In order to measure the effects of grammar instruction, it is common to elicit performance on particular target forms before and after instruction, to see what changes have occurred, and to draw inferences from these findings. However, in much instructed SLA research it seems that certain test types are preferred and that only a limited range of tests is administered. In this paper I will consider what it is that common types of test measure and discuss whether the results of different types of test can be combined to give a more rounded picture of the state of the learners’ target form knowledge. Armed with this more detailed picture, it is argued that one would be in a better position to say how instruction had affected the knowledge which underlies a wider range of performance, thus making the findings more robust and potentially more directly useful to language teachers.

In this paper, firstly, I will highlight the need for test rationales in second language grammar instruction research. Tests are a crucial part of measuring the effects of treatments in a research study, yet although the issue of test validity, and particularly construct validity, is often discussed in language testing circles and in guides to research design, I am going to argue that in this particular field of instructed SLA research, namely research into second language grammar instruction, it is comparatively rare to find a report in which the researchers explicitly state the reasons for using the tests they do, and that this is a major omission. Secondly, I am going to argue for the importance of ensuring as close a match as possible between test type and pedagogic aims. I will show, in fact, that pedagogic aims are often not mentioned at all in
Measuring the effects of explicit and non-explicit grammar instruction: casting the net wide.

research reports, or are extremely general, or are only implied and this means that the tests used to measure the effects of instruction seldom have any direct relationship to the goals of instruction. And finally, I am going to suggest that in measuring the effects of grammar instruction one needs to use a battery of tests in a principled fashion so that learners can display what they know about target features and what they can do with what they know across a range of performance contexts.

Initially, I will look at elicitation procedures that have been used in research into second language grammar instruction. Next, I will discuss some of the principles on which I have based my choice of tests in a research project I am currently designing to measure the effects of four types of grammar instruction. Finally, I will make some tentative suggestions about how elicitation procedures in second language grammar instruction research could be better integrated into research designs.

The term ‘test’ is used throughout to refer to all kinds of elicitation procedure in research studies. Sometimes tests are called ‘instrumentation’, or ‘outcome measures’, and sometimes this term covers how the learner performance is scored. Here the term ‘test’ refers to the ways in which researchers elicit performance so that they can then measure the dependent variable by scoring this performance according to specific criteria.

What tests have second language grammar instruction researchers used?

This brief survey of the tests used by researchers who have looked at second language grammar instruction is guided by Norris and Ortega’s (2000) meta-analysis of experimental and quasi-experimental investigations into the effectiveness of L2 instruction published between 1980 and 1998. In so doing, it is acknowledged that there may be a danger of omitting many influential studies published before or after these dates, of including studies which some may argue are not clearly dealing with gramm-
mar instruction, and of being drawn into Norris and Ortega’s search for studies which reported their findings with sufficient statistical detail to be included in their meta-analysis. However, here I am interested in reviewing empirical studies and the tests that have been used, so it seems appropriate to use Norris and Ortega as a reference list. Norris and Ortega’s paper is an impressive attempt to synthesize research findings from around forty-five different investigations into the effectiveness of formal instruction, with grammar the overwhelming target of instruction.

Before looking at the different types of test used in the studies included on Norris and Ortega’s list, it is interesting to note that Norris and Ortega judge that all the studies they looked at used tests (what they call ‘outcome measures’) which were designed to answer one of only three basic questions:

(a) Did participants acquire the ability to recognize the target form?
(b) Did participants acquire the ability to produce the target form?
(c) Did participants acquire the ability to explain the rule-governed nature of the target form? (N&O 2000:469)

There are perhaps some question marks over whether ‘recognition’ automatically includes ‘comprehension’, but it is useful to keep in mind that there may be only three or four basic types of evidence that one can, in principle, look for when measuring grammar learning.

In Appendix A of their report, Norris and Ortega list the forty-eight unique sample studies they used in their meta-analysis. In this Appendix, amongst other things, they describe the nature of the tests for each study according to the kind of response they required. Thus, tests required either:

1. a metalinguistic judgement response (MJR); or
2. a selected response (SR); or
3. a constrained constructed response (CCR); or
4. a free constructed response (FR).
Measuring the effects of explicit and non-explicit grammar instruction: casting the net wide.

The tests were defined by Norris and Ortega as metalinguistic judgement type tests if the research participant was required to evaluate the appropriacy or grammaticality of L2 target structures as used in item prompts. An example is given from de Graaf (1997) in which participants are asked to judge whether certain morphological and syntactic forms in an artificial language are correct or not.

What Norris and Ortega call selected response tests, their second category of test, required participants to choose the correct response from a range of alternatives, typically either in answer to comprehension questions based on the use of the target L2 forms (e.g. multiple choice tests) or in order to complete a sample segment of the target language with the appropriate target form. An example of a selected response test is given from Leow (1997).

Tests which were coded as constrained constructed response required the participant to produce the target form under highly regulated circumstances, where the use of the appropriate form was essential for grammatical accuracy to occur. This kind of test required learners to produce L2 segments ranging in length from a single word up to a full sentence, but all such tests were designed with the intent to test L2 ability to use the particular form within a highly controlled linguistic context (e.g. sentence completion or sentence combining tests). An example is given from VanPatten and Cadierno (1993).

And finally, free constructed response tests were ones that required participants to produce language with relatively few constraints and with meaningful communication as the goal for L2 production (e.g. oral interviews or written compositions). An example is provided from Harley (1989).

Norris and Ortega thus very helpfully categorize the tests used in the studies they review according to the conditions they impose on learner performance. For the pur-
poses of this paper, I have chosen to call the ‘MJR’ and ‘SR’ tests ‘Type A’ tests and the
‘CCR’ and ‘FR’ tests ‘Type B’ tests (see Figure 1 on the next page). The former seem
to be “what you know” tests. They are recognition or non-production tests in that
learners respond non-verbally to the written stimuli which have been provided. The
responses in these tests indicate the learners’ beliefs about the L2 grammar. Whether
or not the responses also display a degree of comprehension is a moot point. The Type
B tests are, by contrast, “what you can do with what you know” tests. These are pro-
duction-based tests which require learners to write or speak under a wide range of
conditions from the highly constrained to the comparatively free.

It would seem to me to be a reasonable principle to try to include at least one Type
A and one Type B test in a study in order to be able to assess both the learners’ knowl-
edge and/or comprehension of the target form as well as their ability to produce the
form. One would then be in a position to be able to say what the learners know about
the form and what they can do with what they know. In the event of a choice, one
might propose that researchers err on the side of using production tests rather than
recognition tests as these may be more revealing of whether a form has been acquired
or is on the way to being acquired. There are, of course, counter-arguments. The stud-
ies on Norris and Ortega’s list do, in fact, tend to use production tests slightly more
than recognition tests, seventy-seven percent of the studies using at least one produc-
tion test and sixty-five percent using at least one recognition test. However, it is strik-
ing that only just over half of the studies (56%) use both Type A and Type B tests.
Measuring the effects of explicit and non-explicit grammar instruction: casting the net wide.

Notes:
* “not available for independent analysis owing to insufficient reporting by primary researchers.” (Norris and Ortega 2000: 527)

a. MJR was used to assess the learners’ general grammatical knowledge prior to the study proper. I think that the test actually used was much more a ‘constrained constructed response’ test (CCR) in N&O’s terms rather than an MJR test.
It may seem strange to judge studies on this basis when they all have quite different research objectives in mind and when a recognition test used on its own may seem an appropriate and adequate way of eliciting learner performance. My contention is, however, that when measuring the effects of instruction, eliciting a range of performance will enable the researcher to say more about what learners know and what the learners can do with what they know and will enable the researcher to convince a wider audience that something akin to real learning has occurred.

Looking for test rationales.

Having looked at the tests used in these studies according to Norris and Ortega’s categories and having looked in a little more detail at the actual tests themselves, one is, I think, naturally inclined to ask why researchers chose the tests they did. Tests normally impose considerable constraints on learner performance and consequently also impose considerable constraints on what a researcher can say about the learner’s knowledge of the target forms as this is necessarily based on the test performance. One would expect, therefore, to find that researchers explain their choice of tests with some care. Unfortunately, this does not appear to be the case.

Of the twenty-seven studies on Norris and Ortega’s list which were thoroughly reviewed, sixteen (59%) do not explain the rationale for the choice of tests at all. This is a somewhat conservative estimate because those which were credited with providing a rationale often only did so minimally and somewhat vaguely. For example, study number nineteen on the Norris and Ortega list used a written picture narration task as its elicitation procedure. Several pages after the description of the ‘Procedure’ the researchers state that:

“The second research question of this study sought to investigate the immediate effect of textual input enhancement on the learner’s subsequent production.” (p203) (My emphasis)
Measuring the effects of explicit and non-explicit grammar instruction: casting the net wide.

This was taken as a rationale for using the written picture narration task although it is neither expansive nor specific. Other studies explained the rationale for one of the tests used, but not for others. Study number eight, for example, used a grammaticality judgement test, a picture description test requiring single sentence statements from the learner, and a metalinguistic rule completion test. A rationale is given for the first test:

“to have a record of developing knowledge under speeded conditions, that is, where monitoring (in the sense of drawing on explicit knowledge) is difficult.” (p389).

But no rationales are given for choosing the other tests. Nevertheless, this study was credited in this survey with providing at least some explanation for the choice of tests.

It might be objected that it does not matter that test rationales are difficult to find. The rationale is often implicit in the choice of test itself. I think it is important, however, that researchers explain why they chose the tests they did. The performance elicited by one type of test is going to be different from that elicited by another type of test. Not stating the performance conditions defined by the tests on which one is going to base one’s evaluations of the learner’s developing knowledge of the target form seems to me to be a major oversight.

There are reasons which might be suggested perhaps for why so few studies explicitly state the rationale for the choice of tests. I am going to propose that it may be most closely related to having only a limited conception of the instructional goals. Although these studies nominally deal with the effects of instruction, my contention is that only a minority fully conceptualise instruction as a real-world construct and that this has implications for the tests which are used.
Looking for pedagogic goals.

In fact, identifying pedagogic goals in studies which are concerned with the effects of instruction is almost as difficult as identifying test rationales. Only three of the twenty-seven studies surveyed (11%) provided reasonably clear statements about instructional goals. These were all Canadian studies looking at the effects of functional-analytic teaching in French immersion contexts. It has to be said, however, that even these studies were not totally explicit about the effect the teaching was supposed to have on the learners’ knowledge, use or understanding of the target structures and only one specified which skill areas were to be worked on. Study number sixteen on Norris and Ortega’s list states that the aim of the instruction is that:

“the grammatical competence of immersion students, with respect to use of the imparfait and the passe compose, can be enhanced” (p335)

This is clear in so far as it goes, but it would be more helpful if we were told under what general conditions the learners would be expected to use, recognize or judge these structures.

Sometimes hints are given about pedagogical aims when researchers discuss their choice of target structure. For example, the authors of study number five on Norris and Ortega’s list say:

“We wanted our subjects to learn the particular constraints on the English dative alternation.” (p363)

But again, we are not told what conditions might pertain when the learners demonstrate what they have learned about English dative alternation.

Of course, it might appear to be almost irrelevant, perhaps a little unfair, to look for instructional goals particularly in the case of the more laboratory-based, strictly ex-
Measuring the effects of explicit and non-explicit grammar instruction: casting the net wide.

Experimental studies, and especially when volunteers are used. Twelve of the twenty-seven studies reviewed fell into this category. How could there be an authentic teaching goal, it might be objected, for learners studying an artificial or semi-artificial language outside a curriculum framework? Equally, surely the pedagogic goal is implicit in most of these studies? It is obviously to help participants learn the target structures. But, even in the case of laboratory-based studies, I would argue that it is important that researchers state clearly what they expect the instructional treatment will achieve. Are they aiming at establishing some rather static, declarative knowledge in the learner, or are they expecting that the learner will be able to deploy this newly gained information under specified performance conditions? Real-world pedagogic goals normally specify some target language use context defined by various conditions. It would be useful in research reports if performance conditions were at least discussed, however abstract these conditions might be, or alternatively that statements were made justifying why performance conditions do not need to be specified.

Difficulty in identifying the instructional goals is also exacerbated as a result of a lack of clarity about the relationship between the instruction and the tests used to measure the effects of the instruction. If the conditions imposed by the tests are not explained, it is difficult to understand what kind of performance the instructional treatment is seeking to affect. Thus there are good research design principles to be invoked here too: If one is clear about what kind of performance one is trying to affect via instruction, then one can incorporate tests into the research which explicitly provide conditions which elicit that performance.

To summarize the argument so far: It would be helpful if

- explicit test rationales were given in research reports; and
- pedagogic goals were clearly stated in specific terms.
Also, it may be constructive for researchers to think of pedagogic goals ahead of tests and to

- make pedagogic goals part of their test rationale.

As a bare minimum, it would be very useful if they would
- make it absolutely clear what they think the tests measure.

Some pedagogical goals, a test rationale and a proposed battery of tests.

It is easy to criticize research in this area for the lack of test rationales and clearly stated pedagogical goals. Therefore, it is only fair that I lay my own choice of tests open to scrutiny and make explicit the principles behind this choice in a research project I am at present planning.

Very briefly, my study is aimed at comparing the effectiveness of the four instructional treatments shown in Figure 2 below. The broad aims are to see if more explicit grammar instruction techniques are more effective than less explicit techniques and to see if techniques which are integrated into communicative interaction are more effective than techniques which are separated from such interaction. However, following on from the points made above, there are several important questions which need to be answered in order to properly specify the pedagogic goals and, once these are established, it will also be necessary to define the tests that will be employed in such a way that they demonstrably measure what I am proposing to teach.
Measuring the effects of explicit and non-explicit grammar instruction: casting the net wide.

<table>
<thead>
<tr>
<th>Integratedness (+/-)</th>
<th>Explicitness (+/-)</th>
</tr>
</thead>
</table>
| Whether or not the intervention is integrated into a communicative task; | Whether or not the intervention is an explicit attempt to draw learners’ attention to form.

Notes:
First, I need to make explicit my pedagogic goals. I need to be clear about what aspects of performance I want to enhance: Which skills will I be focussing on? Do I want to enhance the ability to produce the target forms, or to comprehend the target forms, or will I be satisfied if the learners show that they know the rules they have been taught by stating those rules in some way? And if I expect the learners to do something with their newly-gained knowledge, what conditions would I expect to pertain to that performance? That is, what performance conditions should the tests impose?

I need tests of explicit knowledge and of implicit knowledge because I want to gauge the effectiveness of explicit instructional treatments and of non-explicit treatments. Also, because data will be collected from learners engaged in an actual language course and because this course emphasizes the development of speaking skills, it is appropriate that the grammar which is taught is tested mainly in terms of how it contributes to spoken production. However, given that the learners will most probably be at an elementary to pre-intermediate level of overall proficiency and on entry to the course will probably not be comfortable communicating orally with a native-speaker examiner/teacher, it seems prudent to provide somewhat less demanding types of test than spoken production tasks. These might enable learners to demonstrate what they know at the outset in the pre-tests without having to communicate with an examiner-teacher and may allow learners to show what they have learned in the post-tests even though they may not be able to produce the target structures yet.

I therefore want tests of what the learners know and of what they can do with what they know. I want tests of what the learners can produce and of what they can demonstrate they know receptively. I want tests which make learners access explicit knowledge and tests which make them access implicit knowledge. And I want tests which attempt to obviate the effects of test method for groups receiving either integrated or separated form-focussed instruction. Thus, drawing again on the suggestions made in
Han and Ellis (1998) and Ellis (2001, 2004), I have pared my tests down to the four presented in Figure 3.

Tests 1 and 2 I have dubbed ‘Communicatively contextualized spoken production tests’ because learners will produce target forms while communicating with the teacher as they complete a task. The tasks will be similar to the ones learners receiving Treatments 3 and 4 will undertake in class while receiving recasts and corrective recasts. In Test 1 the learners will not be given time to plan their utterances either before or during the interaction. The task will be speeded, thus favouring access of implicit knowledge over explicit knowledge. In Test 2, the learners will be able to take time over what they say. This will be an unspeeded task. Some planning time will be available before the learner makes each utterance. This task will thus favour access of explicit knowledge over implicit knowledge. For both tests learners will be told to try to be as accurate as possible, but for Test 1 learners will additionally be told that the task needs to be completed as quickly and as smoothly as possible. Both Tests are in line with the course objectives. Naturally, a good performance on Test 1 would be the ideal which teachers might hope for, but demonstrating that one is able to communi-
cate under the less pressurized conditions of Test 2 might also be seen as performance which is moving towards the ideal.

Tests 3 and 4 have been called ‘Communicatively decontextualized ‘recognition’ tests’. They take language out of a communicative context and ask learners to make judgements about the grammaticality of sentences under speeded and unspeeded conditions. Test 3 should favour access of implicit knowledge and Test 4 access of explicit knowledge. These are not “comprehension” tests necessarily, although Ellis (2004) assumes that semantic processing is, in fact, the first operation involved in answering a GJT item (p. 256). They are, however, receptive tests which may be arguably less demanding than the production tests within a communicative context. They also give learners an opportunity to demonstrate a receptive understanding of the target structures, something which may develop ahead of the ability to use the target structures, possibly in time for the post-tests which immediately follow the treatment period.

The four tests are an attempt to bridge the divide between what are sometimes called competence tests and performance tests allowing learners the chance to show what they know and what they can do with what they know across a range of conditions. They are also an attempt to bridge or subvert other dichotomies represented here in Figure 4.
The reader may wonder, however, why it is necessary to attempt to bridge or subvert these dichotomies. Any demonstration that learning of some kind has taken place is usually an extremely impressive achievement, so it seems overly ambitious, perhaps, to look for a range of learning outcomes. The answer is, of course, that what one decides to test depends totally on how one defines the ability one is trying to measure and I think that what I am trying to measure is “grammatical ability” which itself straddles complex divides.

**Grammatical ability**

Bachman (1990) has composed a general interactionalist construct definition of communicative language ability, which includes “both knowledge, or competence, and the capacity for implementing, or executing that competence in language use” in context (Bachman 1990: 84). Bachman proposes that a component called “strategic competence” controls the interaction between the language knowledge and fundamental processes regarded as traits and the conditions imposed by the context.

Based on Bachman (1990) and Bachman and Palmer (1996), Purpura (2004) defines ‘Grammatical ability’ in the following way:
“Grammatical ability is [ ] the combination of grammatical knowledge and strategic competence; it is specifically defined as the capacity to realize grammatical knowledge accurately and meaningfully in testing or other language-use situations.” (p. 86)

Purpura defines grammatical performance as “the observable manifestation of grammatical ability in language use” (2004: 87) and presents the figure above (Figure 5) to show the relationships among grammatical knowledge, ability, and performance.
Measuring the effects of explicit and non-explicit grammar instruction: casting the net wide.

If I could reconfigure this slightly I would emphasize the lower-most oval because, in my view, the processing implications resulting from test method/task characteristics/performance conditions are critical to how grammatical performance is manifested and I tend to think that the factors presented as dichotomies before such as speeded/unspeeded, productive/receptive, and so on exert a stronger influence than Purpu-ra perhaps acknowledges.

Here, however, is a framework, a theoretical basis for the four tests I have proposed. I am attempting to test grammatical ability as part of communicative language ability and this ability consists of grammatical knowledge and the capacity to realize grammatical knowledge accurately and meaningfully in use. My ‘Communicatively de-contextualized ‘recognition’ tests’ are aimed at assessing grammatical knowledge and the ‘Communicatively contextualized spoken production tests’ are aimed at assessing the capacity to use this knowledge.

There is, of course, a requirement to pitch the tests at roughly the right level for the group of learners in question. Douglas has argued that rather than fight against the effects of test method we should use test method in a principled way to design tests appropriate for a particular population (Douglas 1998:152-153). With my four tests, I think that I have followed this advice. The students I will be teaching may not perform particularly well on a timed communicative oral test (Test 1). But Tests 2, 3 and 4 will, I think allow them to show what they know and what they can do. It seems perfectly legitimate in my teaching context to assess planned/monitored performance via Test 2. I do not see this monitored performance as something which should necessarily be denigrated as some kind of disabled form of speech, especially as it may characterize many classroom learners’ speech for some considerable part of their language learning experience. Monitored speech is at any rate not unusual among the students I teach. This suggestion does seem to be supported to some extent in the arguments put forward by Tarone (1998).
Of course, the principles for testing in second language grammar instruction research are well-established. The theoretical groundwork is already highly developed in works such as Bachman and Palmer (1996) and Purpura (2004) and in articles such as Han and Ellis (1997), Ellis (2001; 2004) and Hulstijn and de Graaf (1994). Also, interesting discussions such as that in Rounds and Schachter (1996) show how tests in research that is not only intended to be interesting and justifiable from linguistic, psychological, and SLA perspectives, but also pedagogically appropriate can only be arrived at after proposals have passed through a series of theoretical and practical screens (Rounds and Schachter 1997: 101).

By way of conclusion, however, let me make some proposals. Firstly, that if ‘instruction’ is part of what we are interested in, then my contention is that the pedagogic goals of the instructional treatments used in second language grammar instruction studies should be clearly stated and should play an important role in shaping the measurement instruments we use.

Secondly, providing rationales for the tests we use is crucial both in the research design process and for the interpretation of results. If the reasons for using a particular test are not explained in the research report and if it is unclear what the researchers thought that test measured, it becomes very difficult to understand what performance on the test is meant to reflect. Consumers of research reports are required to engage in heavy inferencing from the results given under these circumstances.

Thirdly, because the performance conditions defined by the tests constrain the inferences which can be drawn from test performance, providing a range of tests which impose different performance conditions on a principled basis can provide learners with better opportunities to demonstrate what they know and what they can do.

Finally, many language teachers may be more impressed by results from second
Measuring the effects of explicit and non-explicit grammar instruction: casting the net wide.

Language grammar instruction studies which are based on tests that cover a spectrum of performance conditions. Many language teachers may want learners to display an enhanced grammatical ability under the taxing conditions of spoken communicative interaction as the result of instruction, but they may also be quite forgiving about the actual standards learners achieve and the performance produced. Performance which is monitored may be a quite acceptable consequence of instruction for certain populations of learners and may be the only kind of performance which will be seen as an immediate result of instruction. Second language grammar instruction research which attempts to provide the kinds of evidence that language teachers might be impressed by or with which they might identify is an example of applied research which helps to bridge the gap between the ‘academy’ and ‘practitioners’.

************

Note

This is an adapted version of a presentation given at the British Association of Applied Linguistics/Cambridge University Press Seminar ‘Instructed Second Language Learning: State of the Art’ held at the University of York, UK, on April 11-12 2005.

************

References


Tim Ashwell


Measuring the effects of explicit and non-explicit grammar instruction: casting the net wide.