http://dx.doi.org/10.18778/7969-032-9.16

Jakub Bielak, Mirosław Pawlak, Anna Mystkowska-Wiertelak
Adam Mickiewicz University, Kalisz, Poland

Testing the use of grammar: Beyond grammatical accuracy

1. Introduction

If one happens to be a learner of something these days in a context that is not totally informal, one is very likely to be tested, and is virtually bound to be tested if the context of learning is at least slightly institutionalized. Even if the learning is neither required by nor taking place within a formal environment of a school, training program or work arrangement, and is instead conceived, instigated and conducted in the privacy of one's home or other informal setting, self-testing is very likely to be part of the experience. The omnipresence of testing is also a conspicuous feature of language education at all levels and in most institutional and geographical settings. Language testing may tap learners' knowledge of a given language as a whole, as in language proficiency tests, the ability to learn a language, as in language learning aptitude tests, or the command of all the basic subsystems of language, such as vocabulary or pronunciation, among other things. The present paper is restricted to discussing selected issues in testing grammar.

The aim of the paper is to stress the importance of testing the meaning-and--use dimension of grammar, which is unfortunately all too often neglected by teachers, researchers and other language testers. Teaching and subsequently testing grammatical meaning should be regarded as a priority now that most, or at least a steadily growing number of language teachers and researchers view language primarily as a tool for communication, and it is highly surprising that still all too often they are not. As the meaningful use of grammar for communication

often takes the form of spontaneous, usually oral exchanges, in the present paper we also highlight the need to test grammatical knowledge of the implicit sort (Ellis 2005). In this contribution we thus set ourselves the task of illuminating selected challenges encountered in the design of grammar tests tapping not only grammatical form but also grammatical meaning in both more controlled and more spontaneous use. For this purpose, we provide an illustrative discussion of the design and scoring of the data collection tools used in a pilot study we recently conducted.

To provide some background to this discussion, we first present the most conspicuous general problems of contemporary grammar testing, which also includes the exposition of the considerable complexity of language testing in general. Subsequently, the slow shift of emphasis among language professionals and researchers from testing solely or mostly grammatical form/accuracy to measuring grammatical meaning/use as well is discussed. This is followed by a brief presentation of the importance of and techniques of testing both explicit and implicit grammatical knowledge. Next, we provide an extended illustrative discussion of the measures designed for and used in a pilot study focusing on a selected aspect of English grammar, with special emphasis on construct definition, testing both explicit and implicit knowledge and scoring. As is customary, several concluding remarks are offered towards the end of the paper.

2. General grammar testing problems

According to Purpura (2004: 4), despite the large amount of research since the mid-1980s on the teaching and learning of grammar, there is still "a surprising lack of consensus" as to what type of assessment tasks to use and how to design tasks for specific assessment purposes, which will at the same time constitute reliable and valid measures of performance. As Purpura (2004: 4) goes on to say:

> [i]n other words, there is a glaring lack of information available on how the assessment of grammatical ability might be carried out, and how the choices we make in the assessment of grammatical ability might influence the inferences we make about our students' knowledge of grammar, the decisions we make on their behalf and their ultimate development.

This fundamental problem of grammar testing stems in large measure from the existence of considerable uncertainty, or at least a lack of precise wording, as to what exactly constitutes grammatical knowledge. What testifies to this is what was found by Norris and Ortega (2000) in their meta-analysis of research on form-focused instruction, namely that there is often a mismatch between what is claimed by researchers about grammar acquisition [which should be strongly linked to implicit/automatized knowledge (Ellis 2008b)] and what has really been tapped by tests (which is often just explicit knowledge). There is therefore an

acute need to always clearly define the constructs relating to the components of grammatical knowledge one is trying to measure, a point which is also highlighted in the discussion of the stages of test design included later in the paper.

Another problem, or perhaps just difficulty, of language testing and assessment, and therefore grammar testing, is that they are extremely complex processes, especially in the case of large scale high-stakes standardized testing, but also in the case of classroom assessment and testing for research purposes. What gives one an idea of the complexity of language and grammar assessment is the following non-exhaustive summary of steps which may be part of language test design and development, compiled predominantly on the basis of Fulcher (2010) and also Bachman and Palmer (1996) and Mislevy, Almond and Lukas (2004).

At least three critical steps that need to be made right at the outset of test design may be distinguished, namely the establishment of *test purpose*, *criterion* and *construct*. First of all, at the test conception stage, the *purpose of a test* has to be stated such as motivating the learners, providing feedback on progress to the teacher or making some high-stakes decision such as admission to a program of study. Also early in the process of test design, it is customary to settle on *test criterion*, which is usually some real-life performance the mastery of which is attempted to be tapped by a test, such as ordering food in a restaurant, or, if it is not possible to match what is intended to be tested with an easily-defined real-life activity, a more instruction-based performance such as using active and passive voice to package information on the sentence and discourse level. Essential to designing a good language test and validating it is the next move, *construct definition*, which consists in spelling out the underlying abstract concept that the test measures; good examples of language constructs are such notions as grammatical ability, the mastery of some grammatical form or the ability to use a given form in a meaningful manner. Following the establishment of test purpose and test criterion, as well as defining the construct, a series of more practical activities are undertaken in the process of test design.

The first of these more pragmatically oriented steps is *task/item design*, which is normally in the form of *specification writing* for a range of test features. They include task/item specifications, evidence (test taker response) specifications (including specifications for scoring), test assembly specifications (e.g. how many items of different types are to be included), presentation specifications (e.g. margin, font type and size, use of color, amount of text on a single page/screen), and delivery specifications (e.g. the number of invigilators/proctors, timing). All of these specifications relate more or less directly to the most specific and detailed level of test architecture, that is test tasks, items and format.

The next series of procedures in the process of test design and development have to do with ensuring that the specifications, and, especially, sample tasks and items work the way they had been envisaged to work. The first one is the *evaluation* of the specifications and sample items by external experts such as

language teachers or applied linguists. The next stage is *prototyping*, which involves trying out a small number of tasks/items with small groups of learners. What follows is *piloting*, that is, trying out the samples with a larger group of learners and calculating the relevant statistics. *Field testing*, which may follow and complement piloting proper, is a name that is sometimes given to large scale piloting involving a complex test which has received its final shape in terms of the number of items in different tasks. The steps of evaluation, prototyping and piloting are normally wrapped up by *assembling* the items into the final test format.

Finally, the actual *writing* of tasks and items is carried out, and several additional decisions are usually made. Once written, the pool of items for test tasks must be thoroughly *checked* so that any of the following are eliminated: item-specification incongruences, problems with the key (the correct answers), bias against some subpopulations of test takers, and language mistakes. One kind of decisions that might be made concern the possible inferences following the administration of the test, e.g. decisions concerning the pass/fail threshold, the possibility to retake the test or lack thereof, the time lap before test retake, and the like. The complexity of the whole enterprise of test design and development which has just been sketched is exacerbated by the fact that in reality it is usually not a linear process undertaken by a single actor, but rather a constant back-and-forth affair instigated by a team of collaborators (Davidson and Lynch 2002: 57-59; Fulcher 2010: 154).

What contributes to the complexity of language testing is the fact that test *administration* is a highly complicated affair, too. It involves a lot of complex planning having to do with booking classrooms, arranging tables, preparing equipment and making sure that it works properly (e.g. beamers, heating), and so on. Test administrators also need to make efforts to minimize the impact of extraneous variables such as noise and fatigue. In addition, sometimes accommodations for the sake of disabled test takers must be made. What is more, particularly in situations in which the test is to be scored by individuals not directly involved in its design and development, rater training may be necessary. Thus, although it is not part of test design, test administration contributes to the general complexity of language and grammar testing.

A related problem afflicting language testing has to do with an unrealistic view of the activity as relatively simple, which was succinctly expressed by Fulcher (2010: 93):

> [w]hen asked to produce a test, many teachers start with writing test items. Managers frequently encourage this because they expect a teacher to produce the test in an afternoon, or over the weekend at best. It is something that is perceived to be the easiest part of the role of a teacher.

This kind of an approach to testing is likely to result in highly imperfect tests, although fortunately one may reasonably expect this attitude to be less widespread

among researchers and teacher-researchers, compared to language professionals not undertaking research.

The most significant deficiency of contemporary grammar testing noted by many researchers concerns the lack of balance in grammar tests with respect to the two major facets of language, namely linguistic form and meaning. As Purpura (2004) noted, despite the large scale embracement of communicative language teaching, grammar testing has been surprisingly resistant to change and has been in far too many cases focused on traditional testing of grammatical form in the guise of such mainstays of testing as multiple-choice and gap-filling tasks, to the neglect of testing grammatical meaning. According to Larsen-Freeman (2009), in this conservative approach grammatical knowledge is defined in terms of accurate comprehension and production and testing is done with the help of decontextualized, discrete point items. What is more, (explicit) knowledge of grammar is usually tested, but not the ability to employ grammar in real-life communication. Thus, it seems that there is a need to raise the awareness of teachers as well as researchers of the problem identified here as well as the possible ways of overcoming it and the necessity of doing so in at least some situations. If grammar testing is to reflect the emphasis of contemporary language teaching on the semantics of grammar and meaningful language use at all levels of advancement, it must adopt a more balanced approach towards the formal and semantic facets of grammar.

3. Towards testing grammatical meaning/use

In a more recent approach informed by communicative language teaching the ability to use grammar in writing and real-time communication is emphasized and tested (McNamara and Roever 2006) in a more holistic manner. Here, rating scales assessing accuracy, complexity, the range of structures used, and meaning/use are deployed (Larsen-Freeman 2009). However, there are serious disadvantages of this kind of assessment. As Larsen-Freeman (2009: 533) noted, "[t]he judgments are subjective, and because the assessment formats are more open-ended, they are subject to possible inconsistencies". Another problem associated with this kind of testing that Larsen-Freeman (2009) mentioned is that it is difficult to distinguish between grammar and other factors such as textual organization. Also, on the basis of such testing it may be difficult to diagnose grammar difficulties and provide appropriate feedback to learners (Purpura 2004). Larsen-Freeman (2009) concluded that both the integrative approach to testing grammar just described as well as the traditional form oriented approach have a contribution to make.

In the present day of heavy emphasis on language instruction with a communicative bent and increased focus on meaning, which have been around for quite a while, it may be surprising that according to Larsen-Freeman (2009: 534)

redefining the construct of grammar for assessment purposes so that it includes grammatical meaning/use is an innovation in grammar testing. Purpura's (2004: 89) view of grammatical ability as involving "the capacity to realize grammatical knowledge *accurately* and *meaningfully* in test-taking or other language-use contexts [emphasis ours, JB, MP, AMW]" is a notable incarnation of this (relatively) novel trend. Examples of what is covered by Purpura's grammar constructs for grammatical assessment, that is grammatical form and grammatical meaning, as well as what he views as essentially the extra-grammatical area of pragmatic meaning, are given in Figure 1. Other researchers (e.g. Larsen-Freeman 2002, 2003) view grammatical ability as involving, in addition to form and meaning, also *use*, which seems to partially overlap Purpura's grammatical meaning and pragmatic meaning. This use dimension is in fact closely related to meaningful and (pragmatically) appropriate choices language users constantly make in their linguistic performance from among different grammatical structures. It should be signaled at this point that we attempted to focus on the meaningful use of grammar defined along these lines in the tests of our design which are discussed in Section 5. In addition to giving the meaning/use dimension of grammatical competence its due, of late grammar testing experts and language acquisition researchers have also called for testing yet another hitherto neglected facet of this core linguistic ability.

Grammatical form (accuracy)	**Grammatical meaning**	**Pragmatic meaning (appropriateness/ /conventionality/ /naturalness/acceptability)**
Prosodic forms	Interrogation	Social identity
Inflections	Time	Power
Voice	Passivization (focal meaning, foregrounding)	Sarcasm
Cohesive forms (demonstratives)	Informational links to avoid redundancy	

Dependence on context

Figure 1. Examples of Purpura's (2004: 91) grammar constructs for assessment purposes

4. The importance and means of testing both implicit and explicit grammatical knowledge/ability

Because of the previous neglect or even exclusion of spontaneous and automatic employment of grammar, that is, its implicit knowledge, from language tests, a relatively vocal support of its inclusion has been recently heard in applied linguistics and second language acquisition. What transpires from Norris and Ortega's (2000) meta-analysis and is at the same time lamented or at least admitted by many (e.g. Doughty 2003; Erlam 2006; Ellis 2008a; Reindeers and Ellis 2009) is that often a particular instructional option is claimed to favorably affect acquisition, but the testing instrument used taps mostly explicit knowledge. However, in his monograph on grammar testing, Purpura (2004: 45) emphasized that "comprehensive assessments of grammatical ability should attempt to test students on both their explicit and their implicit knowledge of grammar". What may be seen as legitimizing this view is that according to Ellis (2008a: 5) second language acquisition theories and „current approaches to language testing" view implicit knowledge as primary because it enables fluent communication. In a similar vein, Doughty (2003) said that a true test of an instructional method is whether it results in implicit knowledge. Not only have authors called for the testing of implicit knowledge in addition to explicit knowledge, but they have also suggested practical means of doing so.

Generally, two well known ways of measuring implicit grammatical ability may be distinguished. There seems to be agreement that probably the best way to tap the command of implicit grammatical knowledge is to deploy tests involving fluent and spontaneous language use tasks, which have been variously referred to as, for instance, extended-production tasks (Purpura 2004) or free constructed response tasks (Norris and Ortega 2000). The second general option in measuring highly automated command of grammar are a few other options endorsed by Ellis (2005) and collaborators (Ellis et al. 2009). They include such "surgical" tests as the oral elicited imitation test or the timed grammaticality judgment test, which are characterized by relatively crisp and uncomplicated form/design, elicit short responses targeting the relevant forms in a precise manner, and are seemingly easy to implement. Both spontaneous production tests and tests eliciting only short responses involve a number of disadvantages.

Reflecting their contrasting natures, the problems associated with the two broad types of measuring implicit grammatical knowledge are also quite distinct. The most important problem with the extended production tests is that their design, administration and scoring are highly time-consuming if task essentialness (Loschky and Bley-Vroman 1993) of particular grammatical features, that is, their actual elicitation from test takers, is to be achieved. The major inconvenience with the short response tests, on the other hand, is that they are considerably divorced

from spontaneous performance, which may compromise their ecological validity. Also, serious problems concerning their implementation and not insignificant questions concerning their reliability and usefulness have been reported by the present authors (Bielak 2011; Bielak and Pawlak 2013; Mystkowska and Pawlak 2012). Because the use of either of the general types of tests tapping learners' implicit grammatical knowledge involves significant problems, if one intends to gauge this kind of grammatical ability, one either has to use one of them and admit the disadvantages which are involved, or one may attempt to circumnavigate the difficulties by designing a test which tries to combine the advantages of both of these options and simultaneously tries to steer clear of the problems. The former latter was done in the case of the tests whose illustrative discussion is provided in Section 5.

5. Illustrative discussion of the measures designed for and used in a pilot study

Before the details of the tests which are used here to exemplify the incorporation of the semantic dimension into grammar testing are offered, the pilot study of which they were a part should be briefly introduced. The study involved a quasi-experiment with a pretest-posttest design exploring the effects of teaching the meanings/use and additionally the form of the English passive and active voice based on their description offered by the framework of Cognitive Grammar (Langacker 1987, 1991). The participants were 27 year two university students of elementary education with English at a Polish university. There were two experimental mixed-level groups; while one of them, the cognitive group (COG; n = 14), received treatment based on Cognitive Grammar, the treatment offered to the other one, the traditional group (TRAD; n = 13), was based on traditional ELT materials. This pilot study did not involve a control group. As already stated, the target forms were the English passive and active voice. Importantly, however, not only the form, but also the meaning/use of these structures were the focus of instruction. In particular, the treatment in the two groups focused on what might be described as the *meaningful use* of passive and active voice in (extended) discourse depending on thematic structure (topic, focus) and foregrounding/backgrounding of information. In other words, the treatment taught the participants to make meaningful choices from among the options of active and passive voice in their language performance. Additionally, in a somewhat incidental manner, the *form* of active and passive voice in the simple present, present progressive, simple past and present perfect tenses was also taught. Importantly, the tests used in the study were designed with an express aim of measuring the above mentioned constructs, namely the *meaningful use* of passive and active voice in discourse as well as the *form* of these structures, which will be explicitly demonstrated in Section 5.2. The

treatment used in the study was feature-focused, took approximately 100 minutes, and involved explicit instruction of both inductive and deductive sorts followed by input- and output-oriented practice and feedback.

5.1. Data elicitation

The fist data collection tool used in the study was a written limited production response test, which was at the same time a written narrative test. It included a series of prompts, several of which are given here for illustrative purposes:

1. *The house I want to describe is very interesting.*
2. *locate / somebody / the house / in the suburbs / near a beautiful lake and park,*
3. *protect / a high fence / the house / and / surround / an oriental garden / the house*
4. *design / Japanese gardeners / this garden,*
5. *later / learn / these gardeners / a new job,*
6. *attend / these gardeners / many golf courses taught by the best golfers.*

The participants were required to write a narrative including sentences based on the prompts. The sentences were expected to be either passive or active depending on the surrounding discourse determining topic/focus relations. The underlined sentences such as the first one on the list above were to be reproduced in an unchanged form and were included in the test because they were necessary to make the narrative realistic and cohesive. The test created 16 obligatory contexts for passive voice and the same number for active voice. This measure was intended as an explicit knowledge test as it was in the written mode, which allows ample reflection and multiple reediting of responses, especially if it is not strictly timed. This was not the case with the test under discussion, with most participants completing the test within approximately 30 minutes, but some taking as long as 45 minutes.

The second test used in the study was an oral limited production response test, which was, just as the written test, an oral narrative test. However, it was intended as a measure of mostly implicit knowledge, or, given the controversial nature of the concept of implicit knowledge (cf. Bielak and Pawlak 2013: 166; DeKeyser 2003; Purpura 2004: 117), highly automatized explicit knowledge. Its form was very similar to the form of the written test (that is why no sample items/prompts are provided), but it was expected to tap the participants implicit rather than explicit knowledge because it was in the oral mode, meaning that there was no time for pondering or reediting the responses, which were recorded. The relatively strong focus on implicit knowledge was achieved not only by the employment of the oral mode, but also by the fact that the time limits of 90 seconds for getting acquainted with the vocabulary in the prompts and 6 minutes for completing the narrative,

despite not being as strict as to necessitate highly rushed responses and induce undue levels of stress in the participants, were rigorously respected.

At this point, the first major dilemma or difficulty we encountered in the design of our testing instruments needs to be mentioned. While there was no doubt that the written test tapping mostly explicit knowledge should not be strictly timed, the question whether the oral test should be speeded or not was a serious consideration and in fact something of a query. This was because one finds conflicting views concerning this issue in the literature. On the one hand Ellis (2001, 2005) and the like-minded researchers who worked on the Marsden Project (Ellis et al. 2009) were of the opinion that time pressure is necessary to ensure that implicit knowledge does get tested. On the other hand, other authors have voiced important caveats and doubts concerning this feature of measuring implicit linguistic knowledge, which should not be ignored. Purpura (2004: 45) for instance said that caution should be exercised with respect to speededness in tests

> since it is often difficult to determine the impact of speed on the test taker. In effect, speed may simply produce a heightened sense of test anxiety, thereby introducing irrelevant variability in the test scores. If this were the case, speed would not necessarily provide an effective means of eliciting automatic grammatical ability.

In a similar vein, Bachman (1990) claimed that speeded tasks gauge not only grammatical knowledge but also the ability to respond quickly. Purpura (2004) also reported that it is extremely difficult to judge whether and for whom a test is speeded. The perception of speededness is individualistic and even if testees have enough time to complete tasks, there may still be a subjective feeling of speededness, which might result in anxiety and poor performance, which, in turn, confounds scores. These considerations, as well as the disadvantages of the two major types of measures of implicit knowledge discussed in Section 4 had to be taken into account when we were finding a way out of the speeded-unspeeded dilemma and settling for a tool to be used in our study.

The rationale behind the decision to use a timed oral limited production response test, rather than any other test, was as follows. Because of the practical requirement we encountered in the setting where the pilot study was conducted that the implicit knowledge test not be exceedingly time-consuming in its administration and scoring, it was decided not to use an extended constructed response real-time communication test. This decision was also prompted by the fact that such tests do not easily succeed in eliciting target structures. Instead, the decision was made to use a moderately speeded oral limited production response test, which combined the advantages of a short response speeded test and a constructed response spontaneous communication test. As already mentioned, the test was reasonably timed, which put a moderate amount of pressure on the participants, which may be believed to have resulted in their reliance on the

implicit system to a large extent and not have made them overly stressed and anxious. Simultaneously, although it elicited a series of short constrained responses rather than free responses, being a narrative-formation test it resembled a real-life communication free constructed response test to a certain degree. In sum, the oral test intended as a measure of mostly implicit knowledge was speeded to a certain, arguably reasonable extent, and it also elicited responses which bore some features of spontaneous communication.

5.2. Scoring: tapping the form and meaning of grammar

Lado (1961) viewed scorability, defined as the ease with which test tasks and items may be scored, as a desirable feature of language tests. However, scorability is not easy to achieve, and, as Fulcher (2010: 197) notes,

> [p]roblems with the measurement component are much more common than we might think, but they are not usually documented and published because they are part of the test development process. This is to be regretted, as research on scoring particular item types is part of the nitty gritty of practical test development.

To fill the gap in the research and reporting of test making, as well as to shed further light on the intricate business of testing the meaning and use of grammar, an extended discussion of the scoring decisions we made with respect to our tests seems to be in order. Importantly, in this account, the constructs we defined for our tests receive a detailed exposition.

First of all, test makers face the decision whether to use right/wrong scoring or partial credit scoring, which involves granting credit on a scale from no credit to full credit, with one or more intermediate levels. For two major reasons, partial credit scoring is recommended by most researchers (e.g. Bachman and Palmer 1996; Purpura 2004) even with respect to selected response tasks, and definitely in the case of constrained and free production tasks. First of all, partial credit scoring enables testers to make inferences concerning interlanguage development. Also, it results in a more fine-grained picture of testees' abilities. Unfortunately, in addition to these important advantages, partial credit scoring has the disadvantage of being less simple than right/wrong scoring, which compromises test scorability.

As these test were essentially the same (except for the mode and timing), for both the written and the oral tests we used partial credit scoring of the same kind. The design of the tests allowed us to use them to tap at least three combinations of two basic constructs, which were the meaning/use of the English passive and active voice and the form of the same structures. For the measurement of each of the three combinations of the two basic constructs, which may themselves be regarded as higher level constructs and might be termed passive/active voice form

and meaning/use, passive/active voice meaning/use and form, and passive/active voice thematic structure, a different scoring scheme was devised. The adoption of the three scoring schemes effectively resulted in the creation in both the written and oral mode of three distinct tests measuring both the form and meaning/use of the target structures with varying relative importance of these two essential constructs in different tests. Thus, what may be called a passive/active voice form and meaning/use test gave much more weight to the form dimension than to the meaning/use dimension of the English passive and active voice so that it might be considered as a measure focusing predominantly on form; the passive/active voice meaning/use and form test increased the importance of the meaning/use construct so that it was on a more or less equal footing with the form dimension, while the passive/active voice thematic structure test placed much more emphasis on the meaning/use dimension than on the form dimension, so it should be viewed as a test tapping predominantly grammatical meaning/use. The way in which these varying weights attached to the two basic constructs were achieved, as well as the general nature of the constructs, will become clearer once we have discussed the details of the scoring schemes we adopted for each of the three tests in either mode, which will now be done.

The form and meaning/use test, tapping mostly grammatical form but also meaning/use to a certain, and a much smaller degree, was scored as follows:

- 3 pts if obligatory occasion was created[1] and passive form (the correct form in the correct tense) and meaning/use (the correct voice) were supplied;
- 2 pts if obligatory occasion was created and a single inaccuracy with respect to form, tense, or meaning/use was present, e.g. *The garden was designed* Ø *Japanese gardeners* (form), *The house was rebuilded* … (form), *The house was built by builders in 1989* (meaning/use);[2]
- 1 pt if obligatory occasion was created and two inaccuracies of the above sort were present, e.g. *The house was rebuilded by builders in 1989* (form and meaning/use);
- 0 pt if no obligatory occasion was created or if more than two inaccuracies of the above sort were present.

The meaning/use and form test, which attempted to tap grammatical form and meaning/use in approximately equal measure, was scored as follows, with the assumption that, in contrast to the previous scoring scheme, neither the use of a wrong tense (e.g. the simple past instead of the present perfect) nor any other problems with form, as long as it was clear which voice was used, was penalized:

- 2 pts if obligatory occasion was created and meaning/use was supplied, e.g. *The house was already rebuilded twice*;

[1] The term *obligatory occasion* is not intended to mean that obligatory occasion analysis of the sort used and described by Brown (1973) was employed.

[2] Here and in the remaining examples of erroneous responses the errors are underlined.

- 1 pt if obligatory occasion was created and a single inaccuracy with respect to meaning/use was present, e.g. *The garden was designed Ø*, *Was built the house 25 years ago*;
- 0 pt if no obligatory occasion was created, or, if obligatory occasion was created but two or more inaccuracies with respect to meaning/use were present, e.g. *The house was building 25 years ago*, *Was built the house 25 years ago by builders*.

The thematic structure test, which put grammatical meaning and use at a premium but also measured form to a much smaller degree, was scored as follows, again with the assumption that tense was not assessed, and also that form was to a large extent not assessed either, except for the presence or otherwise and the order of noun phrases and the verb:

- 2 pts if obligatory occasion was created, which was tantamount to the presence of a verb, and if the order of noun phrases and the verb was correct, e.g. *A few builders killed during construction*;
- 1 pt if obligatory occasion was created and a single inaccuracy with respect to thematic structure was present, or if a lexical verb was absent*, e.g. During construction a few builders killed something* (thematic structure), *Saw the house 20 people* (thematic structure), *The house was in 1989* (lexical verb absent);
- 0 pt if no obligatory occasion was created, or if obligatory occasion was created but two or more inaccuracies with respect to thematic structure were present, e.g. *Two months ago a young couple from Boston bought the house.*

5.3. Correction for guessing

Another issue that testers may sometimes have to consider with respect to scoring is whether to correct for guessing or not. It turned out to be quite an important consideration in the case of the example tests discussed here and for this reason it will now be covered in considerable detail, although the mainstream approach in language testing literature is that in most cases correction for guessing is superfluous.

Correction for guessing is generally discouraged for a number of theoretical and, especially, practical reasons. In a general statement, Fulcher (2010: 218) said that "there is no theoretical or empirical basis in test-taker behavior" for the application of the existing methods of calculating guessing and corrections for guessing. Fulcher (2010: 219) went on to say that "[i]n reality, guessing [in the case of closed response items] only occurs if the test is so speeded that test takers do not have time to complete the test within the time set". Also with reference to selected response measurement instruments, Bachman and Palmer (1996: 205) said that correcting for guessing "is virtually never useful in language tests". Some of their arguments were that we cannot normally be sure if guessing took place or not

and that "there is a difference between random guessing and informed guessing, in which test takers are able to narrow down the number of possible correct responses on the basis of partial knowledge" (Bachman and Palmer 1996: 205), which should be rewarded. Given the lack of evidence for the need for correcting, the lack of certainty that guessing really occurs, and the possible usefulness of guessing, and also assuming that testees are normally allocated a reasonable amount of time, it might seem that correcting for guessing is unnecessary.

What is more, Bachman and Palmer (1996) recommend steps aimed at eliminating or reducing guessing in selected response tests, which are not difficult to make. First, as already hinted at, ample time should be allowed, so that the majority of test takers are able to provide responses without resorting to guessing. Second, the difficulty of the test should be matched with test taker ability levels. Finally, testees should be encouraged to make informed guesses on the basis of partial knowledge. Although the three strategies related to the adoption of the appropriate tame frame, difficulty level and informed guessing are thought to apply to selected response tests such as multiple choice, it will soon become apparent why they might also possibly pertain to the kind of tests we are discussing here for illustrative purposes.

It has to be admitted that what would seem to reinforce the view that in the case of our tests correction for guessing should be forgone are Bachman and Palmer's (1996: 208) words concerning limited production tests such as ours that "the probability of guessing the correct answer is essentially nil".

However, our tests were also grammatical/pragmatic choice tests in the sense that the participants of the pilot study taking them were required to use either passive or active voice in a manner appropriate to a given context/situation (cf. Bielak 2012; Larsen-Freeman 2003). It is useful to consider in this connection Purpura's (2004: 59) remarks concerning the idea of pragmatic choice in grammar testing:

> (…) from an assessment perspective, the notion of pragmatic choice presents an interesting challenge. When a student produces a correct sentence on a test, we might assume that she is choosing from several possible alternatives that she knows and has chosen the one that she feels is accurate, meaningful and appropriate for the context. Unfortunately, we often have no data to examine the alternatives that she has not chosen to produce. In fact, it may be that the student knows only one way of expressing the message.

In our case, it seems that we can assume that test takers were aware of two alternatives because test instructions asked them to use either active or passive voice. However, on the posttests (the study involved two posttests) quite a few of the participants, probably under the influence of the treatment, which focused more heavily on passive voice rather than on active voice because the latter may be considered as a default option, attempted to use in their responses only or mostly the passive. This guessing-like strategy hypothesized to have been induced by the

treatment serving a kind of a priming function resulted in inflated scores for these participants on the posttests. Their scores were unduly inflated especially in the case of the meaning/use and form test and scoring scheme and the thematic structure ones, where just the plain decision to use one voice rather than another, without necessarily demonstrating the mastery of grammatical form, was sufficient to earn one a lot of points for a given test item, and consequently for the whole test if these decisions were repeated. This was obviously undesirable, as it might not have reflected the state of these participants' knowledge following the treatment.

It should be noted that all testing instruments which involve a binary choice of this sort run the risk of test takers employing this kind of guessing strategy. Obviously, even if a given participant does not attempt to employ exactly this strategy of using the passive (or any other of two possible options) all or most of the time, more randomized guessing is still possible, that is, which voice (or some other option) should be used might be guessed rather than worked out with resort to the participant's knowledge with respect to every single item. Similarly to the guessing strategy we noticed in the pilot study, which, if followed throughout the test, may result in 50% of correct choices, this one also gives one a chance of getting 50% of one's choices right. However, while this kind of guessing is probably impossible to be detected, given the arguments against correcting for guessing presented earlier, its occurrence is much less likely. In contrast, the all-out priming-induced guessing strategy discussed here is easy to discern, and probably quite likely to appear given that quite a few of the participants indulged in it.

Therefore, the question whether to correct for this (guessing) strategy or not turned out to be a valid one for our tests, even though we have also showed that usually correction for guessing will be a non-issue. As we have also showed, the case of our tests is different from the majority of tests in that the tests we have designed were simultaneously constructed response tests and selected response tests. One possible way of correcting for guessing in this particular case would be using a rather special scoring system. In this scheme, the number of points awarded for the items whose keys are, say, the passive, must not exceed the number of points scored for the items whose keys are the active. Obviously, we could have worded it differently by placing *the active* in the subject noun phrase and *the passive* in the object noun phrase of the previous sentence; the alternative sentence would indicate basically the same scoring system. In still other words, under this scheme the number of points scored for the items whose keys include the passive will always be the same as the number of points scored for the items whose keys include the active. A possible disadvantage of this scoring method aimed at correcting for the above mentioned guessing strategy is that it compromises the ecological validity of the test to a certain extent, as some responses which are essentially correct—when considered in isolation—will not be awarded any points. For his reason we are still uncertain whether this scoring scheme should be employed in the research in the pilot of which the tests under discussion were tried.

Alternatively, rather than correct for the kind of guessing we are discussing here, it is possible to introduce some changes to the design of the tests, each of which, unfortunately, also brings with itself certain not insignificant disadvantages. One possibility is not mentioning the two options (passive/active voice) in the test instruction. The downside here is that some test takers might include in their responses forms other than active and passive voice such as, for instance, noun phrases, as in *The construction of the house happened in 1989* used instead of the expected *The house was constructed in 1989*. Another strategy that might be used, possibly along with the one just discussed, is the inclusion in the test of a large number of diverse distarctors, namely items apparently testing and eliciting other grammatical structures, so that testers' attention is drawn away from the two target structures (especially the passive). An obvious disadvantage of this is an inflated test length, as well as the possible underuse of the target forms already mentioned with reference to the first strategy. Because neither the change of test instructions nor the inclusion of diverse distractors is not without serious disadvantages and does not guarantee the disappearance of guessing, some other steps aimed at reducing the likelihood of guessing which are not directly related to the tests themselves might be made.

It has to be admitted that in addition to decisions having to do strictly with the tests discussed here, certain changes to the design of the whole study and the way it was conducted might also be considered in order to prevent guessing. First of all, the introduction of more balance into the study's treatment with respect to the relative attention paid to passive and active voice might significantly reduce the priming effect mentioned earlier and therefore prevent the participants from indiscriminate employment of the passive in all or most of the responses. Secondly, a simple act of excluding these participants who did this from the sample might be employed. This would obviously negatively affect the generalizability of the study, especially if the sample size is not impressive in the first place, but may be necessary anyway since it is not certain whether the strategies aimed at preventing guessing considered above bring a bout the desired effect or not. This may also be a solution of choice for those who do not want to run the risk of compromising the ecological validity of the test, which may accompany the scoring scheme aimed at correcting for guessing described above. It appears then that in addition to attending to the design and scoring of the test, modifying the design of the entire study and the way it was carried out may also go some way toward reducing or eliminating guessing.

6. Conclusions

Although grammar has always been tested as part of language teaching and research, what has gradually changed over the years are the specific aspects of grammatical ability which have been subject to measuring. In this paper we have

tried to stress the importance of testing the semantic-pragmatic dimension of grammar, or, in other words, grammatical meaning and use, which does not in reality complement gauging the formal aspects of grammar as often as it should despite the slowly growing awareness among testers of the need to do so. We have also emphasized the importance of clearly defining the constructs subject to testing, which normally relate to grammatical accuracy and form, as well as grammatical meaning and use. What has also been stressed is the necessity to gauge the explicit and implicit dimensions of grammatical knowledge. In addition, we have also tried to present selected problems often encountered when engaging in the highly complex process of designing and developing language tests, with special emphasis on those intended to tap grammatical meaning and use.

Following Fulcher and Davidson (2007: xix), it is interesting to note that the field of language testing is different from other areas of applied linguistics in that it involves a highly practical activity of creating something very palpable and serviceable, that is, language tests. In this paper we have discussed the practicalities of the design of two major tests we have used in a recent pilot study concerning grammar teaching and learning. This discussion served as an opportunity to focus in more detail on selected problems of designing grammar tests which tap not only grammatical form but also the meaning/use dimension and to demonstrate how they might be solved. In particular, it was showed how the explicit and implicit dimensions of both formal and semantic grammatical knowledge may be tested by two similar tests. Their close similarity might in fact be considered an advantage, as it may facilitate making comparisons between learners' explicit and implicit knowledge of grammatical features. The tests we created also demonstrated how relatively good scorability might be achieved in a grammar test measuring both form and meaning even if one uses partial credit scoring, which is generally desirable but which nonetheless tends to compromise scorability. In addition, some general arguments against correcting for guessing were presented, as well some tentative ways of doing so if a need arises.

At the very end, it should be restated that the frequently encountered absence from grammar testing of a focus on meaning and use is particularly surprising given the contemporary omnipresence of language tests and in most cases should not be condoned. We hope to have demonstrated here that tapping grammatical meaning/use does not constitute an impossible challenge but is instead doable, and that it may be done without compromising other important requirements such as setting oneself a clear goal for testing, creating a precise scoring scheme or unambiguously defining test constructs. Importantly, although we have designed our test for research purposes, there is no reason why most if not all of our considerations, examples, tips, etc. should not apply to grammar testing in the classroom or testing performed for some other purposes.

References

Bachman, L.F., Palmer, A.S. 1996. *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Bachman, L.F. 1990. *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bielak, J. 2011. Oral elicited imitation test—problems and challenges. Presentation delivered at the Topics in Applied Linguistics conference, Opole, Poland, November 2011.

Bielak, J. 2012. Cognitive grammar in the service of teaching linguistic flexibility and creativity. In H. Lankiewicz and E. Wąsikiewicz-Firlej (Eds.), *Informed teaching: Premises of modern foreign language pedagogy*, pp. 155-173. Piła: Państwowa Wyższa Szkoła Zawodowa im. Stanisława Staszica.

Bielak, J., Pawlak, M. 2013. *Applying cognitive grammar in the foreign language classroom: Teaching English tense and aspect*. Heidelberg: Springer.

Brown, R. 1973. *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Davidson, F., Lynch, B.K. 2002. *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.

DeKeyser, R.M. 2003. Implicit and explicit learning. In C.J. Doughty, M.H. Long (Eds.), *The handbook of second language acquisition*, pp. 310-348. Malden, MA: Blackwell.

Doughty, C. 2003. Effects of instruction on learning a second language: A critique of instructed SLA research. In B. VanPatten, J. Williams and S. Rott (Eds.), *Form-meaning connections in second language acquisition*, pp. 181-202. Mahwah, NJ: Lawrence Erlbaum.

Ellis, R. 2001. Some thoughts on testing grammar: An SLA perspective. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara and K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies*, pp. 251-263. Cambridge: Cambridge University Press.

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

Ellis, R. 2008a. Investigating grammatical difficulty in second language learning: Implications for second language acquisition research and language testing. *International Journal of Applied Linguistics* 18: 4-22.

Ellis, R. 2008b. *The study of second language acquisition*. (2nd edition.) Oxford: Oxford University Press.

Ellis, R., Loewen, S., Elder, C., Erlam, R., Philp, J. and Reinders, H. (Eds.) 2009. *Implicit and explicit knowledge in second language learning, testing and teaching*. Bristol: Multilingual Matters.

Erlam, R. 2006. Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics* 27: 464-491.

Fulcher, G. 2010. *Practical language testing.* London: Hodder Education.

Fulcher, G., Davidson, F. 2007. *Language testing and assessment: An advanced resource book*. London: Routledge.

Lado, R. 1961. *Language testing*. London: Longman.

Langacker, R.W. 1987. *Foundations of cognitive grammar*. Vol. 1: *Theoretical prerequisites*. Stanford: Stanford University Press.

Langacker, R.W. 1991. *Foundations of cognitive grammar*. Vol. 2: *Descriptive applications.* Stanford: Stanford University Press.

Larsen-Freeman, D. 2002. The grammar of choice. In E. Hinkel, S. Fotos (Eds.), *New perspectives on grammar teaching in second language classrooms*, pp. 103-118. Mahwah, NJ: Lawrence Erlbaum.

Larsen-Freeman, D. 2003. *Teaching language: From grammar to grammaring*. Boston: Thomson and Heinle.

Larsen-Freeman, D. 2009. Teaching and testing grammar. In M.H. Long, C.J. Doughty (Eds.), *The handbook of language teaching*, pp. 518-542. Malden, MA: Wiley-Blackwell.

Loschky, L., Bley-Vroman, R. 1993. Grammar and task-based methodology. In G. Crookes, S.M. Gass (Eds.), *Tasks and language learning*, pp. 123-167. Clevedon: Multilingual Matters.

McNamara, T., Roever, C. 2006. Language testing: The social dimension. *Language Learning* 56, Supplement 2.

Mislevy, R.J., Almond, R.G. and Lukas, J.F. 2003. *A brief introduction to evidence-centred design*. Research report RR-03-16. Princeton, NJ: Educational Testing Service.

Mystkowska-Wiertelak, A., Pawlak, M. 2012. *Production-oriented and comprehension-based grammar teaching in the foreign language classroom*. Heidelberg: Springer.

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

Purpura, J. 2004. *Assessing grammar*. Cambridge: Cambridge University Press.

Reinders, H., Ellis, R. 2009. The effects of two types of positive enhanced input on intake and L2 acquisition. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp and H. Reinders (Eds.), *Implicit and explicit knowledge in a second language*, pp. 281-302. Clevedon: Multilingual Matters.