upon which data-driven learning processes could operate, facilitating the proceduralization of emerging L2 knowledge (N. Ellis, 2005). This could explain why learners who scored high on the PSTM tests performed better on the oral posttests, as the oral assessments were likely to draw more extensively on the application of procedural knowledge. In contrast, learners with high complex working memory capacity, due to their greater ability to focus, divide, and switch attention among various task demands, might have been more apt at allocating conscious attention to recasts. As a result, they were probably better able to develop metalinguistic, declarative knowledge based on the grammatical information contained in the feedback (N. Ellis, 2005). Arguably, this superior ability to derive metalinguistic information from recasts allowed high-RS learners to achieve greater gains on the GJT and written description task, as these tests were more conducive to the use of declarative knowledge than the oral assessments were.

An alternative, or additional, explanation of the results is related to the nature of the outcome measures rather than to differences in learning processes. It is possible that the three types of tests differentially favored the deployment of learner knowledge as a function of complex working memory and PSTM capacity. The GJT and written descriptions might have proved easier for high-RS learners, as they entailed considerable use of literacy skills. It is well attested that complex verbal working memory plays an important role in the acquisition of both first (Gathercole, Brown, & Pickering, 2003) and second language literacy (Harrington & Sawyer, 1992; Kormos & Sáfár, 2008; Leeser, 2007; Walter, 2004), probably because literacy-related activities typically require learners to hold verbal information in phonological memory while performing other cognitive processes (Kormos & Sáfár, 2008). In the present study, for example,

learners had to hold the processed sentences in memory as well as judge the grammaticality of the sentences when completing the GJT. Likewise, in the written description task, participants, being unpressured for time, probably had the opportunity to reread and check their written production, which is likely to have posed memory demands not only in terms of storage but also in terms of other higher-order, conscious cognitive activities.

Although complex working memory capacity has also been found to predict L2 speaking ability in previous research (Kormos & Sáfár, 2008), performance on the oral assessments might have been relatively little dependent on it in the present study. Given that the oral tasks predominantly elicited the use of the same construction (past progressive), participants might often have been able to retrieve a recently processed chunk from short-term memory when forming their subsequent utterances, which probably reduced the need for conscious encoding and retrieval from memory (Kormos, 2006). It is possible, however, that the oral tasks favored learners with high PSTM. Assuming a similar degree of declarative knowledge, it might have been easier for high-PSTM learners to access their knowledge of the target construction under real operating conditions. It has recently been shown that L2 oral fluency tends to vary as a function of PSTM (Kormos & Sáfár, 2008; O'Brien et al., 2007).

Conclusion

This study has extended existing feedback research by providing further empirical evidence that outcome measures used as dependent variables might influence the results of recast studies. In the present research, oral recasts led to the greatest gains on oral production tests, resulted in considerably less evidence of improvement on a written production test, and engendered the least gains on an untimed GJT. These results were interpreted as suggesting that, while recasts promoted the acquisition of both declarative and procedural knowledge, they were more likely to foster development in procedural knowledge.

The novelty of this study was to demonstrate that working memory can mediate the observed effectiveness of recasts on various outcome measures. As a result of receiving recasts, learners with high RS (i.e., complex working memory capacity) tended to display more growth on the written tests, whereas those with high DS and NWS (i.e., PSTM) showed more substantial improvement on the oral tests. These findings were tentatively explained by potential differences in the extent to which learners engaged in different types of learning processes as a function of their complex working memory and PSTM capacities. Additionally, it was suggested that learners might have been more or less able to

deploy their L2 knowledge on the three types of outcome measures depending on their complex working memory and PSTM abilities.

In evaluating these findings, it is also important to acknowledge the limitations of the study. A major shortcoming lies in the fact that it was hypothesized rather than established that the three types of outcome measures differentially drew on the use of declarative versus procedural knowledge. This inevitably made some interpretations of the results speculative. Another weakness concerns the nature of the RS test used here. While participants were required to recall sentence-final words as well as to answer comprehension questions in an attempt to reduce trade-off effects between storage and processing performance (Waters & Kaplan, 1996), learners were allowed to read the sentences at their own speed, that is, the test was untimed. This, in turn, might have put the complexity of the task at risk (Sagarra, 2007a). A third limitation is that only a single linguistic feature, the past progressive construction in English, was investigated. Therefore, the results may not generalize to other constructions and languages. Replicating the study with constructions that are nontransparent and that involve nonlocal associations between linguistic elements would be especially interesting, as the acquisition of such constructions is likely to entail more conscious, attention-driven analysis (N. Ellis, 2005; Kempe & Brooks, 2008).[5] The target construction in the present study was relatively simple and transparent; thus, it could be argued that L2 learners might have been able to learn it solely from exposure to input, without instruction[6]. A related limitation is that the treatment here primarily elicited obligatory contexts for only one usage linked to the past progressive. Thus, it may have provided learners with a biased representation of the form–meaning mappings associated with this form, reinforcing the general tendency of L2 learners to assume one-to-one correspondences between forms and meanings. A fifth limitation is that participants in this study were Hungarian learners of English, which further constrains the generalizability of the findings to learner populations with other first languages. Finally, learner responses to recasts (Mackey & McDonough, 2006; McDonough, 2005) and characteristics of feedback (Loewen & Philp, 2006; Philp, 2003) were not considered in the analyses here; thus, a follow-up study could examine whether these factors would modulate the results obtained. Despite these limitations, this study has generated some novel insights into the complex links between recasts, working memory, and various L2 outcome measures, as well as some new and potentially fruitful avenues for further exploration.

Revised version accepted 26 March 2010

Notes

1 Importantly, none of the existing recast studies can be claimed to have used pure measures of declarative or procedural knowledge. Nonetheless, in terms of the declarative–procedural continuum, it can be hypothesized whether certain instruments drew more heavily on the application of procedural or declarative knowledge.
2 Note that, although RS tasks continue to be used as a measure of verbal working memory capacity, their reliability and validity have recently been questioned (Waters & Caplan, 2003).
3 It appeared justified to collapse the two recast groups and the two nonrecast groups for the purposes of this research, given that the impact of task complexity was found small compared to the large effects of the variable *recasts* in the original study (see Révész, 2007, 2009).
4 *Average measures* are defined as the average of the logit estimates for all participants in the sample who produced a particular score, whereas *step calibrations* are the difficulties estimated for achieving a particular test score over another. Both measures are expected to increase monotonically as a variable increases in size; disordered average measures and step calibrations suggest that a certain rating scale does not adequately represent the data.
5 As an anonymous reviewer pointed out, given that participants had knowledge of the present progressive construction at the onset of the study, they only had to learn the meaning associated with the auxiliaries "was" and "were" in past progressive contexts. This means that the learning problem was similar to when learners acquire new vocabulary items, not necessitating the learning of nonlocal associations.
6 I am grateful to one of the anonymous reviewers for this comment.

References

Ammar, A., & Spada, N. (2006). One size fits all? Recasts, prompts and L2 learning. *Studies in Second Language Acquisition*, *28*, 543–574.

Baddeley, A. D. (1986). *Working memory*. Oxford, UK: Oxford University Press.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*, 417–423.

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*, 829–839.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. Bower (Ed.), *The psychology of learning and motivation* (pp. 47–90). New York: Academic Press.

Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning, and use.* Malden, MA: Blackwell.

Bond, T. G., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Cheung, H. (1996). Non-word span as a unique predictor of second-language vocabulary learning. *Developmental Psychology*, *32*, 867–873.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

DeKeyser, R. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, *19*, 195–221.

DeKeyser, R. (2007a). Situating the concept of practice. In R. DeKeyser (Ed.), *Practicing in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 1–18). New York: Cambridge University Press.

DeKeyser, R. (2007b). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 97–113). Mahwah, NJ: Erlbaum.

DeKeyser, R. (2007c). The future of practice. In R. DeKeyser (Ed.), *Practicing in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 287–304). New York: Cambridge University Press.

DeKeyser, R. (2009). Cognitive-psychological processes in second language learning. In M. H. Long & C. J. Doughty (Eds.), *Handbook of second language teaching* (pp. 119–138). Oxford, UK: Blackwell.

Doughty, C. J. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 206–257). Cambridge, UK: Cambridge University Press.

Doughty, C. J. (2003). Instructed SLA: Constraints, compensation, and enhancement. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 256–310). Oxford, UK: Blackwell.

Doughty, C. J., & Varela, E. (1998). Communicative focus on form. In C. J. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 114–138). New York: Cambridge University Press.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, *18*, 91–126.

Ellis, R. (2004). The definition and measurement of L2 explicit knowledge. *Language Learning*, *54*, 227–275.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, *27*, 305–352.

Ellis, N. C., & Schmidt, R. (1997). Morphology and longer distance dependencies: Laboratory research illuminating the A in SLA. *Studies in Second Language Acquisition*, *19*, 145–171.

Ellis, N. C., & Sinclair, S. (1996). Working memory in the acquisition of vocabulary and syntax. *Quarterly Journal of Experimental Psychology*, *49*, 234–250.

Ellis, R. (1995). Interpretation tasks for grammar teaching. *TESOL Quarterly*, *29*, 87–103.

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

Ellis, R. (2007). The differential effects of corrective feedback on two grammatical structures. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 339–360). Oxford, UK: Oxford University Press.

Ellis, R., Basturkmen, H., & Loewen, S. (2001). Learner uptake in communicative ESL lessons. *Language Learning*, *51*, 281–318.

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explict corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, *28*, 339–368.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*, 354–375.

Gathercole, S. E. (1999). Cognitive approaches to the development of short-term memory. *Trends in Cognitive Sciences*, *3*, 410–419.

Gathercole, S. E., Brown, L., & Pickering, S. J. (2003). Working memory assessments at school entry as longitudinal predictors of National Curriculum attainment levels. *Educational and Child Psychology*, *20*, 109–122.

Gathercole, S. E., Hitch, G. J., Service, E., & Martin, A. J. (1997). Phonological short-term memory and new word learning in children. *Developmental Psychology*, *33*, 966–979.

Goldschneider, J., & DeKeyser, R. (2001). Explaining the "natural order of L2 morpheme acquisition" in English. A meta-analysis of multiple determinants. *Language Learning*, *51*, 1–50.

Han, Z-H. (2002). A study of the impact of recasts on tense consistency in L2 output. *TESOL Quarterly*, *36*, 543–572.

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skills. *Studies in Second Language Acquisition*, *14*, 25–38.

Jeon, K. S. (2007). Interaction-driven L2 learning: Characterizing linguistic development. In A. Mackey (Ed.), *Conversational interaction and second language acquisition: A series of empirical studies* (pp. 379–403). Oxford, UK: Oxford University Press.

Juffs, A. (2004). Representation, processing and working memory in a second language. *Transactions of the Philological Society*, *102*, 199–226.

Kempe, V., & Brooks, P. J. (2008). Second language learning of complex inflectional systems. *Language Learning*, *58*, 703–746.

Kim, J. H., & Han, Z-H. (2007). Recasts in communicative EFL classes: Do teacher intent and learner interpretation overlap? In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 269–297). Oxford, UK: Oxford University Press.

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, N. J.: Erlbaum.

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, *11*, 261–271.

Krashen, S. (1985). *The Input Hypothesis: Issues and implications*. London: Longman.
Leeman, J. (2007). Feedback in L2 learning: Responding to errors during practice. In R. DeKeyser (Ed.), *Practicing in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 111–137). New York: Cambridge University Press.
Leeser, M. J. (2007). Learner-based factors in L2 reading comprehension and processing grammatical form: Topic familiarity and working memory. *Language Learning*, *57*, 229–270.
Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*, 309–365.
Lightbown, P. M. (2007). Transfer appropriate processing as a model for classroom second language acquisition. In Z-H. Han (Ed.), *Understanding second language process* (pp. 27–44). Clevedon, UK: Multilingual Matters.
Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.
Linacre, J. M. (2006). *FACETS*, Version 3.61 [Computer program]. Chicago: MESA Press.
Loewen, S., & Nabei, T. (2007). Measuring the effects of oral corrective feedback on L2 knowledge. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 361–377). Oxford, UK: Oxford University Press.
Loewen, S., & Philp, J. (2006). Recasts in the adult L2 classroom: Characteristics, explicitness and effectiveness. *Modern Language Journal*, *90*, 536–556.
Long, M. H. (2007). Chapter 4: Recasts: The story so far. In M. H. Long, *Problems in SLA* (pp. 75–116). Mahwah, NJ: Erlbaum.
Long, M. H., Inagaki, S., & Ortega, L. (1998). The role of implicit negative evidence in SLA: Models and recasts in Japanese and Spanish. *Modern Language Journal*, *82*, 357–371.
Lyster, R. (1998). Recasts, repetition, and ambiguity in L2 classroom discourse. *Studies in Second Language Acquisition*, *20*, 51–81.
Lyster, R. (2004). Different effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition*, *26*, 399–432.
Lyster, R., & Mori, H. (2006). Interactional feedback and instructional counterbalance. *Studies in Second Language Acquisition*, *28*, 321–341.
Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in Second Language Acquisition*, *19*, 37–66.
Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*, 265–302.
Mackey, A. (Ed.). (2007). *Conversational interaction in second language acquisition: A collection of empirical studies*. Oxford, UK: Oxford University Press.
Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning*, *60*, 501–533.

Mackey, A., Gass, S. M., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*, *22*, 471–497.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 407–452). Oxford, UK: Oxford University Press.

Mackey, A., & McDonough, K. (2006). Responses to recasts: Repetitions, primed production, and linguistic development. *Language Learning*, *56*, 693–720.

Mackey, A., & Philp, J. (1998). Conversational interaction and second language development: Recasts, responses, and red herrings? *Modern Language Journal*, *82*, 338–356.

Mackey, A., Philp, J., Egi, T., Fujii, A., & Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback and L2 development. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 181–209). Amsterdam: John Benjamins.

Masoura, V. M., & Gathercole, S. E. (1999). Phonological short-term memory and foreign language learning. *International Journal of Psychology*, *34*, 383–388.

Masoura, V. M., & Gathercole, S. E. (2005). Phonological short-term memory skills and new word learning in young Greek children. *Memory*, *13*, 422–429.

McDonough, K. (2005). Identifying the impact of negative feedback and learners' responses on ESL question development. *Studies in Second Language Acquisition*, *27*, 79–103.

Miyake, A., & Friedman, D. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy & L. E. Bourne, Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339–364). Mahwah, NJ: Erlbaum.

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

Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. In C. J. Doughty & M. H. Long, (Eds.), *Handbook of second language acquisition* (pp. 716–761). London: Blackwell.

Nuevo, A. (2006). *Task complexity and interaction: L2 learning opportunities and interaction*. Unpublished doctoral dissertation. Georgetown University, Washington, DC.

O'Brien, I., Segalowitz, N., Collentine, J., & Freed, B. (2006). Phonological memory and lexical narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics*, *27*, 377–402.

O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language fluency gains in adults. *Studies in Second Language Acquisition*, *29*, 557–582.

Papagno, C., & Vallar, G. (1992). Phonological short-term memory and the learning of novel words: The effect of phonological similarity and item length. *The Quarterly Journal of Experimental Psychology*, *44*, 47–67.

Philp, J. (2003). Constraints on "noticing the gap": Non-native speakers' noticing of recasts in NS-NNS interaction. *Studies in Second Language Acquisition*, *25*, 99–126.

Politt, A., & Hutchinson, C. (1987). Calibrated graded assessment: Rasch partial credit analysis of performance in writing. *Language Testing*, *4*, 72–92.

Racsmány, M., Lukács, Á., Németh, D., & Pléh, Cs. (2005). A verbális munkamemória magyar nyelvû vizsgálóeljárásai [Verbal working memory testing procedures in Hungarian]. *Pszichológiai Szemle*, *60*, 479–506.

Révész, A. (2007). *Focus on form in task-based language teaching: Recasts, task complexity, and L2 learning*. Unpublished doctoral dissertation. Teachers College, Columbia University, New York.

Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition*, *31*, 437–470.

Révész, A., & Han, Z-H. (2006). Task content familiarity, task type, and efficacy of recasts. *Language Awareness*, *3*, 160–179.

Révész, A., Sachs, R., & Mackey, A. (2011). Task complexity, uptake of recasts, and second language development. In P. Robinson (Ed.), *Researching second language task complexity: Task demands, language learning and language performance* (pp. 203–236). Amsterdam: John Benjamins.

Robinson, P. (2005). Aptitude and second language acquisition. *Annual Review of Applied Linguistics*, *25*, 45–73.

Sagarra, N. (2007a). From CALL to face-to-face interaction: The effect of computer-delivered recasts and working memory on L2 development. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 229–248). Oxford, UK: Oxford University Press.

Sagarra, N. (2007b). Working memory and L2 processing of redundant grammatical forms. In Z-H. Han (Ed.), *Understanding second language process* (pp. 133–147). Clevedon, UK: Multilingual Matters.

Sawyer, M., & Ranta, L. (2001). Aptitude, individual differences and instructional design. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 319–353). New York: Cambridge University Press.

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

Schwartz, B. D. (1993). On explicit and negative data effecting and affecting competence and linguistic behavior. *Studies in Second Language Acquisition*, *15*, 147–163.

Service, E., & Kohonen, V. (1995). Is the relation between phonological memory and foreign language learning accounted for by vocabulary acquisition? *Applied Psycholinguistics*, *16*, 155–172.

Sheen, Y. (2004). Corrective feedback and learner uptake in communicative classrooms across instructional settings. *Language Teaching Research*, *8*, 263–300.

Sheen, Y. (2006). Exploring the relationship between characteristics of recasts and learner uptake. *Language Teaching Research*, *10*, 361–392.

Sheen, Y. (2007). The effect of corrective feedback, language aptitude and learner attitudes on the acquisition of English articles, In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 301–322). Oxford, UK: Oxford University Press.

Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, *25*, 293–321.

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

Trofimovich, P., Ammar, A., & Gatbonton, E. (2007). How effective are recasts? The role of attention, memory, and analytical ability. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 171–195). Oxford, UK: Oxford University Press.

Walter, C. (2004). Transfer of reading comprehension skills to L2 is linked to mental representations of text and to L2 working memory. *Applied Linguistics*, *25*, 315–339.

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology*, *49A*, 51–79.

Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments & Computers*, *35*, 550–564.

Williams, J. N., & Lovatt, P. (2003). Phonological memory and rule learning. *Language Learning*, *53*, 67–121.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Facets Summary for the Grammaticality Judgment Task.
Figure S2. Facets Summary for the Written Description Task.
Figure S3. Facets Summary for the Oral Description Tasks.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.