MEASURING IMPLICIT AND EXPLICIT KNOWLEDGE OF A SECOND LANGUAGE

A Psychometric Study

Rod Ellis
University of Auckland

———

A problem facing investigations of implicit and explicit learning is the lack of valid measures of second language implicit and explicit knowledge. This paper attempts to establish operational definitions of these two constructs and reports a psychometric study of a battery of tests designed to provide relatively independent measures of them. These tests were (a) an oral imitation test involving grammatical and ungrammatical sentences, (b) an oral narration test, (c) a timed grammaticality judgment test (GJT), (d) an untimed GJT with the same content, and (e) a metalinguistic knowledge test. Tests (a), (b), and (c) were designed as measures of implicit knowledge, and tests (d) and (e) were designed as measures of explicit knowledge. All of the tests examined 17 English grammatical structures. A principal component factor analysis produced two clear factors. This analysis showed that the scores from tests (a), (b), and (c) loaded on Factor 1, whereas the scores from ungrammatical sentences in test (d) and total scores from test (e) loaded on Factor 2. These two factors are interpreted as corresponding to implicit and explicit knowledge, respectively. A number of secondary analyses to support this interpretation of the construct validity of the tests are also reported.

———

Two of the major goals of SLA research are to define and describe second language (L2) linguistic knowledge and to explain how this knowledge develops over time by specifying the external and internal variables involved (R. Ellis, 1994). There is no agreement among SLA researchers regarding the theoretical model that should inform the first of these goals and, I will argue, there has been little real progress in achieving the second goal because of a general failure to address how learners' L2 knowledge can be measured. Thus, SLA as a field of inquiry has been characterized by both theoretical controversy and by a data problem concerning how to obtain reliable and valid evidence of learners' linguistic knowledge. This article is primarily concerned with the second of these problems, but, first, it is necessary to consider the theoretical question.

THE NATURE OF LINGUISTIC KNOWLEDGE

What is meant by *linguistic knowledge*? Broadly speaking, there are two competing positions. The first, drawing on the work of Chomsky (e.g., Chomsky, 1976), claims that linguistic competence consists of a biological capacity for acquiring languages, commonly referred to as Universal Grammar (UG). According to this position, linguistic knowledge consists of the knowledge of the features of a specific language that are derived from impoverished input (positive evidence) with the help of UG and learning principles, such as the subset principle (Wexler & Mancini, 1987). This combination ensures that learners do not need to rely on negative evidence to eliminate nontarget features. This view of language acquisition is largely restricted to grammar and is mentalist in orientation, emphasizing the contribution of a complex and highly specified language module in the mind of the learner.

The second position, drawing on the connectionist theories of language learning as advanced by cognitive psychologists such as Rumelhart and McClelland (1986), does not view language learning as cognitively different from other forms of learning, in that it draws on a general mental capacity for registering and storing phonological, lexical, and grammatical sequences in accordance with their distributional properties in input. Linguistic knowledge emerges gradually as learners acquire new sequences, restructure their representation of old sequences, and, over time, extract underlying patterns that resemble rules.[1] Linguistic knowledge in this sense comprises an elaborated network of nodes and internode connections of varying strengths that dictate the ease with which specific sequences or rules can be accessed. Thus, according to this view, learning is driven primarily by input, and it is necessary to posit only a relatively simple cognitive mechanism (a kind of sensitive pattern detector) that is capable of responding both to positive evidence from the input and to negative evidence available through corrective feedback.

These positions are generally presented as oppositional. Gregg (2003), for example, dismissed connectionist theories on the grounds that they do not

account for linguistic competence at all and polemically argued that no rational scientist could possibly abandon "some sort of classical property theory of linguistic competence" (p. 123). More moderately, Major (1996) and Ioup (1996), responding to N. Ellis' (1996) connectionist account of L2 learning, suggested that both empirical evidence and innate grammatical resources are involved in language learning and that the differences between their views and that of Ellis rest largely on the relative importance they attach to each.

In one important respect, however, the two positions are in agreement. Both the innatist and connectionist accounts of L2 learning acknowledge that linguistic competence comprises *implicit knowledge*. For example, in his prolegomena for a generative theory of linguistic competence, Gregg (1989) pointed to the need to distinguish *knowing that* and *knowing how*, the latter of which Chomsky (1976) called *cognizing*. Gregg argued that *knowledge that* is "basically accidental" and that the study of learners' linguistic competence must focus on that knowledge that enables them to cognize what is grammatical and ungrammatical. For Gregg, acquisition is evident in what learners know intuitively—in their implicit and not their explicit knowledge. N. Ellis (1996, this issue) has also distinguished implicit and explicit learning of an L2, and he sees the former as primary; explicit knowledge is typically "the end product of acquisition, not its cause," although he also considers a number of ways in which explicit knowledge can foster implicit knowledge, as claimed by the weak interface position to be discussed later. Thus, connectionist accounts, like generative accounts, conceive of linguistic knowledge as intuitive and tacit rather than conscious and explicit in nature. Further, both accounts discuss the distinction between implicit and explicit knowledge in very similar terms.

In this respect, innatist and connectionist modeling of language share a common base. They both make a clear distinction between implicit and explicit linguistic knowledge. From my perspective in this article, this fundamental similarity is important. It obviates the need to address the theoretical disputations that have colored SLA. It points to a common need, irrespective of one's theory of linguistic knowledge and language learning, for empirical researchers to distinguish whether what individual learners know about a language is represented implicitly or explicitly. In the next section, I will consider why this is so important.

THE RELATIONSHIP BETWEEN IMPLICIT AND EXPLICIT KNOWLEDGE

As we have seen, there is broad consensus that the acquisition of an L2 entails the development of implicit knowledge. However, there is no consensus on how this is achieved; nor is there consensus on the role played by explicit knowledge. The UG position, as articulated by Gregg (1989), is that acquisition has nothing whatsoever to do with explicit knowledge. Cognitive accounts of L2 acquisition, however, are much more mixed. Traditionally, the relation-

ship between the two types of knowledge has been discussed in terms of the interface between them, as shown in the following discussion of three distinct cognitive perspectives.

According to the *noninterface position*, implicit and explicit L2 knowledge involve different acquisitional mechanisms (Hulstijn, 2002; Krashen, 1981), are stored in different parts of the brain (Paradis, 1994), and are accessed for performance by different processes, either automatic or controlled (R. Ellis, 1993). In its pure form, this position rejects both the possibility of explicit knowledge transforming directly into implicit knowledge and the possibility of implicit knowledge becoming explicit. However, in a weaker form of the noninterface position, the possibility of implicit knowledge transforming into explicit is recognized through the process of conscious reflection on and analysis of output generated by means of implicit knowledge (Bialystok, 1994).

In contrast, *the strong interface position* claims that not only can explicit knowledge be derived from implicit knowledge but also that explicit knowledge can be converted into implicit knowledge through practice; that is, learners can first learn a rule as a declarative fact and then, by dint of practice, can convert it into an implicit representation, although this need not entail the loss of the original explicit representation. The interface position was first formally advanced by Sharwood Smith (1981) and has subsequently been promoted by DeKeyser (1998). Differences exist, however, regarding the nature of the practice that is required to effect the transformation from explicit to implicit knowledge; in particular, researchers disagree on whether this practice can be mechanical or needs to be communicative in nature.

The *weak interface position* exists in three versions, all of which acknowledge the possibility of explicit knowledge becoming implicit but posit some limitation on when or how this can take place. The first version posits that explicit knowledge can convert into implicit knowledge through practice only if the learner is developmentally ready to acquire the linguistic form (R. Ellis, 1993). This version draws on notions of learnability in accordance with attested developmental sequences in L2 acquisition (e.g., Pienemann, 1989). The second version holds that explicit knowledge contributes indirectly to the acquisition of implicit knowledge by promoting some of the processes believed to be responsible. N. Ellis (1994), for example, suggested that "declarative rules can have 'top-down' influences on perception" (p. 16), in particular by making relevant features salient and thus enabling learners to notice them and to notice the gap between the input and their existing linguistic competence. Finally, according to the third version, learners can use their explicit knowledge to produce output that then serves as auto-input to their implicit learning mechanisms (Schmidt & Frota, 1986; Sharwood Smith, 1981).

Irrespective of the role played by explicit knowledge in the acquisition of implicit knowledge, there is wide acceptance that explicit knowledge can contribute to performance. Krashen (1977) argued that explicit knowledge is available to the *monitor*, the production mechanism that enables learners to edit

their own performance by drawing on what they consciously know to be correct. Bialystok (1982) showed that different performance tasks are likely to induce L2 learners to draw differentially on their implicit and explicit knowledge. Fairly obviously, for example, she found that formal writing tasks are likely to induce learners to draw more extensively on their analyzed knowledge of an L2 than tasks calling for unplanned, oral communication.

These positions all have their adherents and have been the topic of much argument in the SLA literature. However, they have not been subjected to empirical inquiry. One reason for this is the lack of agreed instruments designed to ascertain the nature of knowledge acquired by learners: implicit, explicit, or some amalgam of the two. To illustrate this problem, I will consider a number of representative studies that have attempted to examine the relationship between learners' implicit and explicit knowledge.

The Relationship Between Implicit and Explicit Knowledge: Past Studies

A number of early studies examined the relationship between learners' implicit and explicit knowledge (e.g., Hulstijn & Hulstijn, 1984; Seliger, 1979; Sorace, 1985). In all of these studies, explicit knowledge was operationalized as the learners' explanation of specific linguistic features, whereas implicit knowledge was determined by examining the learners' use of these features in oral or written language. In this subsection, I will examine a number of later studies and then discuss the measurement problem that underlies them.

Green and Hecht (1992) presented a set of sentences containing grammatical errors to 300 German students who were learning English in a secondary school or a university setting. The learners were asked to correct each sentence and to state the rule that had been violated. They found that although the learners were able to correct 78% of the sentences, they could only state the correct rule in 46% of the cases (although the university learners in the sample were able to do so in 86% of cases). In other words, the learners' ability to correct the errors exceeded their ability to explain the rules. Green and Hecht suggested that one interpretation of these results is that these learners' explicit rules constituted only a subset of their available implicit knowledge.

Macrory and Stone (2000) investigated students from British secondary schools and examined their perceptions of what they knew about the formation of the French perfect tense (measured by means of self-report), their actual knowledge of the tense (measured by means of gap-filling exercises), and their ability to use the tense in an informal interview and in free written production. They found that the students had a fairly good explicit understanding of the perfect tense (e.g., they understood its function, knew that some verbs required different auxiliaries, were familiar with the forms required by different pronouns, and were aware of the need for a final accent on the past par-

ticiple). In general, this study found only weak relationships among students' perceptions, their performance in the gap-filling exercise, and their use of the tense in free oral and written production. For example, whereas they typically supplied an auxiliary (not always the correct one) in the gap-filling exercise, they typically omitted it in free production except in formulaic expressions involving *j'ai* "I have." Macrory and Stone concluded that what they term *language-as-knowledge* and *language-for-use* might have derived from different sources—instruction about the rule system and routines practiced in class—thus explaining the observed disparity.

Hu (2002) conducted a study of 64 Chinese learners of English. His main purpose was to investigate to what extent explicit knowledge was available for use in spontaneous writing. He asked the learners to complete two spontaneous writing tasks and then to carry out an untimed error correction task and a rule-verbalization task before again completing two similar spontaneous writing tasks and a timed error correction task. The assumption was that the untimed correction and rule-verbalization tasks would serve a consciousness-raising function, thereby making the learners aware of the structures that were the focus of the study. Hu focused on six structures, selecting a prototypical and peripheral rule for each (e.g., for articles, specific reference constituted the prototypical rule and generic reference constituted the peripheral rule). Overall, when correct metalinguistic knowledge was available, the participants were more accurate in their prototypical use of the six structures. Also, accuracy in the use of the six structures increased in the second spontaneous writing task, suggesting that when made aware of the need to attend to specific forms, the learners made fuller use of their metalinguistic knowledge. However, Hu admitted that it is not possible to claim that the participants actually used their metalinguistic knowledge in the writing tasks, although he argued that the results are compatible with such an interpretation.

All of the studies reviewed in this subsection were correlational in design; that is, they either sought to establish whether there was any relationship between learners' explicit and implicit knowledge (Green & Hecht, 1992; Macrory & Stone, 2000) or whether explicit knowledge was available for use in tasks that were hypothesized to require implicit knowledge (Hu, 2002). Such studies do not constitute tests of the interface position (nor were they intended to do so), as demonstrating a relationship does not show that explicit knowledge subsequently transformed into implicit knowledge. Such a demonstration would necessitate an experimental study in which learners were first taught a specific rule explicitly, subsequently developed explicit knowledge of this rule, and, ultimately, developed implicit knowledge of it as a result of opportunities to practice. Again, such a study is only possible if valid and reliable means of measuring explicit and implicit knowledge are available.

One study that has directly tested the interface position is by DeKeyser (1995). DeKeyser examined the effects of two kinds of form-focused instruction (explicit-deductive and implicit-inductive) on two kinds of rules in an artificial grammar (*simple categorical rules* and *fuzzy prototypical rules*). Learning

outcomes were measured by means of a computerized judgment test, which required the learners to indicate whether a sentence matched a picture, and a computerized production test, which required them to type in a sentence to describe a picture. DeKeyser suggested that the production test was to some extent *speeded*; in other words, the learners had 30 seconds to respond. The learners were also asked to complete fill-in-the-blank tests to demonstrate their understanding of the grammatical rules. The learners in the explicit-deductive condition provided clear evidence of their ability to produce the simple categorical rules in new contexts and did better than the learners in the implicit-inductive condition. On the face of it, this study suggested that—at least in the case of simple grammatical forms—production is facilitated when learners are taught explicit knowledge about the forms and then practice them. However, as DeKeyser admitted, it was not clear to what extent the production task allowed for monitoring by explicit knowledge.

How did the four studies discussed operationalize the two types of knowledge? As in the earlier studies, explicit knowledge was typically elicited by asking learners to verbalize specific grammatical rules. Additionally, Macrory and Stone (2000) used a fill-in-the-blank exercise to tap into explicit knowledge. The studies showed more variation in their means of determining implicit knowledge. Two of the studies used spontaneous production tasks—oral and written in the case of Macrory and Stone and a fast-writing task in the case of Hu (2002). In the remaining two studies, implicit knowledge was tapped with either an untimed error correction task (Green & Hecht, 1992) or a cued sentence-based written production task (DeKeyser, 1995). There are some obvious problems with all these methods. To verbalize rules, learners must have at least some productive metalanguage and the ability to provide clear explanations of abstract phenomena. Importantly, learners' explicit knowledge exists independently of both the metalanguage they know and their ability to explain rules (R. Ellis, 2004).[2] Thus, as Bialystok (1979) pointed out many years ago, having learners verbalize rules provides a rather conservative picture of what they know explicitly. Likewise, a fill-in-the-blank exercise might invite the use of explicit knowledge, but it does not guarantee it, as learners are obviously able to complete the exercise by drawing on their implicit knowledge. With regard to implicit knowledge, spontaneous production tasks are probably the best means of elicitation (R. Ellis, 2002), but, again, we cannot be sure that learners do not access at least some explicit knowledge, especially when the task involves writing. Hu, in fact, claimed that within certain constraints, metalinguistic knowledge is available for use in spontaneous production. An error correction task, especially the kind of untimed task used by Green and Hecht, seems unlikely to produce a good measure of implicit knowledge, as the very nature of the task invites learners to access their explicit knowledge.

To date, there has been no empirical test of the interface positions for the simple reason that researchers have failed to give due consideration to implicit and explicit knowledge as constructs. Only DeKeyser (1995) discussed the validity of his chosen instrument for measuring learning outcomes in terms of the

type of knowledge it taps, and only then as a possible limitation of his study (i.e., there is no discussion of the construct validity of the instrument in the Methods section of his article). As Douglas (2001) noted, this failure to consider construct validity of testing instruments is widespread in SLA. In lamenting this, Douglas pointed to what is needed:

> . . . construct validity may be demonstrated by the construction of theoretical arguments linking hypothesized aspects of language ability to features of the test tasks, demonstrating the appropriacy of the tasks for making interpretations regarding the construct, and then providing empirical evidence that the links are in fact present. (p. 447)

It is with a view to meeting Douglas' requirement that the next subsection attempts to examine the constructs of implicit and explicit knowledge, as a preliminary step toward the development of instruments designed to provide separate measures of them.

Distinguishing Implicit and Explicit Knowledge

I will briefly consider seven ways in which implicit and explicit knowledge of language can be distinguished (R. Ellis, 2004) as a way of arriving at a conceptual account of the two constructs.

Awareness. Karmiloff-Smith (1979) distinguished two kinds of data for the study of child language development: epilinguistic data and metalinguistic data. Both involve awareness but of different kinds. Epilinguistic behavior arises when a child can demonstrate intuitive awareness of implicit grammatical rules (e.g., gender concord or the use of one article in preference to another). Karmiloff-Smith suggested that this type of behavior is evident when the child can recognize instantly that a sentence is ungrammatical. On the other hand, metalinguistic behavior is evident when the child has conscious awareness of why a sentence is ungrammatical and can demonstrate this understanding with an explanation for the ungrammaticality. Developmental psycholinguists, such as Karmiloff-Smith, suggest that children first display epilinguistic behavior and only later (5 years old or later) manifest metalinguistic behavior. Thus, as children develop, their implicit knowledge becomes increasingly analyzed, which allows for its explicit representation. Bialystok (1991) suggested that L2 acquisition is a similar process and that teaching learners explicit rules would only prove effective if the learners are ready to incorporate them into their "emerging representational structure" (p. 71).

Type of Knowledge. Anderson (1983) distinguished between declarative and procedural knowledge, suggesting that knowledge is gradually restructured from one form of representation to another. Declarative knowledge is explicit and encyclopedic in nature. It is factual in the same sense as knowledge of when the Normans invaded England or the number of degrees in the

angles of a triangle. Declarative knowledge of language involves both knowledge of abstract rules (e.g., relating to the use of articles) and knowledge of fragments and exemplars (Eichenbaum, 1997). Procedural knowledge is highly automated. This type of knowledge results when the learner gains greater control over the fragments and exemplars and also restructures declarative knowledge of rules into if-then productions of increasing delicacy. This dimension of the implicit versus explicit distinction corresponds to what Bialystok (1991) called control and constitutes the skill component of language. It involves three functions: selective attention, integration, and the ability to handle the language within real-time constraints (p. 72).

Systematicity and Certainty of L2 Knowledge. Reber, Walkenfeld, and Hernstadt (1991) convincingly argued that implicit knowledge displays lower variability than explicit knowledge. Similarly, SLA researchers have claimed that learners' interlanguages (i.e., their implicit knowledge) are highly systematic (Tarone, 1988). Although there is some disagreement as to whether interlanguage grammars contain some linguistic forms that are in free variation (R. Ellis, 1985; Tarone 1988), there is general agreement that these grammars are largely systematic; they contain categorical rules or variable rules of a probabilistic nature, or both, although not necessarily those found in the target variety. Explicit knowledge, in contrast, is often imprecise, inaccurate, and inconsistent (Sorace, 1985). Learners frequently have hunches, rather than a clear understanding, about how specific rules work. Thus, even though both types of knowledge involve some degree of nonsystematicity and uncertainty, implicit knowledge is considered to be more structured than explicit knowledge and, thus, is employed with greater certainty as to its correctness. Zobl (1995) suggested that this difference would be apparent in the standard deviations of test scores used to measure L2 learners' implicit and explicit learning, with greater variation evident in the latter.

Accessibility of Knowledge. Implicit knowledge involves automatic processing; explicit knowledge entails controlled processing. This difference is an epiphenomenon of the distinction between declarative and procedural knowledge already discussed. Krashen (1981) suggested that when communicating, learners formulate messages using their implicit knowledge; if they are focused on form, have the required explicit knowledge, and have time to access them, they can then monitor them for accuracy. SLA researchers differ, however, in whether they see the two knowledge types as clearly distinguished by accessibility. Clearly, it is possible for explicit knowledge to be accessed more or less quickly. DeKeyser (2003) suggested that explicit knowledge can be fully automatized and thereby become functionally equivalent to implicit knowledge. In contrast, Hulstijn (2002) suggested that practice will only "speed up the execution of algorithmic rules to some extent" (p. 211) and that there remains a fundamental difference between automated explicit knowledge and implicit knowledge in terms of their accessibility.

Use of L2 Knowledge. Bialystok (1982) provided evidence that the use of the two types of knowledge varies according to the specific tasks learners are asked to perform. Bialystok distinguished task demands in terms of *analysis* and *control* (i.e., tasks can require knowledge that is +analyzed/+automatic, +analyzed/−automatic, −analyzed/−automatic, −analyzed/+automatic). For example, she provided evidence to show that a written task that requires learners to detect and then correct errors will tap into +analyzed/−automatic knowledge, whereas an aural task requiring the same functions elicits +analyzed/+automatic knowledge. From a different perspective in accordance with sociocultural theory (Lantolf, 2000), explicit knowledge can be viewed as a tool that learners use to achieve control in demanding situations. Explicit knowledge manifests itself in the private speech that learners use to grapple with a communicative or linguistic problem. When asked to perform a think-aloud task (e.g., while completing a grammaticality judgment test), learners typically access declarative information to assist them (R. Ellis, 1991).

Self-report. Explicit knowledge is potentially verbalizable, although it exists in the minds of the learners independently of whether they can verbalize it. Butler (2002) found that her learners (adult Japanese learners of English) were generally able to provide some kind of explanation for the choice of articles in a cloze task. Learners' skill in verbalization might, in part, depend on their knowledge of metalanguage, although as James and Garrett (1992) pointed out, learners can verbalize a linguistic rule using nontechnical language. On the other hand, implicit knowledge is not verbalizable. Indeed, any attempt to verbalize it will entail forming an explicit representation first. As Dienes and Perner (1999) showed, it might be possible to establish degrees of explicitness or implicitness depending on how a proposition about a linguistic feature is encoded. For example, a statement such as "I know *that* is a relative pronoun that can refer to both animate and inanimate nouns" is more explicit than saying "*that* is a relative pronoun," which in turn is more explicit than "I used the word *that*." All of these statements, however, are to some degree explicit.

Learnability. More contentiously, it can be claimed that explicit knowledge is learnable at any age, whereas implicit knowledge is not. For example, Bialystok (1994) claimed that "explicit knowledge can be learned at any age" (p. 566) but that there are age-related limitations on L2 learners' ability to learn implicit knowledge. For example, learners whose first languages lack morphological markers of key grammatical functions (such as articles) will find these difficult to acquire as implicit knowledge past a certain sensitive age although they may well be able to develop explicit knowledge of them. The extent to which explicit knowledge is learnable is also controversial: On the one hand, Krashen (1982) argued that most learners are capable of learning only rules that are formally and functionally simple, but, on the other hand, Green and Hecht (1992) demonstrated that university-level German learners of English are capable of developing highly sophisticated explicit knowledge.

Table 1. Key characteristics of implicit and explicit knowledge

Characteristics	Implicit knowledge	Explicit knowledge
Awareness	Intuitive awareness of linguistic norms	Conscious awareness of linguistic norms
Type of knowledge	Procedural knowledge of rules and fragments	Declarative knowledge of grammatical rules and fragments
Systematicity	Variable but systematic knowledge	Anomalous and inconsistent knowledge
Accessibility	Access to knowledge by means of automatic processing	Access to knowledge by means of controlled processing
Use of L2 knowledge	Access to knowledge during fluent performance	Access to knowledge during planning difficulty
Self-report	Nonverbalizable	Verbalizable
Learnability	Potentially only within critical period	Any age

These seven ways of distinguishing implicit and explicit knowledge are summarized in Table 1.

Operationalizing Implicit and Explicit Knowledge

How, then, can we operationalize these constructs in order to design tests to measure them? I suggest that operationalization be based on seven criteria (based on but not identical to the seven characteristics already discussed). Following R. Ellis (2004), explicit knowledge is conceptualized as primarily involving both analyzed knowledge (i.e., structured knowledge of which learners are consciously aware) and secondarily metalanguage (i.e., knowledge of technical terms such as *verb complement* and semitechnical linguistic terms such as *sentence* and *clause*). The following seven criteria are framed to account for explicit knowledge as analyzed knowledge.

> *Degree of awareness.* This criterion refers to the extent to which learners are aware of their own linguistic knowledge. This clearly represents a continuum but can be measured by asking learners to report retrospectively about whether they made use of *feel* or *rule* in responding to a task.

> *Time available.* This criterion is concerned with whether learners are pressured to perform a task online or whether they have an opportunity to plan their response carefully. Operationally, this involves distinguishing tasks that make significant demands on learners' short-term memories and those that lie comfortably within their L2 processing capacity.

Focus of attention. Does the task prioritize fluency or accuracy? Fluency entails a primary focus on message creation in order to convey information or attitudes, as in an information or opinion-gap task. Accuracy entails a primary focus on form, as in a traditional grammar exercise.

Systematicity. This criterion requires examination of whether learners are consistent or variable in their response to a task. Learners should be more consistent in a task that taps their implicit knowledge than in a task that elicits explicit knowledge.

Certainty. How certain are learners that the linguistic forms they have produced conform to target language norms? Given that learners' explicit knowledge has been shown to be often anomalous, some learners are likely to express more confidence in their responses to a task if they have drawn on their implicit knowledge. However, other learners might place considerable confidence in their explicit rules. Thus, this criterion of explicit knowledge needs to be treated with circumspection.

Metalanguage. This criterion focuses on the relationship between metalanguage and explicit knowledge. Learners' knowledge of metalingual terms will be related to their explicit (analyzed) knowledge but not to their implicit knowledge.

Learnability.[3] This final criterion relates to the learnability of implicit and explicit knowledge. Learners who began learning the L2 as a child are more likely to display high levels of implicit knowledge, whereas those who began as adolescents or adults—especially if they were reliant on instruction—are more likely to display high levels of explicit knowledge.

It should be noted that these criteria refer both to the degree of awareness involved and conditions of use. This reflects the fact that the constitutive features of the two types of knowledge incorporate their manner of use. The criteria and their operationalizations in terms of implicit and explicit (analyzed) linguistic knowledge are summarized in Table 2.

Table 2. Operationalizing the constructs of L2 implicit and explicit knowledge

Criterion	Implicit knowledge	Explicit knowledge
Degree of awareness	Response according to feel	Response using rules
Time available	Time pressure	No time pressure
Focus of attention	Primary focus on meaning	Primary focus on form
Systematicity	Consistent responses	Variable responses
Certainty	High degree of certainty in responses	Low degree of certainty in responses
Metalinguistic knowledge	Metalinguistic knowledge not required	Metalinguistic knowledge encouraged
Learnability[a]	Early learning favored	Late, form-focused instruction favored

[a]The extent to which knowledge can be internalized by learners at whatever stage of development they have reached.

A PSYCHOMETRIC STUDY

Background

The study reported in this section originated in an earlier study (Han & Ellis, 1998). This earlier study analyzed scores derived from a battery of tests (an oral production test, a timed grammaticality judgment test [GJT],[4] and an untimed GJT) and a measure of metalinguistic ability based on learners' verbalizations of a grammatical rule. In a principal component factor analysis, scores from the oral production test and the timed GJT loaded on one factor, whereas the untimed GJT and the metalinguistic comments score loaded on a second factor. Han and Ellis labeled these two factors *implicit* and *explicit L2 knowledge*, respectively. This study was limited, however, in that it focused on a single grammatical structure (verb complementation); nevertheless, despite its narrow scope, statistically significant correlations between the measures of implicit and explicit knowledge and measures of two widely used English language tests (i.e., the test of English as a foreign language [TOEFL] and the SPEAK test) were obtained. The present study builds on the work of Han and Ellis in two ways—it investigates a larger range of grammatical structures and it explores different ways of measuring implicit and explicit knowledge.

Purpose

The purpose of the study was to develop a battery of tests that would provide relatively separate measures of implicit and explicit knowledge. It was acknowledged from the start, however, that even if task conditions that inclined learners to use one type of knowledge in preference to the other could be identified, it would be impossible to construct tasks that would provide pure measures of the two types of knowledge. As a number of researchers have noted, there can be no guarantee that the task-as-workplan will correspond to the task-as-process (e.g., Breen, 1989; Coughlan & Duff, 1994). Further, learners are likely to draw on whatever resources they have at their disposal irrespective of which resources are the ones suited to the task at hand. At best, then, the tests designed were expected to predispose learners to access one or the other type of knowledge only probabilistically.

Research Question and Hypotheses

In accordance with the general purpose of the study, the following research question was formulated: To what extent is it possible to develop tests that provide separate measures of implicit and explicit L2 knowledge? A number

of more specific hypotheses were also formulated as a way of examining the construct validity of the tests as tests of implicit and explicit knowledge. These hypotheses, which were based on the characteristics of the two types of knowledge specified in Table 1, are presented in the Results section.

Participants

A total of 111 participants completed the battery of tests to be described in this section. The participants group was made up of 20 native speakers (NSs) of English and 91 learners of L2 English.[5] The NSs were undergraduates enrolled in arts or engineering courses, graduate students, or former students of a university in New Zealand. Thirteen were male and seven female. Fifteen had studied a foreign language, including 11 of them for a period of 2 years or longer. Ten of the NSs had studied two or more foreign languages. The L2 learner group showed mixed language proficiency. Some were enrolled in low-level courses in the university's English Language Academy ($n = 21$), some were taking more advanced courses in English as a second or other language as part of an undergraduate degree program ($n = 30$), whereas others had been tested through the International English Language Testing System (IELTS), with an overall mean of 6.24 out of a possible 9.0 ($n = 44$). Of the L2 learners, 36 were male and 58 female (1 participant failed to indicate gender). They had been learning English for 10 years on average, mostly in a foreign language context—the learners had spent an average of only 1.9 years living in an English-speaking country. Most of the L2 learners were NSs of Chinese (70.5%).

Test Content

The tests were designed to provide measures of learners' knowledge of 17 English grammatical structures. The choice of the grammatical content was motivated by a number of considerations. First and foremost, an attempt was made to select target language structures that were known to be universally problematic to learners (i.e., to result in errors). To this end, the SLA literature was consulted (e.g., Burt & Kiparsky, 1972). Second, the structures were selected to represent both early and late acquired grammatical features according to what is known about the developmental properties of L2 acquisition (e.g., Pienemann, 1989). Third, the structures were selected to represent a broad range of proficiency levels according to when they were introduced in ESL courses covering beginner, lower intermediate, upper intermediate, and advanced levels. Fourth, the structures were chosen to include both morphological and syntactic features. Table 3 lists the structures and summarizes their properties in terms of the various selection criteria.

Table 3. Experimental grammatical structures

Structure	Example of learner error	Acquisition	Pedagogic introduction	Type
Verb complements	Liao says he wants *buying* a new car.	Early	Lower intermediate	S
Regular past tense	Martin *complete* his assignment yesterday.	Intermediate	Elementary/lower intermediate	M
Question tags	We will leave tomorrow, *isn't it*?	Late	No clear focus at any level	S
Yes/no questions	Did Keiko *completed* her homework?	Intermediate	Elementary/lower intermediate	M
Modal verbs	I must *to brush* my teeth now.	Early	Various levels	M
Unreal conditionals	If he had been richer, she *will* marry him.	Late	Lower intermediate/intermediate	S
Since and *for*	He *has been living* in New Zealand *since* three years.	Intermediate	Lower intermediate	S
Indefinite article	They had *the* very good time at the party.	Late	Elementary	M
Ergative verbs	Between 1990 and 2000 the population of New Zealand *was increased*.	Late	Various levels	S
Possessive -s	Liao is still living in his rich *uncle* house.	Late	Elementary	M
Plural -s	Martin sold a few old *coin* to a shop.	Early	No clear focus at any level	M
Third person -s	Hiroshi *live* with his friend Koji.	Late	Elementary/lower intermediate	M
Relative clauses	The boat that my father bought *it* has sunk.	Late	Intermediate/advanced	S
Embedded questions	Tom wanted to know what *had I done*	Late	Intermediate	S
Dative alternation	The teacher explained *John the answer*.	Late	No clear focus at any level	S
Comparatives	The building is *more bigger* than your house.	Late	Elementary/intermediate	S
Adverb placement	She writes *very well* English.	Late	Elementary/lower intermediate	S

Note. S = syntactic; M = morphological.

Test Battery

Following the criteria established previously, a total of five tests were developed.

Imitation test. This test consisted of a set of belief statements involving both grammatical and ungrammatical sentences containing the target structures. In the original version of this test, there were 68 statements. However, to shorten the time it took to administer this test, the number was subsequently reduced to 34 statements (one grammatical and one ungrammatical sentence per structure). The sentences retained were those that correlated most strongly with total test scores in an initial sample of 50 L2 learners and 10 NSs and were therefore considered the best measures of the underlying construct. The sentences were presented orally to the participants, who were then required to say first whether they agreed with, disagreed with, or were not sure about the content of each statement. This was intended to focus their attention on meaning. Next, the participants were asked to repeat the sentences orally in correct English and their responses were audio recorded. The responses were then analyzed by identifying obligatory occasions for the use of the target structures. Test takers' failure to imitate a sentence at all or to reproduce the sentence in such a form that they did not create an obligatory context for the target structure of a sentence was coded as avoidance. Each imitated sentence was allocated a score of 1 (the target structure was correctly supplied) or 0 (the target structure was either avoided or attempted but incorrectly supplied). Scores were expressed as percentage correct.

Oral narrative test. The story used in this test was designed to elicit the use of a number of the target structures (i.e., regular past tense, modal verbs, third person -s, plural -s, indefinite article, and possessive -s). The participants read a story twice and were then asked to retell the story orally within 3 minutes. Their narratives were audio recorded and subsequently transcribed. An obligatory occasion analysis was carried out to establish the percentage of correct suppliance of each target structure. Total scores were calculated by averaging the percentage scores for each structure.

Timed GJT. This was a computer-delivered test consisting of 68 sentences, evenly divided between grammatical and ungrammatical. The sentences, which were different from those in the imitation test, were presented in written form on a computer screen. Thus, there were 4 sentences to be judged for each of the 17 grammatical structures. Participants were required to indicate whether each sentence was grammatical or ungrammatical by pressing response buttons within a fixed time limit.[6] The time limit for each sentence was established on the basis of NSs' average response time for each sentence in a pilot study, to which was added an additional 20% of the time taken for each sentence to allow for the slower processing speed of L2 learners. The time allowed for judging the individual sentences ranged from 1.8 to 6.24 seconds. Each item was scored dichotomously as correct/incorrect, with items left unanswered scored as incorrect. A percentage accuracy score was calculated.

Untimed GJT. This was a computer-delivered test with the same content as the timed GJT. Again, the sentences were presented in written form. Participants were required to (a) indicate whether each sentence was grammatical or ungrammatical, (b) indi-

cate the degree of certainty of their judgment (as proposed by Sorace, 1996) on a scale marked from 0% to 100%, and (c) to self-report whether they used rule or feel for each sentence. This test provided three separate measures: a percentage judgment accuracy score based on the participants' dichotomous responses, a percentage certainty score, and a percentage score based on the participants' reported use of rule in judging each item.

Metalinguistic knowledge test. This was an adaptation of an earlier test of metalanguage devised by Alderson, Clapham, and Steel (1997). It consisted of an untimed computerized multiple-choice test in two parts. The first part presented participants with 17 ungrammatical sentences (one sentence per target structure) and required them to select the rule that best explained each error out of 4 choices provided. The second part consisted of two sections. In section 1, the participants were asked to read a short text and then to find examples of 21 specific grammatical features from the text (e.g., *preposition* and *finite verb*). In section 2, they were asked to identify the named grammatical parts in a set of sentences. A total percentage accuracy score was calculated.

These tests were designed in accordance with four of the criteria for distinguishing implicit and explicit knowledge discussed previously[7]; that is, it was predicted that each test would provide a relatively separate measure of either implicit or explicit knowledge according to how it mapped out on these criteria. Table 4 sets out these predictions. The imitation test and the oral narrative test were predicted to measure implicit knowledge because the participants would rely predominantly on feel, they would be under pressure to perform in real time, they would be focused primarily on meaning, and they would have no reason to access their metalanguage. In contrast, the metalinguistic knowledge test was predicted to measure explicit knowledge because it involved a high degree of awareness, was unpressured, focused attention on form, and, obviously, required the use of metalinguistic knowledge. Both of the GJTs required participants to focus attention primarily on form (as judging the correctness of sentences necessarily entails this). However, the two GJTs differed insofar as the timed task was predicted to measure primarily implicit knowledge, whereas the untimed GJT was predicted to measure primarily explicit knowledge. The timed task encouraged the use of feel, it was time-pressured, and there was little need or opportunity to access metalin-

Table 4. Design features of the tests

Criterion	Imitation	Oral narrative	Timed GJT	Untimed GJT	Metalanguage
Degree of awareness	Feel	Feel	Feel	Rule	Rule
Time available	Pressured	Pressured	Pressured	Unpressured	Unpressured
Focus of attention	Meaning	Meaning	Form	Form	Form
Metalinguistic knowledge	No	No	No	Yes	Yes

guistic knowledge; the untimed task encouraged a high degree of awareness and was unpressured, both of which predicted that responses would likely involve metalinguistic knowledge.

Procedure

The tests were completed in the following order: imitation test, oral narrative test, timed GJT, untimed GJT, and metalinguistic knowledge test. All tests included a number of training examples.

The imitation test was completed in one-on-one meetings between a researcher and a participant. Each participant listened to the sentences one at a time on a cassette recorder, completed an answer sheet indicating his or her response to the belief statement, and then orally reproduced the sentence, which was audio recorded. The oral narrative test required participants to listen to a narrative and then provide an oral retelling of the narrative, which was recorded on a computer. The timed GJT, the untimed GJT, and the metalinguistic knowledge test were completed individually on a computer in a private office. All of the tests were completed in a single session that lasted approximately 2.5 hours.

The nonnative participants also completed a background questionnaire that contained questions about their mother tongue, age they began to learn English, number of years in an English-speaking country, other languages they had studied, and the kind of instruction in English they had received at school.

Analysis

Descriptive statistics for the five tests were calculated. The reliability of the different test measures was calculated using Cronbach alpha. Pearson product moment coefficients were computed to examine the interrelationships between the various test measures. A principal component factor analysis (SPSS Version 11.5) was then carried out with a view to investigating the predictions about the type of knowledge each test measured. In a two-factor solution, it was predicted that the imitation test, oral narrative test, and the timed GJT would load on one factor (implicit knowledge) and that the untimed GJT and metalinguistic knowledge test would load on the other factor (explicit knowledge). Further factor analyses were carried out when it became clear that the grammatical and ungrammatical sentences were functioning differently in the untimed GJT. These are explained in the Results section. A number of additional analyses were also carried out to examine predictions based on the operationalization of implicit and explicit knowledge summarized in Table 3. These analyses addressed the construct validity of the tests.

Table 5. Reliability measures for the five tests

Test	Items	Participants[a]	Reliability
Imitation	44	91	$\alpha = .88$
Oral narrative	Variable	83	$r = .85$ (interrater agreement)
Timed GJT	68	91	$\alpha = .81$
Untimed GJT	68	91	$\alpha = .83$
Metalinguistic knowledge	41	91	$\alpha = .90$

[a]Only nonnative participants are included in this count.

RESULTS

Table 5 shows the measure of reliability for each test. These varied between .90 for the metalinguistic knowledge test and .81 for the timed GJT, indicating that each test was reliable.

Table 6 presents means and standard deviations for each of the five measures completed by the NSs and L2 learners. The NSs achieved scores close to 100% on all measures except the test of metalinguistic knowledge and the ungrammatical sentences in the timed GJT. Their scores exceeded those of the L2 learners on all measures except metalinguistic knowledge. On the other hand, the L2 learners scored highest on the untimed GJT measures. Interestingly, both the NSs and the L2 learners scored markedly higher on the grammatical than on the ungrammatical sentences in the timed GJT. Finally, as might be expected, the L2 learners manifested considerably greater intergroup variance than the NSs on all the tests, as reflected in the standard deviations.

Table 7 shows the correlation matrix for the L2 learners' performance on the five tests. Each possible pair of tests was intercorrelated, with coefficients that reached statistical significance at the .05 level or higher. However, the correlation between metalinguistic knowledge and the other tests was gen-

Table 6. Descriptive statistics for the five tests

Test	Native speakers			Nonnative speakers		
	%	*SD*	*n*	%	*SD*	*n*
Imitation	94	4.1	20	51	17.20	91
Oral narrative	99	2.1	15	72	14.25	83
Timed GJT	80	10.02	18	54	11.80	91
Untimed GJT	96	1.55	19	82	10.50	91
Metalinguistic knowledge	57	7.37	20	53	20.73	91

Table 7. Correlational matrix for the five tests (L2 learners)

Test	Imitation $n = 91$	Oral narrative $n = 83$	Timed GJT $n = 91$	Untimed GJT $n = 91$	Metalinguistic knowledge $n = 91$
Imitation	—	.48**	.58**	.59**	.28**
Oral narrative		—	.36**	.36**	.27*
Timed GJT			—	.57**	.24*
Untimed GJT				—	.60**
Metalinguistic knowledge					—

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

erally not as strong as the correlations found for the pairings between the other tests.

Table 8 shows the eigenvalues of the two factors, whereas Table 9 shows the results of a principal component factor analysis of the L2 learners' test scores. A decision was made to specify a two-factor solution. This decision followed from the original design of the tests, which was to measure two distinct constructs, as well as from an inspection of the eigenvalues for the first two factors. Thus, although the eigenvalue for the second factor was below 1.0 (.822), it accounted for a substantial increase in the shared variance (i.e., 16.4%). Overall, the two factors accounted for 74.6% of the total variance. The imitation test, oral narrative test, and timed GJT all loaded heavily at .7 or higher on Factor 1. The untimed GJT and the metalinguistic test loaded heavily on Factor 2 (i.e., higher than .7).

A decision was then made to examine the psychometric properties of the grammatical and ungrammatical sentences in the untimed GJT separately. This was motivated in part by the fact that the untimed GJT loaded quite strongly on Factor 1 (.522) and on Factor 2 (.730) as well as by previous research (Bialystok, 1979; Hedgcock, 1993), which has pointed to the fact that L2 learners respond differently to the grammatical and ungrammatical sentences in a GJT. Hedgcock, for example, commented that although "it would be ill-advised to

Table 8. Principal component factor analysis

Component	Eigenvalue	Variance[a]	Cumulative[a]
1	2.909	58.187	58.187
2	0.822	16.437	74.624

[a]Percentage.

Table 9. Loadings for principal component factor analysis

Test	Component 1	Component 2
Imitation	.824	.270
Oral narrative	.805	.065
Timed GJT	.721	.357
Untimed GJT	.522	.730
Metalinguistic knowledge	.103	.929

claim that subjects rely on *different* [italics added] L2 data bases or cognitive processes in approving well-formed strings and in rejecting ungrammatical strings," nevertheless "such a possibility is not entirely implausible" (p. 15). He then went on to suggest that "positing autonomous L2 knowledge systems . . . is an attractive way of accounting for variable performance across learners and tasks." Pearson product moment coefficients were calculated between the grammatical and ungrammatical sentences in the untimed GJT and all other test measures. The results are shown in Table 10. The grammatical sentences score correlated significantly with the other tests but more strongly with the imitation test, oral narrative test, and timed GJT than with the metalinguistic knowledge test. In contrast, the ungrammatical sentences score correlated strongly with the metalinguistic knowledge test ($r = .67$) and less strongly with the other tests, especially the imitation and oral narrative tests. A second factor analysis was then computed, substituting the scores for the ungrammatical sentences in the untimed GJT for the total untimed GJT scores (as in Table 9). This decision was taken because the correlational matrix in Table 10 showed that the ungrammatical sentences measure was more clearly related to the metalinguistic knowledge score than to the imitation and oral narrative

Table 10. Correlations between scores for the grammatical and ungrammatical sentences in the untimed GJT and other test measures

	Untimed GJT	
Test	Grammatical	Ungrammatical
Imitation	.58**	.38**
Oral narrative	.37**	.26*
Timed GJT	.62**	.33**
Metalinguistic knowledge	.27*	.63**

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

Table 11. Principal component factor analysis

Component	Total	Variance[a]	Cumulative[a]
1	2.719	54.389	54.389
2	0.982	19.637	74.026

[a]Percentage.

Table 12. Loadings for principal component factor analysis

Test	Component 1	Component 2
Imitation	.825	.253
Oral narrative	.803	.091
Timed GJT	.760	.262
Untimed GJT (ungrammatical items)	.265	.869
Metalinguistic knowledge	.161	.898

scores. Again, a two-factor solution was specified. The eigenvalues for the two factors are shown in Table 11 and the results of the factor analysis are given in Table 12. The imitation test, oral narrative test, and timed GJT all load at .75 or higher on Factor 1. The loading of both the untimed GJT (ungrammatical items) and the metalinguistic knowledge test are below .3. Scores for both of these tests load at higher than .85 on Factor 2, with loadings for all the other tests below .3. The cumulative variance in test scores accounted for by these two factors is very similar to that of the first factor analysis (i.e., 74%).

Hypotheses

Subsequent analyses explored hypotheses concerning the relationship of specific characteristics of the learners and the test items to implicit and explicit knowledge. The properties investigated were based on the operationalization of implicit and explicit knowledge summarized in Table 2. For the purposes of these analyses, tests of implicit knowledge were assumed to be those loading on Factor 1 in Table 12, and tests of explicit knowledge were assumed to be those loading on Factor 2.

Hypothesis 1: Tests of Explicit Knowledge Will Encourage the Use of Rule, Whereas Tests of Implicit Knowledge Will Favor Feel. To test this hypothesis, Pearson product moment coefficients of correlation were computed between the measure of the learners' application of rule in the untimed GJT and all the other measures. It was predicted that rule would correlate more strongly with

Table 13. Correlations between use of rule and the test measures

Test	Rule
Imitation	−.03
Oral narrative	.03
Timed GJT (grammatical items)	.12
Timed GJT (ungrammatical items)	.07
Untimed GJT (grammatical items)	.08
Untimed GJT (ungrammatical items)	.32*
Metalinguistic knowledge	.37*

*p < .05

accuracy of judgment in the untimed GJT (ungrammatical items) and also with scores on the metalinguistic judgment test than with scores on the imitation test, oral narrative test, and the timed GJT (grammatical and ungrammatical items). Table 13 shows the results of this analysis. This supports the hypothesis. Low correlations between rule and the measures of implicit knowledge were found, whereas statistically significant correlations at the .01 level were observed between rule and untimed GJT (ungrammatical items) and metalinguistic knowledge. Rule, however, was not related to untimed GJT (grammatical items), but, as we have already seen, this did not constitute a strong measure of explicit knowledge.

Hypothesis 2: Time-Pressured Tests Will Require Learners to Rely on Their Implicit Knowledge, Whereas Tests Without Time Constraints Will Permit Learners to Draw on Their Explicit as Well as Their Implicit Knowledge. The time-pressured tests were the imitation test, oral narrative test, and the timed GJT, whereas the tests without time pressure were the untimed GJT and the metalinguistic test. As we saw in Table 12, the principal component factor analysis indicated that the unpressured and pressured tests loaded on different factors. If we assume that, in general, learners will perform better on the unpressured tests than the pressured tests because they will be able to supplement their implicit knowledge with their explicit knowledge, a difference in mean scores on the two groups of tests can be expected. The mean score for all learners' performance on the pressured tests was 57.3%, and on the unpressured tests, it was 65.9%. This difference was statistically significant, $t(84) = 4.54$, $p < .01$. The effects of time pressure can be assessed best by comparing the timed and untimed GJTs, as these tests were otherwise identical in content and method. The mean score for the timed GJT was 53.9%, whereas the average for the untimed GJT was 82.1%. Again, this difference was statistically significant, $t(87) = 12.60$, $p < .001$.[8]

Hypothesis 3: Tests That Require Learners to Focus on Meaning Will Elicit Implicit Knowledge, Whereas Tests That Encourage Learners to Focus on Form Will Elicit Explicit Knowledge. Two tests required a focus on meaning: the imi-

tation and oral narrative tests. Both tests loaded heavily on Factor 1 in the principal component factor analyses reported in Tables 9 and 12. The two tests that required a focus on form showed less consistent results; the timed GJT also loaded on Factor 1, but less heavily, whereas the untimed GJT (ungrammatical items) loaded heavily on Factor 2. However, this hypothesis cannot be properly tested in this study, as the focus and time pressure variables were confounded in the design of the tests.

Hypothesis 4: Tests of Implicit Knowledge Will Elicit More Systematic (Less Variable) Responses Than Tests of Explicit Knowledge. This hypothesis was tested by inspecting the standard deviations for the different measures. The hypothesis predicts that the tests of implicit knowledge would result in lower standard deviations than the tests of explicit knowledge. Table 6 shows that, on average, the standard deviations were in fact higher on the tests of explicit knowledge, especially in the case of the metalinguistic knowledge test. In line with this hypothesis, it is expected that the standard deviations for the untimed GJT, which tested explicit knowledge, will be higher than the standard deviation for the timed GJT. However, a direct comparison of the standard deviations of the timed and untimed GJTs shows a higher standard deviation in the former (i.e., 11.80 vs. 10.50), possibly because the time pressures for this test induced random behavior. Thus, the evidence does not provide clear support for this hypothesis.

Hypothesis 5: Tests of Implicit Knowledge Will Elicit More Certain Responses from Learners Than Tests of Explicit Knowledge. The untimed GJT asked participants to indicate the degree to which they were certain of their judgments using a percentage scale. Given that the grammatical sentences correlated more strongly with the measures of implicit knowledge and the ungrammatical sentences with the measures of explicit knowledge, a comparison of the certainty judgments for the two sets of sentences allows us to test this hypothesis. Pearson product moment correlations between the measure of certainty and scores for the grammatical and ungrammatical sentences were computed. Both coefficients were statistically significant at the .01 level ($r = .32$ for the grammatical and $r = .31$ for the ungrammatical sentences). These correlations, therefore, do not indicate that the participants were more certain of their responses to the grammatical than the ungrammatical items and, thus, they do not support the hypothesis. To further test the hypothesis, correlations between the participants' reported use of rule and their certainty scores for each item in the untimed GJT were calculated. In accordance with the hypothesis under consideration, it was predicted that these correlations would generally be negative or very low (i.e., the participants would tend to be less certain when they used an explicit rule to make a judgment). However, most of the correlations (i.e., 55 coefficients out of 68) between certainty and rule were statistically significant at the .05 level or higher, indicating a generally strong relationship between the participants' level of certainty and their use of explicit knowledge in this test. Overall, then, these results do not support this hypothesis.

Hypothesis 6: Tests of Explicit Knowledge Will Make Fuller Use of Metalinguistic Knowledge Than Tests of Implicit Knowledge. The correlations reported in Tables 7 and 10 lend support to this hypothesis. Scores on the metalinguistic knowledge test were more strongly related to scores on the untimed GJT (grammatical and ungrammatical items), $r = .60$, and to untimed GJT (ungrammatical items), $r = .64$, than to scores on the imitation test, $r = .28$, the oral narrative test, $r = .27$, and the timed GJT, $r = .24$. However, it should be noted that although the correlations between the test of metalinguistic knowledge and the tests that I claim measure implicit knowledge were weak, nonetheless, they were statistically significant.

Hypothesis 7: Scores on Tests of Implicit Knowledge Will Relate More Strongly to the Age Learners Started Learning the L2 Than to Years of Classroom Instruction, Whereas the Opposite Will Be the Case for Scores on Tests of Explicit Knowledge. This hypothesis relates to the learnability criterion in Table 1. Table 14 shows the correlations between the variables *starting age* and *years of formal instruction* and the different test measures. Starting age was related negatively to the timed GJT (i.e., the older learners were when they began learning, the less well they performed on this test). However, the correlations between starting age and the other tests deemed to measure implicit knowledge (the imitation and oral narrative tests) did not reach statistical significance. Correlations between starting age and the measures of explicit knowledge were all nonsignificant and very weak. In contrast, years of formal instruction was positively related to untimed GJT (ungrammatical items) but not to the other measure of explicit knowledge (metalinguistic knowledge). No statistically significant relationship was observed between this variable and the measures of implicit knowledge. In general, these results support this hypothesis.

Table 14. Correlations between starting age and years of formal instruction and test measures

	Variables	
Test	Starting age[a]	Formal instruction[a]
Imitation	−.11	.10
Oral narrative	.03	−.05
Timed GJT	−.23*	.17
Untimed GJT (grammatical items)	−.05	.03
Untimed GJT (ungrammatical items)	−.02	.27*
Metalinguistc knowledge	.05	.18

[a] In years.
*$p < .05$

DISCUSSION

The main purpose of this study was to demonstrate that tests could be designed to provide relatively separate measures of L2 implicit and explicit knowledge that were reliable and valid. To this end, operational definitions of the two types of knowledge were constructed. These served to draw up the specifications for five tests. With a view to establishing the validity of these specifications, a number of hypotheses based on the operational definitions were formed and tested using the scores obtained from the tests.

The reliability of four of the tests was verified by means of the internal consistency of responses to the items that made up each test. The Cronbach alpha coefficients all exceeded .80 (generally considered to demonstrate a satisfactory level of reliability in social science research). The reliability of the oral narrative test was determined by means of interrater agreement. This measure was also above .80.

A comparison of the performance of the NSs and L2 learners on the five tests lends further support to the overall validity of the five tests. Whereas NSs can be expected to possess higher levels of implicit knowledge than L2 learners, they cannot necessarily be expected to demonstrate higher levels of explicit knowledge, as L2 learners might have benefited in this respect from formal instruction. The results show that the NSs outperformed the L2 learners on the three tests that measured implicit knowledge (the imitation test, the oral narrative test, and the timed GJT). They also outperformed the L2 learners on the untimed GJT (grammatical and ungrammatical items), a test originally designed to measure explicit knowledge, but the difference in scores on this test was much smaller than the differences found on the tests of implicit knowledge. Further, the NSs and L2 learners performed very similarly on the other test of explicit knowledge, the metalinguistic knowledge test.

The L2 learners' scores on all five tests were intercorrelated (see Tables 7 and 10). However, the shared variance between any pair of tests did not exceed 45% and was as low as 6.4%. Overall, then, the correlations do not support Oller's (1979) claim that L2 proficiency is unitary in nature. Furthermore, the factor analyses reported in Tables 9 and 12 demonstrate that the tests measure two different constructs. As predicted, test scores loaded largely on two factors: the imitation test, the oral narrative test, and the timed GJT on one factor and the untimed GJT and the metalinguistic knowledge test on the other factor. These analyses suggest that the primary purpose of the study has largely been achieved; the tests provide relatively separate measures of implicit and explicit knowledge. The imitation test and the metalinguistic knowledge test can be seen as the tests that best measure implicit and explicit knowledge, respectively (i.e., they load the heaviest on their respective factors). I would also argue that the two tests of explicit knowledge measure the two components of this construct—analyzed knowledge and metalanguage—although additional analyses are necessary for an adequate demonstration of this proposal. The two factors account for a remarkably high proportion of the total

variance in the test scores (nearly 75%), lending support to a model of grammatical proficiency based on the distinction between the two types of knowledge represented by these factors (implicit and explicit). Further, the factor with heavy loadings for the tests of implicit knowledge is clearly dominant in the sample of measures obtained for this study (accounting for 58% of the total variance). This lends some support to the claim of certain current theories of L2 acquisition (see N. Ellis, 2002; Krashen, 1981) that implicit knowledge is primary. The factor with heavy loadings for the tests of explicit knowledge is clearly secondary in the sample of measures, failing to achieve an eigenvalue of 1 and accounting for far less of the shared variance (16.4%). In short, there is a congruence among the results of the factor analyses, the constructs underlying the test specifications, and SLA theory.

The results of the two factor analyses also point to the need to distinguish between the grammatical and ungrammatical sentences in the GJTs. Particularly in the case of the untimed GJT, the grammatical and ungrammatical sentences appear to measure different constructs; grammatical sentences draw on implicit knowledge, whereas ungrammatical sentences tap into explicit knowledge. A more detailed analysis of the GJT scores from this study (see Loewen, 2003) indicated that they differ significantly on two dimensions (timed vs. untimed and grammatical vs. ungrammatical). This has important implications for the use of this kind of test in SLA research. In particular, it suggests that SLA researchers need to take great care to distinguish these two properties in both the design of GJTs and the analysis of scores obtained from tests, as they will influence what is measured.

To demonstrate the validity of the test constructs, a number of hypotheses were investigated. In general, the results of these investigations support the construct validity of the tests. Thus, both tests of explicit knowledge were strongly related to the learners' reported use of rule in the untimed GJT, whereas the tests of implicit knowledge were only weakly related to this measure. The three tests that imposed time pressure all loaded on the implicit knowledge factor, whereas the two unpressured tests loaded on the explicit knowledge factor. In the examination of Hypothesis 4, the difference between the standard deviations of the oral imitation and metalinguistic knowledge tests lent some support to the claim that there is greater systematicity in learners' implicit knowledge. With respect to Hypothesis 6, the untimed GJT (ungrammatical items), a measure of explicit knowledge, was more strongly related to metalinguistic knowledge than the tests that were shown to measure implicit knowledge. Finally, in the discussion of Hypothesis 7, it was demonstrated that one of the tests of implicit knowledge (the timed GJT) was related to learners' starting age, whereas the test of explicit knowledge (the untimed GJT [ungrammatical items]) was related to the number of years of formal instruction. Only one of the seven hypotheses investigated was not supported; the learners appeared to be more certain of their responses to the test items when they had access to their explicit knowledge. This might reflect the fact that many of the participants—especially those with lower levels of

proficiency—lacked confidence in their implicit knowledge of many of the grammatical structures tested, as many of these are known to be late acquired (e.g., question tags and hypothetical conditionals). Also, Hypothesis 3, which was concerned with the distinction between focus on form and meaning, could not be properly tested. Overall, however, the construct validity of the tests receives empirical support from the analyses of the scores obtained. It is clear, however, that further work is needed to provide a clear validation of each design feature. The work undertaken in this study is to be seen as exploratory in this respect.

CONCLUSION

There is an obvious need in both SLA and language testing to construct convincing models of L2 proficiency and, taking these models as a starting point, to develop instruments capable of providing reliable and valid measurements of L2 knowledge. The present study is an attempt to meet this need.

In SLA, irrespective of one's theoretical orientation, it is important to be able to distinguish between learners' implicit and explicit knowledge of an L2. Until this differentiation is achieved, it will not be possible to test the interface and noninterface hypotheses that lie at the center of much current debate in SLA. Surprisingly, however, SLA researchers have made few attempts to develop instruments capable of distinguishing these two types of knowledge. Indeed, as Douglas (2001) remarked, researchers have conspicuously failed to make the effort to demonstrate the validity (and reliability) of their testing instruments. This constitutes a major weakness in the discipline.

In contrast, in language testing there has been a constant and sophisticated examination of the reliability and construct validity of instruments designed to measure language proficiency. However, the models of L2 proficiency that have informed test construction have generally not been supported by analyses of the tests designed to investigate the models. Oller's (1979) unitary competence hypothesis, which claimed that language proficiency is comprised of a single underlying construct (pragmatic expectancy grammar), was rejected on two grounds: Oller failed to include an oral test of proficiency, and the factorial analyses he employed were inconclusive (see Baker, 1989, for discussion). Subsequent attempts to validate models of proficiency based on a modular view of communicative competence have not fared much better. For example, Harley, Allen, Cummins, and Swain (1990) examined the validity of Canale and Swain's (1980) model of communicative competence by developing a battery of tests designed to measure different components of competence (grammatical, discourse, and sociolinguistic) with three different methods (oral, written, and multiple choice). However, a confirmatory factor analysis failed to support the model. Attempts to build models of proficiency based on the construct of ability to use as mediating between underlying competence and performance conditions (e.g., Bachman, 1990; Bach-

man & Palmer, 1996) have also failed to find clear empirical support. More recently, Skehan (1998) attempted to build a psycholinguistic model of proficiency that incorporates both a language dimension (where lexically based and rule-based knowledge of language are distinguished) and a language-processing dimension (based on a limited-capacity short-term memory). Furthermore, Skehan's model attempted to explore how different tasks affect the fluency, complexity, and accuracy of learners' production. However, Iwashita, Elder, and McNamara's (2001) attempt to validate Skehan's model in the context of a tape-based test failed to show the expected differences in the quality of language produced when various dimensions of tasks were manipulated. Obviously, the choice of model of language proficiency to serve as the basis for the development of language tests must take into account a number of factors (e.g., the purpose of the test, the target domain of language use, and the likely backwash effect). However, one such factor ought to be the psycholinguistic validity of the underlying model, as this can be demonstrated empirically. In this respect, the models referred to in this section have not been successful.

The results of this study are of potential significance to the fields of SLA as well as language testing. They demonstrate that it might be possible to develop tests that will provide relatively separate measures of implicit and explicit knowledge. If subsequent research confirms this, SLA will have available the necessary instruments to investigate issues of central theoretical importance in the study of L2 acquisition. For language testers, this study points to an alternative model of L2 proficiency, drawn from SLA and provisionally supported by the results of the factor analyses reported previously. Of particular interest here is the extent to which the kinds of tests of *grammatical* proficiency used in this study are predictive of *general* language proficiency (as measured by recognized proficiency tests such as TOEFL and IELTS).[9] This constitutes an obvious direction for future inquiry.

NOTES

1. It could be argued (as one of the anonymous *SSLA* reviewers pointed out) that linguistic *knowledge* is a mentalistic concept and thus not appropriate as a label for the kind of linguistic representation posited by connectionist models. Nevertheless, *knowledge* is a widely used term by connectionist theorists such as N. Ellis (1996).

2. Green and Hecht's (1992) finding that there was a gap between their learners' ability to correct errors and to verbalize the rules involved was interpreted as a reflection of a difference between implicit and explicit knowledge. However, another equally valid interpretation is that this difference reflected the disparity between what the learners knew explicitly and what they could actually verbalize.

3. *Learnability* has a technical sense in theories of UG. However, the term is used in a more general sense here to refer to the extent to which knowledge can be internalized by learners at whatever stage of development they have reached.

4. One of the anonymous *SSLA* reviewers pointed out that I have used the term *grammaticality judgment test* throughout although, in actuality, the tests referred to call for judgments of acceptability. Although acknowledging this, I have decided to continue to use *grammaticality* in accordance with the previously published literature.

5. Not all the participants completed all tests. The actual numbers completing each test are shown in Table 5.

6. An anonymous *SSLA* reviewer expressed concern that the participants were not asked to correct the sentences in the timed GJT, thus making it difficult to know exactly what they were responding to in the sentences. This must be acknowledged as a weakness of the test. However, in piloting the test, it was felt that the pressured nature of test made it extremely demanding and that to also require participants to also produce corrections would have overloaded the resources of many of them.

7. The other three criteria were systematicity, certainty, and learnability. Systematicity did not inform the design of any of the tests although it was examined in a post hoc analysis of the test scores (see Results section). Certainty was a design feature in only one of the tests (the untimed GJT) and, for this reason, is not included in Table 4. Learnability is a characteristic of learners, which was also considered post hoc.

8. In fact, a detailed analysis suggests that time pressure in the two GJTs interacted with the grammaticality of the sentences. A univariate ANOVA found a significant difference in the four sets of scores, $F(3) = 253.33$, $p < .001$, whereas a post hoc Scheffe test indicated three subsets: (a) timed GJT (ungrammatical items), (b) timed GJT (grammatical items) and untimed GJT (ungrammatical items), and (c) timed GJT (ungrammatical items) and untimed GJT (grammatical items). In other words, the learners' responses in the GJTs were not solely the product of time pressure.

9. Han and Ellis (1998) reported statistically significant correlations between their tests and standard measures of L2 proficiency (IELTS and SPEAK Test).

REFERENCES

Alderson, J., Clapham, C., & Steel, D. (1997). Metalinguistic knowledge, language aptitude, and language proficiency. *Language Teaching Research*, *1*, 93–121.

Anderson, J. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

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

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

Baker, D. (1989). *Language testing: A critical survey and practical guide*. London: Arnold.

Bialystok, E. (1979). Explicit and implicit judgments of L2 grammaticality. *Language Learning*, *29*, 81–103.

Bialystok, E. (1982). On the relationship between knowing and using forms. *Applied Linguistics*, *3*, 181–206.

Bialystok, E. (1991). Achieving proficiency in a second language: A processing description. In R. Phillipson, E. Kellerman, L. Selinker, M. Sharwood Smith, & M. Swain (Eds.), *Foreign/second language pedagogy research* (pp. 63–78). Clevedon, UK: Multilingual Matters.

Bialystok, E. (1994). Representation and ways of knowing: Three issues in second language acquisition. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 549–569). San Diego, CA: Academic Press.

Breen, M. (1989). The evaluation cycle for language learning tasks. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 187–206). New York: Cambridge University Press.

Burt, M., & Kiparsky, C. (1972). *The gooficon: A repair manual for English*. Rowley, MA: Newbury House.

Butler, Y. (2002). Second language learners' theories on the use of English articles. *Studies in Second Language Acquisition*, *24*, 451–480.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*, 1–47.

Chomsky, N. (1976). *Reflections on language*. London: Temple Smith.

Coughlan, P., & Duff, P. A. (1994). Same task, different activities: Analysis of a SLA task from an activity theory perspective. In J. Lantolf & G. Appel (Eds.), *Vygotskian approaches to second language research* (pp. 173–194). Westport, CT: Ablex.

DeKeyser, R. M. (1995). Learning second language grammar rules: An experiment with a miniature linguistic system. *Studies in Second Language Acquisition*, *17*, 379–410.

DeKeyser, R. M. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. J. Doughty & J. Williams (Eds.), *Focus on form in second language acquisition* (pp. 42–63). New York: Cambridge University Press.

DeKeyser, R. M. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *Handbook of second language learning* (pp. 313–348). Oxford: Blackwell.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, *22*, 735–808.

Douglas, D. (2001). Performance consistency in second language acquisition and language testing: a conceptual gap. *Second Language Research*, *17*, 442–456.

Eichenbaum, H. (1997). Declarative memory: Insights from cognitive neurobiology. *Annual Review of Psychology*, *48*, 547–572.

Ellis, N. C. (1994). Introduction: Implicit and explicit language learning—An overview. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 1–31). San Diego, CA: Academic Press.

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

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188.

Ellis, R. (1985). Sources of variability in interlanguage. *Applied Linguistics*, *6*, 118–131.

Ellis, R. (1991). Grammaticality judgments and learner variability. In H. Burmeister & P. Rounds (Eds.), *Variability in second language acquisition: Proceedings of the tenth meeting of the Second Language Acquisition Forum* (pp. 25–60). Eugene, OR: Department of Linguistics, University of Oregon.

Ellis, R. (1993). Second language acquisition and the structural syllabus. *TESOL Quarterly*, *27*, 91–113.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Ellis, R. (2002). Does form-focused instruction affect the acquisition of implicit knowledge? A review of the research. *Studies in Second Language Aquisition*, *24*, 223–236.

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

Green, P., & Hecht, K. (1992). Implicit and explicit grammar: An empirical study. *Applied Linguistics*, *13*, 168–184.

Gregg, K. R. (1989). Second language acquisition theory: The case for a generative perspective. In S. M. Gass & J. Schachter (Eds.), *Linguistic perspectives on second language acquisition* (pp. 15–40). New York: Cambridge University Press.

Gregg, K. R. (2003). The state of emergentism in second language acquisition. *Second Language Research*, *19*, 95–128.

Han, Y., & Ellis, R. (1998). Implicit knowledge, explicit knowledge, and general language proficiency. *Language Teaching Research*, *2*, 1–23.

Harley, B., Allen, P., Cummins, J., & Swain, M. (Eds.). (1990). *The development of second language proficiency*. New York: Cambridge University Press.

Hedgcock, J. (1993). Well-formed versus ill-formed strings in L2 metalingual tasks: Specifying features of grammaticality judgments. *Second Language Research*, *9*, 1–21.

Hu, G. (2002). Psychological constraints on the utility of metalinguistic knowledge in second language production. *Studies in Second Language Acquisition*, *24*, 347–386.

Hulstijn, J. H. (2002). Towards a unified account of the representation, processing and acquisition of second language knowledge. *Second Language Research*, *18*, 193–223.

Hulstijn, J. H., & Hulstijn, W. (1984). Grammatical errors as a function of processing constraints and explicit knowledge. *Language Learning*, *34*, 23–43.

Ioup, G. (1996). Grammatical knowledge and memorized chunks: A response to Ellis. *Studies in Second Language Acquisition*, *18*, 355–360.

Iwashita, N., Elder, C., & McNamara, T. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, *51*, 401–436.

James, C., & Garrett, P. (1992). The scope of awareness. In C. James & P. Garrett (Eds.), *Language awareness in the classroom* (pp. 3–20). London: Longman.

Karmiloff-Smith, A. (1979). Micro- and macro-developmental changes in language acquisition and other representation systems. *Cognitive Science*, *3*, 91–118.

Krashen, S. (1977). Some issues relating to the Monitor Model. In H. Brown, C. Yorio, & R. Crymes (Eds.), *On TESOL '77* (pp. 144–158). Washington, DC: TESOL.

Krashen, S. (1981). *Second language acquisition and second language learning*. London: Pergamon.

Krashen, S. (1982). *Principles and practice in second language acquisition*. London: Pergamon.

Lantolf, J. (2000). Introducing sociocultural theory. In J. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 1–26). Oxford: Oxford University Press.

Loewen, S. (2003, October). *Grammaticality judgment tests: What do they really measure?* Paper presented at the Second Language Research Forum, Tucson, AZ.

Macrory, G., & Stone, V. (2000). Pupil progress in the acquisition of the perfect tense in French: The relationship between knowledge and use. *Language Teaching Research*, *4*, 55–82.

Major, R. C. (1996). Chunking and phonological memory: A response to Ellis. *Studies in Second Language Acquisition*, *18*, 351–354.

Oller, J. (1979). *Language tests at school*. London: Longman.

Paradis, M. (1994). Neurolinguistic aspects of implicit and explicit memory: Implications for bilingualism and SLA. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 393–419). San Diego, CA: Academic Press.

Pienemann, M. (1989). Is language teachable? Psycholinguistic experiments and hypotheses. *Applied Linguistics*, *10*, 52–79.

Reber, A., Walkenfeld, F., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 888–896.

Rumelhart, D., & McClelland, J. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations*. Cambridge, MA: MIT Press.

Schmidt, R., & Frota, S. (1986). Developing basic conversational ability in a second language: A case-study of an adult learner. In R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 237–326). Rowley, MA: Newbury House.

Seliger, H. (1979). On the nature and function of language rules in language teaching. *TESOL Quarterly*, *13*, 359–369.

Sharwood Smith, M. (1981). Consciousness-raising and the second language learner. *Applied Linguistics*, *2*, 159–169.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Sorace, A. (1985). Metalinguistic knowledge and language use in acquisition-poor environments. *Applied Linguistics*, *6*, 239–254.

Sorace. A. (1996). The use of acceptability judgments in second language acquisition research. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 375–409). San Diego, CA: Academic Press.

Tarone, E. (1988). *Variation in interlanguage*. London: Arnold.

Wexler, K., & Mancini, R. (1987). Parameters and learnability in binding theory. In T. Roeper & E. Williams (Eds.), *Parameter setting* (pp. 41–76). Dordrecht: Reidel.

Zobl, H. (1995). Converging evidence for the 'acquisition-learning' distinction. *Applied Linguistics*, *16*, 35–56.