Fluency in second language writing: the effects of enhanced speed of lexical retrieval

Snellings, P.J.F.

Citation for published version (APA):
VALIDATING A TEST OF L2 WRITTEN LEXICAL RETRIEVAL: A NEW MEASURE OF FLUENCY IN WRITTEN LANGUAGE PRODUCTION

ABSTRACT
Lexical retrieval is an essential sub-process of language production and crucial in fluent writing and speaking. In this study we discuss a new measure of the speed of written lexical retrieval in a second language, the Written Productive Translation Task (WPTT). In contrast to Picture Naming tasks, the WPTT is not restricted to concrete nouns and verbs that can be depicted by pictures. The WPTT can also test combinations of words. Moreover, the WPTT can be used in a classroom setting because individual testing is not required. Using Messick’s 1989 framework, we investigated whether the WPTT is an appropriate measure of the construct of written lexical retrieval. In a multiple regression approach we examined the relationship between the WPTT and another test of the same construct, a written Picture Naming task. In addition, we looked at the relationship with other related constructs such as lexical access and orthographic encoding. Results suggest that the WPTT is both a reliable and valid measure of written lexical retrieval. When corrected for typing speed, the WPTT is related to the Picture Naming task and Orthographic Encoding. Finally, the results show that the relationship between the WPTT and the Lexical Decision Task is not very strong. This study makes clear that Lexical Decision Tasks do not constitute a valid substitute for measuring written lexical retrieval. At the same time it underlines the advantages of using a test specifically developed for measuring speed of written lexical retrieval like the WPTT.

INTRODUCTION
Lexical retrieval is an essential subskill of fluent language production. Through the encoding (morphologically, phonologically and phonetically) of selected lexical concepts, lexical material can subsequently be articulated (Levelt, 1989; Levelt et al., 1998). Fluency and sufficient speed in lexical retrieval is important for speaking to hold the floor. Non-fluent speech is effortful and demands attention, resulting in many hesitations because speakers are looking for the right words or have problems in combining them into utterances (Schmidt, 1992). In the case of writing, fluency is especially important in contexts with time restrictions imposed. Moreover, too much attention dedicated to retrieving words may leave little working memory free to attend to generating detailed content and organized discourse (Kellogg, 1999; McCutchen, 2000; McCutchen et al., 1994). Because second language writers lack fluency, writing in a second language (L2) can be a very effortful process for beginning L2 learners (Chenoweth & Hayes, 2001), and it is therefore particularly important to get an insight into the process of lexical retrieval in written second language production. Furthermore, another study (Snellings et al., submitted, Chapter 4) has shown that enhancing lexical retrieval effectively increases production in actual writing. The implication of this finding is that in teaching, attention should be focused on speed of lexical retrieval as well. Simply teaching words until their meaning is known may not be sufficient. Only when students can retrieve words effortlessly, teaching has succeeded in the sense that students will be able to use the words productively. Clearly, to establish whether speed of lexical retrieval has increased, a reliable and valid measure is essential. To this aim, a new instrument to measure L2 lexical retrieval in written language production was developed: the Written Productive Translation Task (WPTT). The goal of the current study is to investigate the validity of this test and to establish what meaning we can attribute to the scores generated by it. Our aim was to measure

lexical retrieval in an educational context. Therefore, the WPTT was designed to be applicable in a classroom setting. So far, vocabulary research has mainly focused on the amount and breadth of word knowledge. Some studies have looked at the productive use of vocabulary (Henriksen, 1999; Laufer, 1998) but there has been no systematic focus on the speed with which this vocabulary knowledge can be retrieved. Although Meara (1996) has made a plea to include “how automatically the items in a lexicon could be accessed” into a theoretical framework of vocabulary knowledge, empirical investigations into the acquisition of this skill are still lacking. Furthermore, existing tests (e.g. Meara, 1997) tap lexical access in reading and do not specifically focus on lexical retrieval in written production. In a first study into the acquisition of speed of L2 written lexical retrieval (Snellings et al., 2002, Chapter 2), we have demonstrated that L2 written lexical retrieval as measured by the WPTT can be effectively enhanced through a short but focused instruction. Responsiveness to experimental treatment provides significant support for the WPTT’s validity but in view of the importance of this test for measuring fluency in L2 written language production, additional research is necessary into its properties and what the interpretation of its scores should be. Additional research is all the more relevant to determine whether the construct (“the concept or characteristic that a test is designed to measure” (AERA/APA/NCME, 1999)) the WPTT intends to measure is sufficiently different from other constructs such as lexical access and orthographic encoding. In the current study we took up these issues by comparing the scores on the WPTT with those obtained on other speed measures tapping related constructs.

Previously available methods

To measure speed of lexical retrieval for speaking Picture Naming tasks (PN) are commonly used. In these tasks a picture is presented and the word denoting the picture has to be named as fast as possible (for an example see the Method section). Reaction times (RTs) are then recorded on the basis of the voice onset time (VOT). Glaser’s (1992) research into picture naming focused on the cognitive structures and processes that underlie speaking. It has made clear that assuming memory to be divided into a semantic system and a verbal system, picture naming always contacts the semantic code and requires access to the meaning of the words. Glaser’s discussion is relevant for writing as well, as access to the meaning of a word can also be assumed to be the first step in the retrieval process in writing. Unfortunately, the sort of words that can be adequately and non-ambiguously depicted by pictures is limited to concrete nouns and verbs. Another drawback is the fact that it is not possible to depict combinations of words (e.g. “a beautiful waterfall”) or collocations. This is a serious problem as the importance of collocations for fluent language production, both for reducing cognitive processing and as a production strategy, has been stressed repeatedly in the literature (Weinert, 1995; Wray & Perkins, 2000).

Another type of measure from psycholinguistic research, the Lexical Decision Task (LDT), can be used to measure speed of lexical access. Lexical access as measured by this task is theoretically related to lexical retrieval for writing because it involves access to the written form. The LDT requires learners to determine as fast as possible whether a word presented on the screen is an existing word in the target language. Subjects respond by typing a Yes or No response (for an example see the Method section). As Harley (1995) shows, there is some controversy about whether the LDT measures subject strategies in addition to lexical access. It is argued that the LDT is more sensitive to strategic factors than naming tasks, partly because apart from accessing the lexicon, a binary response should be given; deciding between word-nonword. Nevertheless, lexical access as measured by the LDT is assumed to be related to lexical retrieval. From theoretical accounts of the bilingual lexicon (Kroll & Stewart, 1994) it can be deduced that lexical access is similar to lexical retrieval but not identical as the process develops in a different direction; from form to meaning (lexical
access) instead of from meaning to form (lexical retrieval). Finally, in written lexical retrieval a writer needs to access not only the word itself but must also retrieve its phonetic form and, crucial in writing, its orthographic form. As we are interested in lexical retrieval for writing purposes, retrieval of the sound of a word and its orthography are part of the construct under investigation. Despite these theoretical differences between lexical retrieval and lexical access, and despite the fact that the LDT aims to measure lexical access, the LDT has been used (presumably for want of a better method) for measuring lexical retrieval (McCutchen et al., 1994). We would like to stress that because the constructs of lexical access and lexical retrieval are theoretically distinct, the assumption that the LDT can also measure lexical retrieval has to be validated empirically.

An existing measure that inspired the development of the WPTT specifically taps lexical retrieval and has a tradition in bilingual research. This oral translation task (De Groot et al., 1994) requires the speeded naming of either the translation of a native language word into the L2 or the other way round. However, as our focus is on written lexical retrieval, the test should require retrieval of the written form including its orthography. Therefore, these oral tasks are not ideal. In addition, due to the fact that the spoken answer has to be recorded the oral translation task cannot be readily implemented in classroom research. Background noises distort the recordings and individual testing is required. Consequently, we had to develop a different task.

Clearly, in view of the theoretical claims made about lexical retrieval and its implications for fluent writing, there is a need for a measure that takes into account the specific characteristics of written lexical retrieval and represents the underlying construct better. In this study we will have a detailed look at such a test to see how it relates to related constructs as speed of lexical access and speed of orthographic encoding, and whether it resembles scores on a picture naming task that also taps speed of lexical retrieval.

A new written measure of L2 lexical retrieval
In a writing task, learners have to retrieve both words and combinations of words in a specified context. We developed a new measure of L2 written lexical retrieval, the Written Productive Translation Task (WPTT), that aims to tap the efficiency of this retrieval process in L2 writing. To elicit a specified set of lexical items we have used native language words (for an example see the Method section). Once the intended concept is clear to the writer, the process of retrieving the words from memory and writing them down in context is similar to the process involved in actual writing. As pointed out, the idea of the WPTT was based on the oral translation tasks used in psycholinguistic research. Similar to these “forward translation tasks” (De Groot et al., 1994), learners have to translate from their native language into the L2. Because lexical retrieval for writing involves retrieval of the orthography, the answers in the WPTT have to be written down. In contrast to picture naming tasks, the stimuli do not have to be restricted to concrete nouns and verbs. In addition, it is possible to test combinations of words. Finally, the written format of the WPTT also makes it possible to test a whole class of students at the same time without the need for laboratory conditions.

In our validation of the test we adopt Messick’s (1989) “unified validity” framework to look at the validity of the test from different angles. In this view, that has been adopted by the measurement profession (AERA/APA/NCME, 1999; Bachman, 1990), different kinds of validity evidence are not alternatives but supplement each other in assessing the unifying concept of construct validity. Construct validity involves an integration of any evidence on what the interpretation or meaning of the test scores is for a particular purpose. In the following sections we will discuss multiple types of validity evidence. We will first discuss the content of the test, followed by its internal structure and finally we will focus on the interrelations of the test scores with other variables.
Content relevance

Because it is possible to use all parts of speech in writing, a restriction to concrete nouns and verbs is inadequate. In contrast to picture naming tasks, the format of the WPTT allows for the use of all parts of speech and is therefore a better representation of the words used in actual written language production. As pointed out before (Weinert, 1995; Wray & Perkins, 2000), fluent writing requires not only the retrieval of words but also of word combinations and collocations. As a result, it is crucial for the representation of the construct that the WPTT not only includes single words but also word combinations. In the picture naming task, a test of lexical retrieval in speaking, Huttenlocher and Kubicek (1983) distinguish four components: visual processing, activation of the concept, retrieval of the form and articulation. They conclude that frequency effects, the finding that frequent words can be named faster than infrequent words, originate in form retrieval. As our current aim is to measure the speed of retrieving lexical material from memory or “personal frequency” (how familiar a learner is with a specific lexical item), form retrieval should be part of the test. Therefore, in the case of the WPTT retrieval of the English form is required after the L1 form has activated the relevant concept. Furthermore, as Daneman and Green (1986) pointed out, processing skill differences in working memory are most likely task specific. They demonstrated that efficiency at sentence production processes rather than efficiency at sentence comprehension processes determines working memory capacity when individuals have to produce vocabulary in context. In addition, DeKeyser (2001) has argued that understanding sentences in a language quickly does not mean that one can produce them fast. It follows that to tap lexical retrieval it is essential to test productive processes. In view of task specificity, a written test format seems more appropriate than a spoken format to test the behavior relevant to writing. By using a productive test in a written format tapping form retrieval of different parts of speech, either on their own or in combinations, we have ensured that the content of the WPTT covers the content domain of written language production.

Structural component of construct validity

As pointed out by Messick (1989, p.43), ‘the degree of homogeneity in the test should be commensurate with this characteristic degree of homogeneity associated with the construct’. At this stage, we assume that the construct of lexical retrieval is unidimensional and therefore implies a single dimension of behavior, tapping the same process. Consequently, when looking at its internal structure, we expect a high internal consistency for measuring inter-individual skill differences. From a psychometric point of view this means that the test has a high reliability.

External component of construct validity

Loevinger (1957, p.643) already argued that “the nature of the outside variables showing varying degrees of interrelationship with the test score are relevant to proposed interpretations”. Both Loevinger and Messick (1989, p.45) call this the external component of construct validity. To investigate the relationships with other tests we also have to look at the internal consistency of these other tests, because using these tests as indicators is conditional on high reliabilities of these tests. As both the newly developed WPTT and the Picture Naming (PN) tasks aim to measure the same construct of speed of lexical retrieval, we expect high correlations. Because the methods used to tap lexical retrieval differ, naming pictures versus translating words, high correlations would constitute corroborative evidence that regardless of method, the same construct is tapped.

In contrast, the theory states that although related, lexical access and lexical retrieval are distinct, involving a different order of the underlying processes. Moreover, Lexical
retrieval for writing involves the fast application of phonetic and orthographic knowledge. In other words, there are two sub processes in lexical retrieval; phonological retrieval and orthographic retrieval. Consequently, as the order of the processes is different and lexical access does not involve orthographic retrieval, the relationship between the WPTT and the LDT is expected to be less strong than that between the WPTT and the written PN task, both productive tasks that require orthographic retrieval.

As orthographic encoding is an essential sub process of lexical retrieval for writing we expect a relationship between an Orthographic Encoding task (OE, for an example see the Method section) and the WPTT. In the WPTT, students have to spell out their answers as fast as possible. Consequently, we expect the WPTT scores to also express speed of spelling, which we tap with the OE speed measure. On the other hand, if the WPTT truly measures the construct of lexical retrieval, the correlation between the WPTT and PN tasks, both productive and tapping phonological as well as orthographic retrieval, should be higher than that between the WPTT and the OE, as the latter only tests the orthographic sub process of written lexical retrieval. In contrast, if the correlation between the WPTT and the OE were higher, this would imply that orthographic encoding (orthographic retrieval) is the main process in the new test of lexical retrieval. As the construct of lexical retrieval not only involves retrieving the orthography but also retrieving the word itself (phonological retrieval) this would mean that the new test is restricted to measuring just one of the sub processes of lexical retrieval and not the other (an instance of construct under representation).

RESEARCH QUESTIONS

It is evident from the discussion above that we need to investigate the scores on the written productive translation test (WPTT). We will study its construct validity from various angles. The current focus will be on the WPTT’s relationship with other theoretically related measures (Messick’s external component). We will study the relationship with a written Picture Naming task, a Lexical Decision Task and an Orthographic Encoding task. We will also look into its internal structure (Messick’s structural component of construct validity) by examining test reliability.

Research questions:
1. What is the internal structure of the WPTT?
2. What is the relationship between the WPTT and another test of written lexical retrieval (a Picture Naming task), a test of lexical access (a Lexical Decision Task) and a test of orthographic encoding (Orthographic Encoding)?

METHOD

Participants
In this study 109 Dutch students took part, all in their third year of secondary education (grade 9, ages 14-15). Students were sampled from four different classes, two classes from lower general secondary education (n=51) and two classes from pre-university education (n=58). All students had received at least two full years of formal English tuition (2-3 lessons a week) at secondary education level, and a two year introductory course at the elementary level focusing mainly on basic oral communication skills in English (maximum 1 lesson per week).
Instruments

In the following section we will discuss the tasks used in this study, beginning with the WPTT. We will then discuss the other tasks; the Orthographic Encoding task, Picture Naming and the Lexical Decision Task. Finally, we will describe the Typing task, a task intended to control for differences in typing speed but that could also function as an indication of general reaction time. All stimuli were presented on laptop screens in a classroom setting. The sequence of events on each trial was as follows: prior to the stimulus, a fixation stimulus (an asterisk) appeared for 1s. Immediately after it disappeared, the stimulus appeared and remained on the screen until a response was given. Ten practice items preceded the target items in the WPTT, OE and PN, 20 in the case of the LDT.

Written Productive Translation Task (WPTT)

In speed tests it is customary to take into account only the speed of correct responses and therefore we aimed at as many correct responses as possible by selecting words not too difficult for the students. According to Wolter (2001), word-frequency ratings, even if based on large corpora, cannot give conclusive evidence in predicting which words are known or not known by a particular individual. Moreover, most larger corpora consist of native speaker data and are therefore not applicable to second language learners. In view of these considerations we did not use frequency measures but chose words that could be elicited from a similar group of students in a writing task. A pilot study was conducted in which cartoon-based stories were written by students of the same age and similar levels of English ability as in the current study. A selection of the words used by the students was made. The words had to have straightforward translations in Dutch. We selected only words that could be assumed, on the basis of the textbooks used in the Dutch education system in grade 7-8, to be familiar and not too difficult. The words belonged to different parts of speech (e.g. noun, article, determiner, adjective, adverb, verb). Some words were used on their own, others in combination with each other. The procedure is as follows; immediately after the fixation stimulus disappears, the target stimulus appears in the same position and a timer starts. The timer stops when the Return key is pressed. The target stimulus disappears and the fixation stimulus reappears. If no answer has been given, an automatic time out message appears after 20 s. The Dutch words are in capital letters and in simple English carrier sentences to make them unambiguous. Students have to type the English translations of the Dutch words as fast as possible on a computer screen. Before pressing the Return key they can use the arrow, the delete or the backspace key to change their answer.

Example WPTT

last month they VERKOCHTEN the farm

sold (Return key)

As the test was preceded by the OE, we ensured that no WPTT items were part of the OE test. There are 67 items. For the items tested in the WPTT see Appendix B. Responses are saved into log files; we record the time the first key is hit, whether the first key is correct, when the Return key is pressed and whether the response is correct. The maximum number of correct alternatives is four.

Orthographic Encoding (OE)

To measure orthographic encoding speed without the interference of typing ability we chose a format in which students have to choose as fast as possible between two spelling alternatives of the same word in a short sentence or group of words. Immediately after the fixation stimulus disappears, the target stimulus appears near the same position and a timers starts.
The timer stops when “1” or “2” is pressed (which do not appear on the screen). The target stimulus disappears and the fixation stimulus reappears. If no answer has been given, an automatic time out message appears after 5 s. Students are encouraged to hold their index finger next to the “1” and “2”. They have to use their right hand if they are right handed, their left if they are left handed.

Example OE
1) Italian COFEE
2) Italian COFFEE

The orthographical problems were selected to be representative of typical problems for these students. The items involve inflection (e.g. simple past) and problems having to do with conversion of sound into script (especially vowels, e.g. ‘fear/feer’). There is no overlap with other tests. The total number of items is 59. The correct item occurs equally often in the first or the second word position.

Picture Naming task (PN)
We used a written version of the oral picture naming method (Schoonen et al., 2002). An advantage of this format is that it can be implemented in a school setting testing students simultaneously in the classroom. In this method, instead of naming the word aloud, learners have to type the first letter of the noun referring to the person or object depicted. For a similar approach using an electronic writing pad see Bonin (2001). Immediately after the fixation stimulus (in the center of the screen) disappears, a picture appears centered at the fixation stimulus position and a timer starts. The timer stops when a letter key is pressed. The picture disappears and the fixation stimulus reappears. If no answer has been given, an automatic time out message appears after 9 s.

Example PN

Answer: C
(taken from Damhuis & Schoonen, 1993)

Words were selected that could be unambiguously depicted by pictures and that had proved to have high hit rates in previous large-scale research (Schoonen et al., 2002). Words that have
the same first letter in Dutch and English ('wandelen'/'walking') were avoided. We made sure that there was no overlap with the words in the carrier sentences of the WPTT nor the WPTT or OE stimuli. The answers and reaction times (RTs) are registered. The total number of items is 18.

**Lexical Decision Task (LDT)**
Students have to decide as fast as possible whether a letter string is a real word or not. The correctness of answers and the length of reaction times are registered. Immediately after the fixation stimulus disappears, the target stimulus appears at the same position and a timer starts. The timer stops when either the space bar (a Yes response) or the letter N (a NO response) is pressed. The target stimulus disappears and the fixation stimulus reappears. If no answer has been given, an automatic time out message appears after 3 s. Students are encouraged to hold one index finger next to the space bar and the other next to the letter N. To give a “Yes” response they have to use their right hand if they are right handed, their left if they are left handed.

**Example LDT**
repribe (a No response)
bat (a Yes response)

The test consists of 120 letter strings of 3-8 letters. The selection of words and pseudo-words was based on lists of non-cognates and fillers in Woutersen (1997) and checked with a learner’s corpus for Dutch learners of English (Willems & Oud-de Glas, 1990). Half of the words were existing mono-morphomatic well-known words (nouns and adjectives). Half of them were (morphologically correct) pseudo-words. The order in which pseudo-words and real words appeared was randomized. The distribution of the length of morphologically correct pseudo-words was identical to the real words.

We tried to avoid overlap in target items between different tests. However, there was a small overlap of three target items between the LDT and the WPTT. In view of a total of 60 LDT items and given the high reliability of the LDT (see Results section), it is safe to assume that this has a negligible effect on our results.

**Typing task (T)**
In the WPTT and PN tasks students have to either type all letters of the words (WPTT) or the first letter of the word depicted by the picture (PN). Apart from knowledge of the words, their RTs reflects the time it takes to look up the letters on the keyboard. Looking up the letters on the keyboard takes time and students will differ in this ability. The time it takes to respond to the OE and LDT also includes general non-language specific differences in reaction time. To measure typing speed the Typing task was administered. Students are presented with individual letters on a laptop screen and have to type these letters as fast as possible. Immediately after the fixation stimulus disappears, the target stimulus appears at the same position and a timer starts. The timer stops when a letter key is pressed. The target stimulus disappears and the fixation stimulus reappears. If no answer has been given, an automatic time out message appears after 3 s. The total number of items is 42.

**Data collection**
All tests were taken on laptop computers in a classroom setting, administered by the principal researcher and a test assistant. All students took the tests in the same order over two sessions. In establishing the order of test administration we tried to keep overlap at a minimum and therefore took into account that some items also occurred in examples and carrier sentences.
As the large number of words in the English carrier sentences of the WPTT might influence response times of the OE, the OE was administered first, immediately followed by the WPTT. In a second session (two days, four days or a week later) the speed of Typing task, the Picture Naming task and the Lexical Decision Task were administered.

**Scoring**

In all tests two scores are relevant: reaction time (RT) and accuracy. Mean RTs were computed across the “hits” (i.e. correct responses on the items). All incorrect responses were recoded into missing values. Since we are not interested in differences in linguistic knowledge but solely in speed of processing, we only used items with sufficiently high hit rates. Therefore, we had to set cut-off percentages for each test.

In the WPTT words with a hit rate lower than 60% were removed. This percentage takes into account that correction for guessing in this task is not necessary and that the stimuli are more difficult to answer correctly than in the other speed tasks. As a result of this procedure 57 items remained in the analysis out of the initial 67. This relatively large loss of items is an indication of the difficulty of the WPTT for our students. In the final analysis, we also removed two items with low-item rest correlations (<.30). This left us with a final set of 55 items.

Because we believed that the stimuli had to be rather easy, we set a minimum level of 75%, uncorrected for guessing. Therefore, words with a hit rate lower than 87.5% were removed in the LDT and OE. This percentage takes into account a chance level of 50% correct guesses. The correction produces the cut-off limit of 87.5%. In the case of the LDT three items were removed, resulting in a total of 46 items. Sixteen items were removed from the OE, resulting in 43 items in the final analysis. So, we can deduce that the OE task also consisted of a relatively large number of (too) difficult items. For PN and T the criterion for item deletion was set to a hit percentage of 75, taking into account the impossibility of guessing. All items satisfied this criterion.

**Missing values**

There were three kinds of sources for missing values in our study. The first consisted of skipping items by individual students. The second source was incorrect responses. The third one was “invalid” responses: reactions that were incredibly fast or extremely slow. Presumably, these responses do not provide valid information on the students’ skill. Students can accidentally touch the response keys or may be distracted while doing the test in a classroom context.

For all tests too slow answers (more than three standard deviations from the mean) were recoded into a missing value. Reaction times faster than the fastest reactions of a group of six advanced L2 speakers were regarded as invalid. In the PN reaction times faster than 550 ms, in the LDT, OE and T reactions faster than 250 ms were recoded into missing values. Because of its open format and the need to type a whole word or word combination, too fast correct answers were not possible in the WPTT.

To arrive at a full data matrix for all students missing values were estimated. The advantage of this procedure is that mean test scores of individual students will be based on the same set of items, making the analysis more reliable. Moreover, analysis of reliability is seriously hampered if few students are left with valid scores on all words. Missing values were estimated according to the EM-algorithm of SPSS. This is an iterative procedure for estimating missing data that has many advantages to other methods of missing value imputation (Acock, 1997; Hox, 1999). The largest amount of missing data occurred in the WPTT, because of the difficulty of the test. Therefore, we checked our estimation procedure.
for the WPTT scores. We compared the mean score on the basis of raw scores and on the basis of EM scores. Their correlation of .98 shows that the estimation procedure does not dramatically alter the structure of the data.

In order to prevent missing value estimation on the basis of too few data for individual students, we had to set a criterion for a minimal amount of hits. For the WPTT, PN and T students had to have 51% hits or more for each test. With this criterion we took into account the difficulty of the tests and the need to keep as much students as possible in the final analysis. As a result four students were dropped from the analysis. For the LDT and the OE, we required a hit rate of 25% as a minimum without guessing. Therefore, the criterion was set on 62.5% hits, taking into account that the chance for guessing the correct answer is 50%. One student failed to meet this criterion but was already dropped from the analysis because the criterion for the WPTT and PN was not met either. In sum, four students were omitted and because another eight students missed a test session, the total number of students in our final analysis was 97.

**First letter versus whole word reaction time in scoring the WPTT**

As an indication of speed of written lexical retrieval, the WPTT yields two different scores; RT for the first letter or the Return key after typing the whole response. We opted for the time it took students to type their answer and press the Return key. By using the Return key at the end of the word instead of the first letter, we deviated from the procedure most commonly used in oral naming tasks. One reason not to look at the first letter was the fact that many items consisted of more than one word; e.g. “a beautiful waterfall” or “called for help”. Clearly, the time it takes to type the first letter “a” or “c” does not necessarily give a good indication of the time required to retrieve the whole item. In addition, students could type the first letter and hesitate about what had to follow before going on and hitting the Enter key. To take these effects into account we decided to focus on responses of the whole word or combination of words. As an additional check, we analyzed the relationship between the Return key scores and the first letter scores. This yielded a correlation of .85, suggesting that the results of the two scoring procedures are strongly related but not identical. The first letter therefore seems to be a fairly good but imperfect indication of total word retrieval. Presumably, students who start typing the first letter normally continue to type without many hesitations. As a consequence, students who are fast at typing the first letter are generally also fast at pressing the Return key.

**Leniency in scoring the WPTT**

The program automatically generates correct scores; those answers that have both correct orthography and correct inflection. The automatic scoring procedure does not allow for a lenient scoring of orthography and inflection errors. However, we did not want orthographic correctness to play a too important role in the scoring of the WPTT. Given the low level of L2 proficiency of our students, they make numerous errors in orthography and inflection, even though they retrieved the correct words and tried to use correct orthography and inflection. A scoring procedure was developed to establish criteria for retrieval of the intended word forms (details of the procedure can be requested from the authors). A correlational comparison between three possible criteria (automatic, correct inflection only, and lenient) revealed a very high correlation between the different criteria when applied to the items that were part of our final analysis and which were selected under a lenient criterion. The lenient approach is nevertheless preferable as we are interested in speed of lexical retrieval and we are not interested in production errors.
Data analysis
To establish the structural properties of the tests, means, standard deviations and an estimate of internal consistency (Cronbach’s $\alpha$) were calculated. To investigate the relations with other tests we carried out three analyses. First, a correlational analysis using uncorrected scores. Next, a correlational analysis on the basis of residual scores from the regression of the WPTT on the speed of Typing scores. For the other tests we also used residual scores from the regression on the speed of Typing scores. Finally, using a standard regression approach (Tabachnick & Fidell, 1996, p.149) with the Residual Written Productive Translation Task scores (corrected for typing speed) as the dependent variable, we entered our independent variables simultaneously, after correcting them for typing speed as well: Residual Picture Naming task scores, Residual Lexical Decision Task scores and Residual Orthographic Encoding scores.

RESULTS

Structural component of construct validity
We looked at the internal structure of our WPTT to examine the internal consistency. As can be seen in Table 1, the internal consistency of the WPTT was extremely high with an alpha of .95.

<table>
<thead>
<tr>
<th>Test</th>
<th>Number of items</th>
<th>M (in ms)</th>
<th>SD</th>
<th>Cronbach’s $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>WPTT</td>
<td>55</td>
<td>7358.41</td>
<td>1461.70</td>
<td>.95</td>
</tr>
<tr>
<td>T</td>
<td>42</td>
<td>988.33</td>
<td>150.68</td>
<td>.95</td>
</tr>
<tr>
<td>PN</td>
<td>18</td>
<td>1799.40</td>
<td>319.73</td>
<td>.86</td>
</tr>
<tr>
<td>LDT</td>
<td>46</td>
<td>734.06</td>
<td>127.07</td>
<td>.95</td>
</tr>
<tr>
<td>OE</td>
<td>43</td>
<td>1693.82</td>
<td>338.35</td>
<td>.96</td>
</tr>
</tbody>
</table>

* Of the initial 109 students four failed to meet our “minimal hits criterion” (see Missing values section) and another eight students had missing values for one of the tests and were omitted from the analysis by using listwise deletion.

External structure
We carried out several analyses, using both correlational and regression methods. The descriptives in Table 1 show that the tests of the predictor variables have sufficiently high internal consistencies and are therefore a reliable basis for our analyses. All tests have very high consistencies ($\alpha>=.95$), only the Picture Naming test had a slightly lower alpha ($\alpha=.86$) that was nevertheless at a satisfactory level.
Table 2 Pearson product-moment correlations between the Written Productive Translation Task (WPTT) scores and speed of Typing (T), Picture Naming (PN), Lexical Decision Task (LDT) and Orthographic Encoding (OE) scores, n=97.

<table>
<thead>
<tr>
<th>Test</th>
<th>WPTT</th>
<th>T</th>
<th>PN</th>
<th>LDT</th>
<th>OE</th>
</tr>
</thead>
<tbody>
<tr>
<td>WPTT</td>
<td>.60*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>.51*</td>
<td>.51*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PN</td>
<td>.71*</td>
<td>.34*</td>
<td>.48*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDT</td>
<td>.43*</td>
<td>.34*</td>
<td>.47*</td>
<td>.49*</td>
<td></td>
</tr>
<tr>
<td>OE</td>
<td>.53*</td>
<td>.30*</td>
<td>.47*</td>
<td>.49*</td>
<td></td>
</tr>
</tbody>
</table>

*statistically significant (p<.01, a one-tailed test because they are all speed measures so we assume that the correlations are higher than zero).

A first step was to establish the relationship between the WPTT and the Picture Naming (PN), Lexical Decision Task (LDT), Orthographic Encoding (OE) and the Speed of Typing (T) tests, by examining the uncorrected correlations. Table 2 summarizes the results of the correlational analysis. As can be seen in Table 2 all predictor variables correlate positively with the dependent variable WPTT (one-tailed p<.01, a one-tailed test because we assume that the correlations are higher than zero). Moreover, we see that the highest correlation is between the WPTT and the PN (r=.71), followed by that between the WPTT and the speed of Typing (r=.60), WPTT and OE (r=.53) and finally, between the WPTT and the LDT (r=.43).

Using the Fisher r-to-Z transformation (Hays, 1988) we calculated the confidence intervals for these correlations. Fisher showed that confidence intervals can be made from moderately large samples from a bivariate normal population if one uses a particular function of r (see for details Hays, 1988, pg 588-593), rather than the sample correlation coefficient itself. The confidence intervals show that the WPTT is most related to another measure of the same construct of lexical retrieval (the Picture Naming task) and less to a measure of the related construct of lexical access (the Lexical Decision Task). The confidence interval of the Orthographic Encoding test, which is a measure of a sub-skill of written lexical retrieval, overlaps with the PN. The same holds for the confidence interval of the OE and the LDT.

In Table 2 we also see a substantial correlation (r=.60) between the WPTT and the speed of Typing test (T). This is the result of our written format that requires students to type their answers on a laptop computer. The relationship with the typing scores is likely to be due to the typing skills necessary in the WPTT: some of the variance in the WPTT scores is the result of differences in typing speed; the speed of finding the letters on the keyboard. The other test of lexical retrieval, the Picture Naming test, similarly relies on typing skills, but here typing is restricted to finding one letter (r=.51). Speed of typing is slightly related to the other tests as well, presumably because it also measures general differences in reaction speed (with LDT r=.34, with OE r=.30). Because the construct we are interested in is “speed of lexical retrieval”, measuring variance in WPTT scores due to typing skills or general differences in reaction speed is not our objective. In addition, the correlation between the WPTT and the PN may be inflated as both involve speed of typing. As a second step, we therefore carried out an additional correlational analysis, using residual WPTT scores from the regression of the WPTT on the speed of Typing scores to obtain a purer measure of the intended construct of lexical retrieval. To get a more valid picture of the relationships between the tests, we also used residual scores from the regression of all other tests on the speed of Typing scores.
Table 3 Pearson product-moment correlations between the Residual Written Productive Translation Task (WPTT') scores and Residual Picture Naming (PN'), Residual Lexical Decision Task (LDT') and Residual Orthographic Encoding (OE') scores, n=97.

<table>
<thead>
<tr>
<th>Test</th>
<th>WPTT'</th>
<th>PN'</th>
<th>LDT'</th>
<th>OE'</th>
</tr>
</thead>
<tbody>
<tr>
<td>WPTT'</td>
<td>.59*</td>
<td></td>
<td>.45*</td>
<td></td>
</tr>
<tr>
<td>PN'</td>
<td></td>
<td>.31*</td>
<td></td>
<td>.38*</td>
</tr>
<tr>
<td>LDT'</td>
<td></td>
<td></td>
<td>.38*</td>
<td>.43*</td>
</tr>
<tr>
<td>OE'</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*statistically significant (p<.01, a one-tailed test because they are all speed measures so we assume that the correlations are higher than zero).

Table 3 shows that all variables correlate positively with the dependent variable WPTT (one-tailed p<.01, a one-tailed test because we assume that the correlations are higher than zero). As predicted, the correlations have changed after speed of typing has been partialled out. We see that all correlations between the WPTT and the other tests have become lower, indicating that part of the relationships were due to differences in typing speed (the relation between the WPTT and PN) or general differences in reaction speed (the relation between the WPTT and both LDT and OE). Interestingly, even though the correlations have become slightly lower, we still see the same pattern. Once more the relationship between the WPTT and the Picture Naming task seems the strongest, followed by that between the Orthographic Encoding task and the WPTT. Again, the relationship between the Lexical Decision Task and the WPTT appears the weakest. We used the Fisher r-to-Z transformation to calculate the confidence intervals for these correlations. These confidence intervals overlap so that differences in correlations are no longer significant. In sum, once potential bias due to typing speed and general reaction speed had been filtered out, the pattern stays the same but the differences are too small to reach significance.

Table 4 Standard multiple regression of Residual Written Productive Translation Task (WPTT') scores on Residual Picture Naming (PN'), Residual Lexical Decision Task (LDT') and Residual Orthographic Encoding (OE') scores, n=97.

<table>
<thead>
<tr>
<th>Variables</th>
<th>B</th>
<th>β</th>
<th>r² (unique)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PN'</td>
<td>2.060*</td>
<td>0.48</td>
<td>.19</td>
</tr>
<tr>
<td>LDT'</td>
<td>0.797</td>
<td>0.01</td>
<td>.00</td>
</tr>
<tr>
<td>OE'</td>
<td>0.965*</td>
<td>0.27</td>
<td>.05</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.197</td>
<td></td>
<td>R²=.41</td>
</tr>
</tbody>
</table>

*statistically significant (p<.05, two-tailed).

So far, we have looked at the bivariate correlations to establish the relationship between the critical measures in this study. We have shown that after partialling out bias from the WPTT as captured by our speed of Typing test, the predicted relations still hold but the differences have become smaller. On the other hand, if we look once more at Table 2 and 3 we see that, apart from the correlations between the WPTT and the other tests, there is also a considerable intercorrelation between the tests themselves. This means that the tests are not independent from each other because the constructs they tap are related. This makes it difficult to determine which part of the variance in the WPTT they account for. In Table 2 the raw scores on the
Picture Naming (PN) task correlate .48 with the Lexical Decision Task (LDT) and .47 with the Orthographic Encoding (OE) respectively, whereas LDT and OE correlate .49. Once speed of Typing is partialled out (Table 3), the large intercorrelations persist: PN' correlates .38 with both LDT' and OE', LDT' correlates .43 with OE'.

As a third step, we carried out a multiple regression analysis to establish the unique variance of the dependent measure explained by each predictor. In particular, we are interested to see whether the contribution to the explanation of the WPTT’ scores by the PN’ is different from that of the other predictors. Another reason to use a regression approach is to find out how much variance can be uniquely ascribed to the WPTT’ once the predictor variables are taken into account. Using a regression analysis has the additional advantage that regression coefficients are less influenced by restriction of range in the sample than are correlation coefficients.

The predictor variables Residual Picture Naming (PN’), Residual Lexical Decision Task (LDT’) and Residual speed of Orthographic Encoding (OE’) are all entered simultaneously in a standard regression approach. Results of evaluation of assumptions indicated that there was no multicollinearity nor singularity and that assumptions of normality, linearity, homoscedasticity and independence of residuals were satisfied. Table 4 displays the unstandardized regression coefficients (B), intercept, the standardized regression coefficients (β), the semipartial correlations (sr²), and percentage explained variance (R²). R for regression was significantly different from zero, F(3,93)=21.26, p<.001. Only two of the predictors contributed significantly to prediction of the WPTT’ scores, Residual Picture Naming (sr² = .19) and Residual Orthographic Encoding (sr² = .05). Altogether, 41% (36% in the population) of the variability in scores on the WPTT’ was predicted by knowing scores on these two predictors. These results indicate once more that the relationship between the WPTT’ and Picture Naming is strongest, with 19% of the variance in WPTT’ scores uniquely contributed by Picture Naming. In contrast, there is no unique contribution of Lexical Decision scores to the variance in WPTT’ scores. The unique contribution of Orthographic Encoding to the WPTT’ scores, although significant, is relatively small with only 5% unique variance. Finally, 59% of the variance in the WPTT’ scores cannot be explained by the scores on the PN’ and OE’ tasks.

DISCUSSION

Summary
A Written Productive Translation Task (WPTT) was constructed to assess the speed with which lexical material can be retrieved for writing in English. Speed of lexical retrieval is an important component of fluent production and the WPTT makes it possible to measure this process efficiently in a large group in a classroom setting. To gauge the meaning of the WPTT scores we administered tests of related skills to a sample of 109 learners of English as a second language. We included a Picture Naming (PN) task as another test of the same construct, a Lexical Decision Task (LDT) as a measure of the related but presumably distinct process of lexical access, and finally, an Orthographic Encoding (OE) test that taps a sub process of written lexical retrieval. To filter out unwanted bias due to differences in typing skill or general non-language specific differences in reaction time, we also measured speed of Typing (T). All tests exhibited high internal consistencies and the results from the correlational analysis showed a strong association between the WPTT scores and the PN scores, indicating that the WPTT taps the same construct of written lexical retrieval. In contrast, the LDT scores which tap the related construct “lexical access” are less strongly associated with the WPTT scores. This outcome is consistent with theories of the mental
lexicon, in which a distinction is made between the processes to obtain meaning from lexical input (lexical access) and the related process of generating lexical output on the basis of intended meaning (lexical retrieval). In addition, we saw that the scores on the OE task are also related to the WPTT scores yet this correlation is somewhat lower than that between the WPTT and the PN scores. This fits in with the idea that orthographic encoding is a sub-process of written lexical retrieval yet does not cover the whole construct of written lexical retrieval. If we take the confidence intervals into account we see that the difference between the WPTT-PN correlation and the WPTT-LDT correlation is significant. The difference between the WPTT-OE and the WPTT-PN correlation is not significant and neither is that between the WPTT-OE and the WPTT-LDT correlation.

The correlational analysis also showed that speed of Typing correlated substantially with all other tests. The relation was strongest with the WPTT and the PN, two measures that both involve typing letters. Speed of Typing is no essential part of the construct of written lexical retrieval but is a consequence of the computerized format chosen. For this reason, we filtered out variance due to speed of typing differences by using the residual scores from the regressions of the individual tests on speed of Typing in an additional correlational analysis. The correlations between these residual scores (PN', LDT' and OE') and the residual WPTT scores (WPTT') were lower but we still saw the same pattern. After removing possible speed of typing bias, the correlation between the WPTT' and PN' was once more highest, followed by that between WPTT' and OE' and the smallest correlation held between the WPTT' and LDT'. Nevertheless, even though the differences are in the predicted direction, the overlap in confidence intervals indicated that these correlations were not statistically different. In addition, the correlational analyses yielded high intercorrelations between PN, LDT and OE, as well as between PN', LDT' and OE'. Therefore, to filter out shared variance due to other factors than the construct involved and to look at the unique contribution of each test once the contribution of other test scores had been taken into account, we carried out a multiple regression analysis with residual scores. These results show that two predictors explained 41% of the variance in WPTT's scores and that there was a large and significant unique contribution of PN'. The unique contribution of OE' was also significant but rather small. Crucially, the unique contribution of LDT' was non-significant. Coupled with the lower correlations between the WPTT scores and the LDT scores this implies that the empirical results corroborate the view that lexical access and lexical retrieval are distinct skills. In addition, the analysis showed that the OE' scores only accounted for a small part of the explained variance, even though the correlation between the WPTT' and the OE' task was .45. This suggests that the WPTT' scores are a good representation of the construct of written lexical retrieval and involve more than getting fast access to orthographic information.

Finally, although the PN' scores accounted for a considerable part of the explained variance on the WPTT', there is still 59% variance that cannot be explained by the combination of scores on the PN' and OE' tasks.

**Implications**

Using the framework advanced by Messick (1989) we have started gathering evidence to meet professional standards for test validation (AERA/APA/NCME, 1999). We have demonstrated that the Written Productive Translation Task (WPTT) can effectively tap the construct of written lexical retrieval as evidenced by its strong relationship to the scores on the Picture Naming task scores. Crucially, the WPTT has the additional advantage of being able to tap aspects of the construct that Picture Naming tasks cannot elicit. Due to its format the WPTT is not restricted to lexical items (concrete nouns and verbs) that can be captured by pictures, and it can also generate information about the retrieval speed of combinations of lexical items.
Fast retrieval of collocations is an important aspect of fluent second language production (Wray & Perkins, 2000) that should be taken into account to obtain content relevance. The finding that the PN’ scores cannot account for the variance in the WPTT’ scores suggests that the WPTT’ encompasses additional processes to those measured by the Picture Naming task. We assume that retrieving words that cannot be depicted by pictures or retrieving combinations of words, taps a wider array of lexical retrieval skills than can be measured in a Picture Naming task. In addition to tapping a broader range of the construct of lexical retrieval, the WPTT can also be applied outside a research context for classroom purposes. Recent research has indicated the importance of lexical retrieval for writing skills (Kellogg, 1999; McCutchen, 2000). Scores on the WPTT can be used to assess the fluency of this essential aspect of language production.

The main focus of this research was to determine which processes are involved in the WPTT. Our study has shown that it is crucial to obtain a measure of speed of typing. As speed of typing is not part of the theoretical construct of written lexical retrieval but an artefact due to the method used, we have to stress the importance of filtering typing speed out. Only if typing speed is taken into account can we have a good look at the construct of written lexical retrieval. A measure of speed of typing can then be used to remove the bias in lexical retrieval times due to differences in typing skills or in general reaction time. Because of the written computerized format of the WPTT it is possible to use the speed of Typing scores. Had we measured speed of writing on an electronic writing pad as Bonin (2001) did, handwriting skills would have been important instead and they should have been controlled for. Similarly, spoken Picture Naming tasks should be corrected for general answering speed, although it is difficult to imagine how this can be done in practice. As the correlations derived from the regression analysis with residual scores show, failing to take these factors into account may lead to inflated correlations due to unwanted method bias.

As pointed out by Messick (1989, p.48), ‘discounting the redundancy of a construct is only powerful, or even sensible, in relation to closely related or rival constructs’. In the case of the WPTT’ this means that the related or rival tests are speed measures as well. Naturally, this leads to high intercorrelations between measures. For this reason, even though bivariate correlations are necessary indications of the relationships between the measures, they should be supplemented by analyses that can demonstrate unique variance once shared variance is filtered out. In the case of the WPTT’, our analyses showed that there was a strong relationship with another test of written lexical retrieval, the Picture Naming task, and that the PN’ scores also contributed a substantial amount of unique variance. This suggests that the WPTT’ scores can be interpreted as an alternative measure for tapping the construct of written lexical retrieval. In contrast, even though there was a medium correlation between the Lexical Decision Task scores and the WPTT scores, the Lexical Decision Task scores did not contribute uniquely. It is reasonable to assume that the speed of finding the orthography (OE’scores) is related to the scores of the WPTT’, a task in which students not only have to retrieve the words but also have to use the orthography in writing them down. Although we expected the WPTT’ scores and the OE’ scores to be related, it turned out that the OE’scores contributed uniquely to the WPTT’ scores whereas the Lexical Decision Task scores did not. It seems that due to the written format of the WPTT’, orthographic encoding speed contributes to the WPTT’ scores. This is an important finding because it demonstrates that a certain amount of the time necessary to retrieve words in writing is spent in finding an orthographic representation. Crucially, the regression analysis results demonstrate that the WPTT’ can tap more than speed in finding the orthography and this is additional evidence that the WPTT’ is an adequate measure of the much broader construct of written lexical retrieval.
The fact that the LDT' scores have no significant unique contribution suggests that we should reconsider the validity of Lexical Decision Tasks as a tool for measuring lexical retrieval. Using Lexical Decision Tasks to tap written lexical retrieval should be critically evaluated, given the evidence about the validity of LDT's in the current study.

In another study, an experimental intervention aimed at increasing lexical retrieval was developed (Snellings et al., 2002, Chapter 2). The effect of this training was measured by the WPTT. The findings of this study were analyzed to see whether test scores were altered in a way that is consistent with the theory on lexical retrieval. In addition to the findings reported in the current study, responsiveness to experimental treatment provides support for the WPTT's validity. The results showed that the trained group improved in both level of correctness and Reaction Times (RTs), indicating that the measure is sensitive to differences in speed of lexical retrieval resulting from experimental training.

The present study is an important step in validating a new and more direct test to establish gains in written lexical retrieval speed, investigating empirically whether the new test of lexical retrieval is capable of measuring what it claims to do. Although we have demonstrated in previous research that our approach is effective in enhancing speed of lexical retrieval, further research has to make clear which factors in a training are crucial, and what is the best way to achieve an increase in retrieval speed. It should be noted that a sole focus on speed may have negative social consequences as well. We would like to stress that a focus on speed should be complementary to an approach focusing on receptive and productive unspeeded knowledge of words. At the moment however, a focus on speed of lexical retrieval is lacking in educational practice whereas previous research (McCutchen, 2000; McCutchen et al., 1994) suggests that speed of lexical retrieval is essential for fluent writing. Still, it is important to realize that the present study is carried out in a research context and as pointed out by Shepard (1993), social consequences are not the main focus at this stage. In addition, whatever may be the best way of increasing retrieval speed, a good measure of the intended construct is still essential. The Written Productive Translation Task is a measure of lexical retrieval that is not restricted to single words, that can incorporate different parts of speech, requires a written response, and that seems a better measure of the intended construct than existing measures.