6

Assessing Grammar

James E. Purpura
Teachers College, Columbia University, USA

Introduction

Although it is generally accepted that much of second language acquisition (SLA) happens incidentally while learners are focused on meaningful input and engaged in interaction, the explicit teaching of a second or foreign language (L2) and the assessment of a learner's development of grammatical ability have always been of critical concern for L2 educators. This interest in grammar is bolstered by findings in SLA, showing that while all instruction does not impact learning positively, learners receiving explicit, form-based instruction are more likely to optimize natural learning processes, develop grammatical ability at more accelerated rates, and achieve higher levels of L2 proficiency (Ellis, 2008) than learners not receiving form-focused instruction. This is especially so if L2 input is rich, abundant, and meaningful; grammar explanations and corrective feedback summon awareness of patterns previously undetected; and instruction is sequenced to promote processing and skill acquisition.

L2 testers have also acknowledged the importance of grammar in assessing communicative language ability (Purpura, 2004). Interest in grammar assessment stems from the fundamental role that it plays in predicting the ability to communicate precisely and effectively in the L2, and from the potential it has for providing learners and teachers with information, at various grain sizes, on the grammar needed to improve. Several researchers (e.g., Hulstijn, Schoonen, de Jong, Steinel, & Florijn, 2012) are also interested in grammar assessment for the potential it offers in helping to characterize L2 knowledge in different contexts, or at diverse proficiency levels, as referenced by some external standard, framework, or proficiency scale. Finally, interest in grammar assessment has increased considerably as a result of the potential role that grammatical

The Companion to Language Assessment, First Edition. Edited by Antony John Kunnan.

DOI: 10.1002/9781118411360.wbcla147

features play in developing speech and writing recognition and processing technologies on the one hand, and automated scoring and feedback systems of L2 assessments on the other (Xi, 2010).

The current chapter examines how grammatical assessment has been conceptualized, implemented, and researched over the years. It also discusses challenges and future directions of grammar assessment.

Previous Conceptualizations of Language Knowledge and Research: Focusing on Grammar

While grammar as a construct has been conceptualized in many different ways with reference to one or more linguistic frameworks (e.g., structural linguistics), L2 educators have generally defined "grammar" as a set of structural rules, patterns, norms, or conventions that govern the construction of well-formed and meaningful utterances with respect to specific language use contexts. And most L2 educators would agree that the ability to generate well-formed and meaningful utterances in context-rich or impoverished situations (e.g., a traditional discrete-point grammar test) depends on a range of linguistic resources involving phonology, morphology, syntax, semantics, discourse, and pragmatics.

Drawing on eclectic but principled descriptions of grammar for educational purposes, several L2 testers have proposed conceptualizations of L2 proficiency in which grammatical knowledge has played a consistently prominent role. The resulting conceptualizations of grammatical knowledge have then been used as a basis for constructing grammar assessments. In other words, they have been used to describe how grammatical knowledge might be represented in a learner's head, described at different proficiency levels, defined with respect to some given assessment purpose, and importantly, conceptualized within a comprehensive framework of L2 proficiency. I will discuss how grammatical knowledge has been defined theoretically in three such conceptualizations before describing four approaches to grammar assessment.

Lado's Conceptualization of Language Knowledge

In an insightful attempt at describing L2 communication, Lado (1961) proposed a model of L2 proficiency in which language is characterized in terms of two individuals who use linguistic *forms* in some variational distribution to create word and sentence *meanings*. These basic elements are then used as resources for communicating cultural and individual meanings. The form–meaning elements for Lado involve phonology, structures, and the lexicon. Cultural meanings refer to concepts or notions associated with a specific culture (e.g., "American breakfast") or speech community (e.g., "business meeting" at a conference). And individual meanings are viewed as outside the culture, referring to the personal associations individuals make with words and concepts (e.g., personal associations with "Christmas"). Lado's depiction of language, culture, and the individual is presented in Figure 6.1.

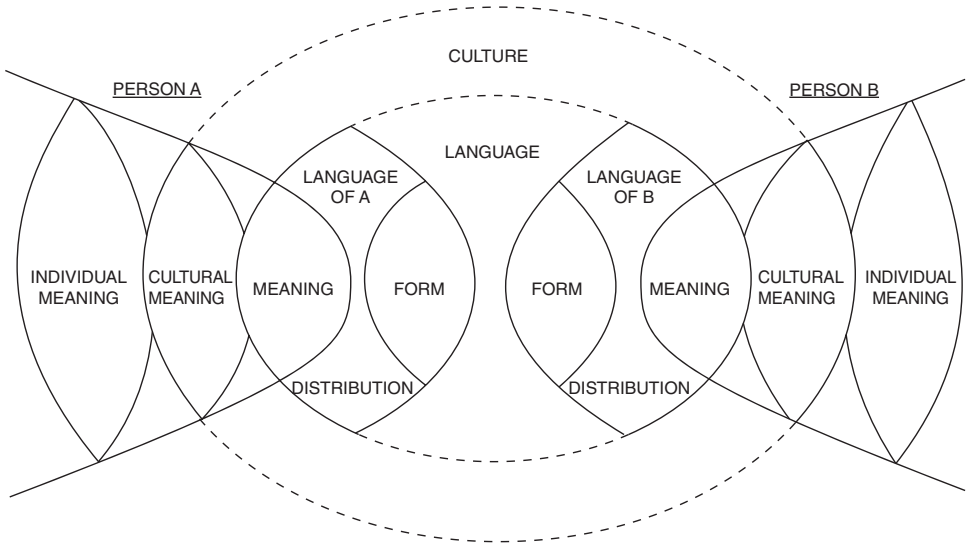

Figure 6.1 **Lado's conceptualization of language knowledge: Language, culture, and the individual (Lado, 1961, p. 6). © Longman. Reprinted with permission**

Lado's view of L2 proficiency was operationalized in terms of a skills-and-elements approach to assessment. This approach viewed L2 knowledge in terms of the language elements (i.e., knowledge of phonology, structures, lexis), measured in the context of the language skills (i.e., reading, writing, speaking, listening). The individual elements were taken to be the principal building blocks of L2 proficiency—the assumption being that L2 proficiency was achieved by internalizing simple, discrete components of the L2 before acquiring more complex units, the accumulation of which constituted "proficiency." This view led to a discrete-point approach to assessment, where discrete linguistic elements (e.g., 20 multiple choice [MC] grammar items) are presented to learners and scored dichotomously for accuracy (e.g., 1 for a right answer, 0 for a wrong one). The scores from the correct responses are then aggregated to produce an overall proficiency estimate.

Probably the best example of a test grounded in Lado's skills-and-elements conceptualization of L2 proficiency is the Comprehensive English Language Test (CELT) (Harris & Palmer, 1986). The grammar subtest assessed five structures: (1) choice of verb forms and modals; (2) form and choice of nouns, pronouns, adjectives, and adverbs; (3) word order; (4) choice of prepositions; and (5) formation of tag questions and elliptical responses. The subtest consisted of 75 discrete-point, MC items with four response options.

The listening section was also organized around different grammatical structures, but assessment focused on the meaning of those structures. For example, the first task aimed to measure the ability to understand *wh-* and *yes/no* questions. The second focused on the comprehension of conditionals, comparisons, and time and number expressions. And the third task targeted the comprehension of lexical

items in two-turn conversations by asking examinees to respond to detail questions (e.g., "on what day"). In sum, the CELT was designed to measure language elements in reading and listening tasks.

Lado's (1961) theoretical conceptualization of proficiency was truly visionary. However, the operationalization of proficiency as knowledge related to discrete structural and lexical items presents a highly restricted view of the construct. Most L2 educators would now want to assess how grammatical forms are associated with a range of semantic meanings, not just lexical meanings, and they would want to target the ability to understand and use pragmatic meanings, where context is a critical resource for meanings specific to a situation. Nonetheless, Lado's approach to grammar assessment remains highly useful for measuring isolated forms, when this is the assessment goal.

In terms of determining what grammatical content to put on grammar tests, Lado (1961) argued that contrastive analysis and transfer from the first language (L1) to the L2 should play a major role in item selection. He maintained that when structures in the L1 and L2 have the same form, meaning, and usage distribution (e.g., the present perfect in French and Italian), learning is assumed to be easier. However, when these features differ across the L1 and L2, the structures are assumed to be more difficult to learn. In sum, Lado believed that L2 assessment should be rooted in SLA theory.

Bachman and Palmer's (1996) Conceptualization of Language Knowledge

Another insightful and well-known conceptualization of L2 proficiency in which grammatical knowledge plays a prominent role was proposed by Bachman and Palmer (1996). They described language use in terms of an interaction between the individual characteristics of the language user on the one hand and the context of language use on the other. The characteristics of the user are further defined as the interaction among an individual's language ability (i.e., language knowledge and strategic competence), topical knowledge (e.g., information on how to book a flight online), and affective schemata (e.g., motivation). Language knowledge is defined in terms of organizational knowledge (involving grammatical and textual knowledge) and pragmatic knowledge (comprising functional and sociolinguistic knowledge). In this framework, grammatical knowledge refers to how individual utterances or sentences are organized with respect to knowledge of phonology or graphology, vocabulary, and syntax. Textual knowledge relates to how utterances or sentences are organized to form texts, and involves knowledge of cohesion and rhetorical or conversational organization. Finally, grammatical and textual knowledge are seen as resources for being able to communicate the goals of a language user in a given L2 use setting. Bachman and Palmer's conceptualization of language knowledge is presented in Figure 6.2.

Bachman and Palmer's model of language knowledge has been used as a heuristic for guiding test development in numerous L2 tests throughout the world, including the Test of English as a Foreign Language (TOEFL) and the Cambridge exams.

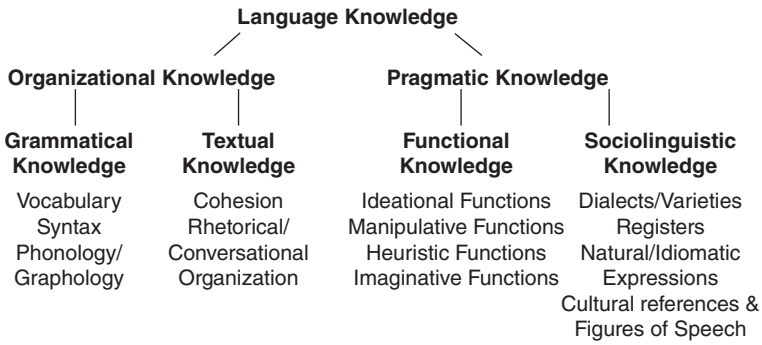

Figure 6.2 Bachman and Palmer's (1996) conceptualization of language knowledge. © Oxford University Press. Reprinted with permission

Current Conceptualizations of Language Knowledge

A more recent depiction of L2 proficiency was proposed by Purpura (2004). His conceptualization of L2 proficiency was inspired by L2 assessment theory, SLA research, and years of experience in L2 teaching and testing. From the L2 assessment perspective, Purpura's conceptualization of L2 proficiency was inspired by the theoretical models of proficiency proposed by Lado (1961), Canale and Swain (1980), Bachman and Palmer (1996), and many others, described in the previous sections. These models helped identify the components of L2 proficiency. Purpura's model was also influenced by Larsen-Freeman's (1991) and Rea-Dickins's (1991) conceptualizations of L2 proficiency as form, meaning, and use in the context of teaching and testing communicative grammar.

From the SLA perspective, L2 proficiency in Purpura's view acknowledges the research on the connections between grammatical forms and their associated semantic meanings (e.g., VanPatten, Williams, Rott, & Overstreet, 2004). Rather than questioning the nature of these two dimensions, SLA research is more concerned with the behavioral and cognitive processes that allow form–meaning mappings to occur and be maintained. Findings from this research have generally shown that low proficiency learners tend to learn simple forms or parts of forms based on the need to communicate lexical meanings (e.g., *going to* vs. *will* to express future time), thereby making learners less likely to process how more complex forms (e.g., *going to*) might encode morphosyntactic meanings such as modality or aspect. Advanced learners, on the other hand, seem more capable of using the linguistic and situational context to connect how forms encode semantic or pragmatic meanings (Bardovi-Harlig, 2000). In sum, as Larsen-Freeman (1991) always reminds us, learners vary on which dimension of grammatical knowledge is acquired on the acquisitional pathway—a finding which, I believe, has serious implications for L2 assessment, and for grammar assessment in particular.

Finally, and just as important, Purpura's conceptualization of L2 proficiency was strongly influenced by years of observing the kinds of linguistic challenges (in terms of forms, meanings, and uses) that learners exhibit in classrooms when attempting to learn an L2 (Purpura & Pinkley, 1991) and on language assessments

when attempting to respond to language tasks—especially as this regards the provision of feedback for formative purposes.

Purpura's Conceptualization of Language Knowledge

Purpura (2004, 2012) describes language knowledge as the interaction between *grammatical knowledge* and *pragmatic knowledge*. Grammatical knowledge is further defined in terms of a range of linguistic forms (e.g., *-s* affix; word order) and semantic meanings associated with these forms, either individually (e.g., plurality with a noun; time reference with a verb) or collectively (e.g., the overall literal meaning of the utterance). These forms and meanings occur at the subsentential, sentential, and suprasentential or discourse levels. Specifically, the forms and meanings can be categorized with respect to (1) phonology or graphology, (2) lexis, (3) morphosyntax, (4) cohesion, (5) information management (e.g., topic or comment), and (6) interaction (e.g., metadiscourse markers like "uh-huh"). In this conceptualization, the form–meaning mappings are assumed to provide fundamental resources for the ability to convey and understand the literal and intended meaning of utterances in L2 use situations. They also provide critical resources for conveying and understanding pragmatic meanings in L2 use, where context plays a major role in interpreting meanings expressed implicitly.

Consider, for example, the form and meaning dimensions of L2 proficiency. The plural *-s* affix added to a noun in English is a grammatical form associated with plurality—its semantic meaning. These two dimensions of the *-s* affix form may present challenges to learners whose L1s use different forms to convey plurality (e.g., Italian uses *-i* or *-e*) or whose L1s have different notions of plurality (e.g., plurality in Arabic treats two entities differently from more than two entities). As a result, English-speaking students learning Italian typically are assumed to have no problem understanding the notion of plurality in Italian, but may encounter challenges using plural forms correctly.

Given learning challenges relating to these two dimensions, it is important for testers to think about test content for grammar assessments in a systematic and principled way, so that specific assessments can be designed for different test purposes. Thus, as described above, we can think of grammar test content in terms of grammatical forms and meanings at the sub(sentential) level (i.e., phonology, lexis, morphosyntax) and at the suprasentential level (i.e., cohesion, information management, interaction). Such a view accommodates both sentence-level and discourse-level spoken and written grammar. Thus, drawing on a comprehensive framework of grammatical knowledge, a tester may choose to measure only the form dimension, understanding that without the meaning dimension, claims can only be made about knowledge of grammatical form, but not about grammatical knowledge in general. In other words, the ability to add the *-ed* affix to verbs does not necessarily mean a learner knows what the past tense verbs mean or how they can be used.

In developing the Oxford Online Placement Test, Purpura and his colleagues used the six categories described above as an organizational frame for creating a taxonomy of test content. They then surveyed English as a second language (ESL) textbooks and pedagogical grammars (e.g., Celce-Murcia & Larsen-Freeman,

Table 6.1 Taxonomy of grammatical forms

Nouns and noun phrases:
- predeterminers, determiners, post-determiners
- nouns (countability, affixation, compounding)

Verbs, verb phrases, tense and aspect:
- tense—present, past; aspect—progressive
- subject–verb agreement

Modals and phrasal modals (*be able to*):
- forms—present, past, future, perfective, progressive
- obligation—*should*, *supposed to*

Phrasal verbs:
- form—two-word, three-word
- separability

Prepositions and prepositional phrases:
- co-occurrence with verb, adjective or noun—*rely on*, *fond of*
- spatial or temporal relationships—*at the store, at 5*

Adjectives and adjectival phrases:
- formation (*-ous*, *-ive*)
- adjective order—*the lovely, little, plastic Cher doll*

Logical connectors:
- relationships of time, space, reason, and purpose
- subordinating and coordinating conjunctions

Relative clauses:
- forms—animate, inanimate, zero, place
- subject noun phrase, (in)direct object noun phrase, genitive noun phrase

Nonreferential *It* and *There*:
- time, distance, environment—*it's noisy in here*
- existence—*there is/are*

Pronouns and reference (cohesion):
- personal, demonstrative, reciprocal
- relative, indefinite, interrogative

Questions and responses:
- yes/no, *wh-*, negative, uninverted
- tags

Conditionals:
- forms—present, past, future
- factual, counterfactual

Passive voice:
- form—present, past, future, perfective
- other passives—*get something done*

Complements and complementation:
- verb + noun phrase + (preposition) noun phrase
- infinitive or gerund complements—*want (him) to*; *believe him to*; *get used to* + gerund

Comparisons:
- comparatives and superlatives
- equatives—*as/so big as*

Adverbials and adverbial phrases:
- forms—adverb phrase, clause, prepositional phrase
- placement—sentence initial, medial, and final

Reported speech:
- backshifting
- indirect imperatives or questions

Focus and emphasis:
- emphasis—emphatic *do*
- marked word order—*him I see*

1999) for grammar points to include in the taxonomy. The resulting taxonomy allowed them to specify what features of grammatical knowledge they wanted on the test, and to balance the content across different categories, so that structures from all the categories could be represented in the test content. A simplified version of this taxonomy appears in Table 6.1.

Besides grammatical knowledge, Purpura's (2004) depiction of L2 proficiency specifies how grammatical forms and their semantic meanings provide resources

for conveying and understanding pragmatic meanings—that is, meanings that occur in language use that are not solely derivable from the literal meanings of words alone or arranged in syntax, but can only be interpreted from a concurrent understanding of the context. For example, the sentence *I'm Italian* changes meanings depending on the context in which it is used. If there were no further context than this sentence (as in many grammar tests), then one would default to the *literal* meaning based on the literal meanings of the words arranged in syntax. The utterance would, therefore, refer to an expression of one's nationality, and would be a plausible response to:

What's your nationality? → *I'm Italian.*
Where are you from? → *I'm Italian.*

The intended, functional meaning of the utterance would be to inform the interlocutor of the speaker's nationality.

In a different context, however, the same sentence could also be a response to:

Do you like red wine? → [smile] *I'm Italian.*
Do you lie about bad pizza? → [condescending look] *I'm Italian.*

In these cases, the response *I'm Italian* would obviously encode more than an expression of nationality. It would simultaneously convey a *sociocultural* association between Italian identity and the presupposition that Italians generally like red wine, or that they are not usually inclined to lie about substandard pizza. Such an utterance could also convey *sociolinguistic meanings* (e.g., informality between friends), and *psychological meanings* (e.g., playfulness). Thus, the utterance *I'm Italian* uses the same grammatical forms to convey literal meaning (i.e., nationality), intended meaning (i.e., to inform), and, other meanings derivable solely from context. Thus, pragmatic meanings are different from, but intrinsically linked to both a learner's grammatical resources and the contextual characteristics of the communicative event.

While this chapter is not specifically about pragmatic knowledge, it is important to distinguish how, in a comprehensive model of L2 proficiency, grammatical forms together with their literal and intended meanings (i.e., grammatical knowledge) provide the fundamental resources for communicating contextual implicatures; metaphor; poetry; social and cultural identity; social and cultural appropriateness—formality, politeness; affective stance—emotionality, irony, humor, sarcasm; and so forth.

Purpura's (2012) theoretical model of language knowledge appears in Figure 6.3.

In order to translate this theoretical model into an organizational framework that can be used flexibly in the design, development, scoring, and validation of grammatical assessments, Purpura proposed an operational model of language knowledge that specifies several types of grammatical forms together with their associated semantic meanings (grammatical knowledge), and a range of possible pragmatic meanings (pragmatic knowledge). The intention was to provide an organized list of features that could be used to design assessments specific to the assessment purpose. In other words, the model could be used to help design and

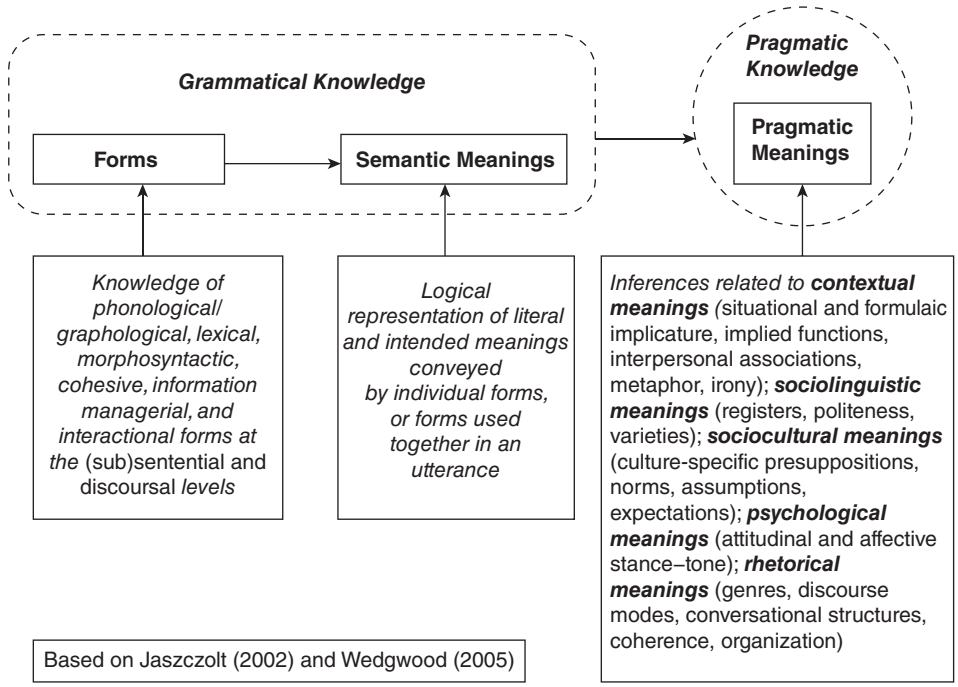

Figure 6.3 Purpura's theoretical model of language knowledge: the grammatical and pragmatic components (based on Purpura, 2012)

score assessments targeting discrete aspects of grammatical knowledge such as lexical forms (e.g., get rid *of*, different *from*) or cohesive meanings (e.g., *therefore*, *however*, *consequently*), should the assessment situation call for it. Or it could be used to design and score grammar assessments targeting the overall meaningfulness of one or more utterances (semantic meaning) and the precision of grammatical resources (forms) used to convey propositions in complex, language use tasks (e.g., the use of the active or passive voice in describing the desalination process). Finally, this model could also serve as a guide for specifying content related to the grammatical and semantic features of L2 production (e.g., accuracy, complexity, meaningfulness, and fluency), or the stages of L2 development (e.g., profiles of features characterizing beginning or advanced learners). Purpura's operational model of language knowledge is presented in Figure 6.4.

While the ultimate goal of grammar assessment is to ascertain a representation of grammatical knowledge in the learner's brain, we need to bear in mind that grammatical knowledge, as one component of language knowledge, combines with many other factors when learners have to use this knowledge to perform tasks involving the four skills. More specifically, grammatical and pragmatic knowledge in a learner's brain (i.e., L2 knowledge) combine with other internal factors (e.g., topical knowledge, sociocognitive ability, personal attributes) to provide the capacity to use this knowledge (L2 ability) to perform tasks (L2 use) involving receptive or productive modalities (L2 skills). The relationships between L2 knowledge, ability, and use appear in Figure 6.5.

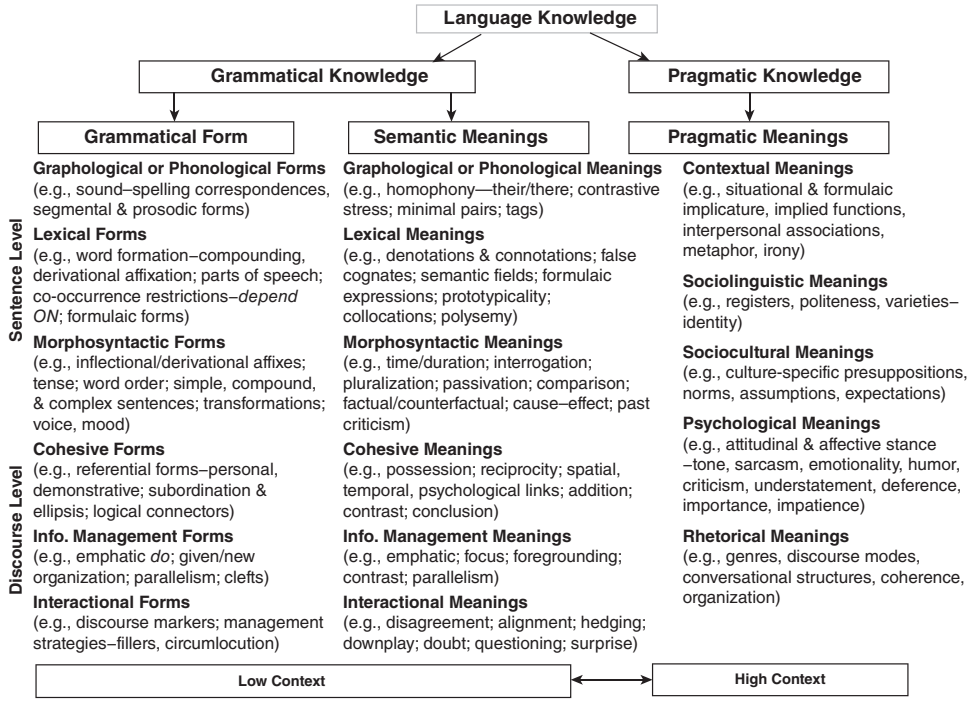

Figure 6.4 Purpura's (2012) operational model of language knowledge

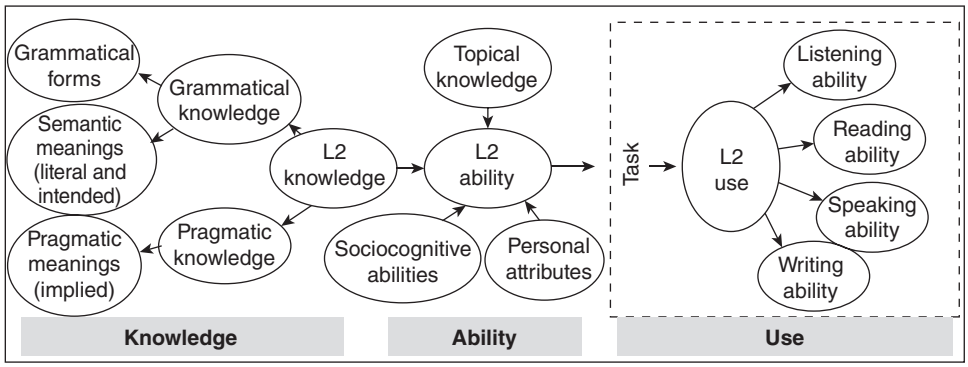

Figure 6.5 Grammatical knowledge as a resource for L2 use

Several studies (e.g., Chang, 2004; Ameriks, 2009; Grabowski, 2009; Kim, 2009; Liao, 2009; Dakin, 2010; Vafaee Basheer, & Heitner, 2012) have used Purpura's conceptualization of language knowledge to examine the nature of L2 grammatical ability in assessment contexts. Some of these studies have examined only the relationships between form and semantic meaning; most, however, have studied form–meaning resources in the context of L2 use. These studies consistently found that the learners' knowledge of grammatical form was unsurprisingly related to their knowledge of the semantic meaning, and, more generally, that knowledge

of the forms is related to the ability to use them as resources for conveying literal and intended meanings (i.e., ideas, propositions, topics), as well as nuanced pragmatic meanings in context.

For example, Vafaee et al. (2012) examined the trait structure of the grammar section of a placement test, where grammatical knowledge was defined in terms of knowledge of form and meaning. The test consisted of 19 MC form and 12 semantic meaning items, constructed around four themes. The test was administered to 144 participants representing multiple proficiency levels. The results of a confirmatory factor analysis showed that the most plausible model of the test construct consisted of two traits (form and meaning) and four methods (the test themes). Interestingly, this study not only confirmed that the form and meaning traits were separate but highly related, as one would expect, but also showed a clear, empirical relationship between grammatical knowledge (defined in terms of form–meaning mappings) and the contexts of language use.

In a much more complex study, Liao (2009) investigated the factorial structure of the grammar, reading, and listening sections of the General English Placement Test—a high stakes test used in student admissions and job screening in Taiwan. The grammar test consisted of 11 MC form and 15 semantic meaning items, and was administered to 609 participants. Liao also found two distinct but highly correlated factors: knowledge of grammatical form and semantic meaning. Furthermore, she observed that while knowledge of grammatical form and semantic meaning in the grammar test provided strong predictors of the ability to understand semantic and pragmatic meanings encoded in the reading and listening texts, knowledge of semantic meaning influenced reading and listening ability to a much greater extent than did grammatical form.

In a beginning ESL program for adult immigrants studying to be US citizens, Dakin (2009) examined the relationships between grammatical knowledge (defined in terms of form and meaning) and knowledge of civics over the course of a semester. Administering a grammar and a civics test to 98 participants before and after instruction, she found a strong relationship between the learners' grammatical knowledge and their development of civics content knowledge, noting that over time, knowledge of semantic meaning was a better predictor of civics knowledge than was grammatical form.

Finally, Grabowski (2009) investigated the nature of grammatical and pragmatic knowledge by means of a high context, reciprocal test of speaking ability designed specifically to elicit grammatical knowledge along with contextually situated pragmatic meanings. She found that knowledge of grammatical form and meaning played a consistent and significant role in interactive speaking ability across all test contexts and at all proficiency levels, whereas the examinees' knowledge of pragmatic meanings was pretty much dependent upon the situation elicited by the task. Lastly, she found that while grammatical knowledge made the most important contribution to the examinees' overall speaking proficiency scores at all levels, this contribution decreased to some extent at the advanced level. She concluded that both grammatical and pragmatic knowledge should be explicitly measured in speaking proficiency assessments at all levels of proficiency.

In sum, these studies provide compelling evidence that grammatical knowledge involves more than a single focus on form, and that the measurement of both

dimensions of form and meaning are critical to a comprehensive assessment of L2 proficiency.

Current Approaches, Challenges, and Research Related to the Measurement of Grammatical Knowledge

Despite the form–meaning research, most L2 testers continue to conceptualize grammatical knowledge uniquely in terms of form, with little or no explicit attention to the measurement of meaning. While a form-focused approach to L2 assessment is certainly appropriate for some purposes, it provides only a partial representation of the grammar construct. As a result, important opportunities for supplying learners with information that could help them develop are missed. Therefore, I believe that grammar test development should be guided by a theoretical model of grammatical knowledge if for no other reason than to contextualize the actual test construct within the larger frame, and to help ensure that important aspects of the construct are represented in the test.

In the next section, I will first discuss some general considerations in the design of grammar assessment tasks. Then, I will discuss four methodological approaches to grammar assessment.

General Considerations in the Design of Grammar Test Tasks

Once we know the test purpose and what aspects of the construct to measure, we need to consider the contexts of target language use (TLU) so that we identify tasks that examinees are likely to encounter in real-life or instructional language use (Bachman & Palmer, 1996). This pool of target-like tasks can then be used for selecting test tasks. The degree to which the tasks on language tests correspond to the tasks in the TLU domain is referred to as *test authenticity* (Bachman & Palmer, 1996). This characteristic of assessment is critical for providing a basis to generalize score-based performance from assessment tasks to performance in the TLU domain.

Therefore, in an effort to maximize authenticity, grammar test development should probably begin with a consideration of the domains (i.e., situations) in which examinees will be likely to function linguistically, so that tasks within that domain can be identified and considered for test inclusion in light of the test purpose. We would also need to think about the grammar examinees would need to use to perform these tasks.

To illustrate, imagine we were designing a placement test in a university setting. Examinees in this context typically need to perform language tasks related to the following four domains: (1) the social-interpersonal (e.g., having a conversation in a café), (2) the social-transactional (e.g., resolving a course registration problem), (3) the academic (e.g., listening to a lecture), and (4) the professional (e.g., making a conference presentation). Within and across each domain, we can think of several features that could guide and control task development to ensure that test tasks align with TLU tasks. Table 6.2 provides an example of how tasks within these

Table 6.2 Linking test tasks with TLU tasks

Target domain	Social-interpersonal	Social-transactional	Academic	Professional
Setting	Business (café)	Business (pharmacy)	School (chemistry lab)	Conference (lecture hall)
Event	Socializing	Service encounter	Lab experiment and report	Conference presentation
Participant roles	Friend/friend	Customer/pharmacist	Student/student	Presenter/audience
Topic of communication	Subway event	Instructions for medicine	Litmus test experiment	Global warming
Goal of communication	Narrate a subway story	Understand prescription instructions	Report lab results	Explain and critique a new policy
Sociolinguistic features	Informal, close friends	Formal, businesslike	Formal, academic	Formal, professional
Sociocultural characteristics	New York City	USA	University	International association
Affective tone	Friendly/polite	Confused/helpful	Supportive/inquisitive	Collegial/supportive
Test input	Written prompt, role play	Instructions, listening text (dialogue), MC questions	Written prompt	Written prompt
Expected response	Extended production	Selected response	Extended production	Extended production
Communicative focus of assessment	Literal/intended meanings: narrate sequence of events	Literal/intended meanings: understand how to take medicine, side effects, and interactions	Literal/intended meanings: report: describe actions, observed results, and conclusions	Literal/intended meanings: provide a logical argument for topic: pros, cons
Linguistic focus of assessment	Grammatical (and pragmatic) resources: use a range of grammatical forms (e.g., simple past and past continuous tenses, logical connectors) accurately and meaningfully	Grammatical (and pragmatic) resources: use a range of grammatical forms to understand	Grammatical (and pragmatic) resources: use a range of grammatical forms (e.g., active and passive voice, *when* clauses, connectors) accurately and meaningfully	Grammatical (and pragmatic) resources: use a range of grammatical forms (e.g., formal registers, hedges) accurately, meaningfully, and appropriately

four domains can be generated, specified with respect to several features, and used to create grammar assessments.

In designing grammar tasks, we also need to consider how to elicit test performance so that examinees can display their grammatical knowledge. In other words, we need to consider the types of responses examinees might be expected to make in relation to the instructions and questions on the test—i.e., the *task input*. The type of *expected response* is critical since inferences about grammatical knowledge will be based on the scores associated with these responses. Test tasks can require examinees either to *select* a response from two or more options or to *construct* a response. Selected response tasks (SR) allow us to make inferences about the learners' receptive knowledge of the learning point; constructed response (CR) tasks allow us to make inferences about the examinee's language production. In constructing responses, examinees may need to produce a *limited* amount of language (i.e., anywhere from a word to a sentence) or an *extended* amount (i.e., more than a sentence). Limited production (LP) tasks allow us to make inferences about the learners' *emergent knowledge* of the learning point, while extended production (EP) tasks allow us to make inferences about learners' full production or their overall L2 performance. (For more information on writing items and tasks and on different response formats, see Chapter 48, Writing Items and Tasks, and Chapter 52, Response Formats.) Examples of *SR*, *LP*, and *EP* or *performance* tasks are presented in Table 6.3.

Table 6.3 Ways of eliciting grammatical performance (Purpura, 2012)

SR tasks	*CR tasks*			
	LP tasks	*EP tasks*		
• noticing (circle the verbs) • matching • same/different • true/false • agree/disagree • MC • error detection • ordering • categorizing • grouping • judgment tasks	• labeling • listing • gap-filling • cloze • sentence completion • discourse completion task (DCT) • short answer • sentence reformulation	Product focused: • essay • report • project • poster • portfolio • interview • presentation • debate • recital • play	Performance focused: • role play • improvisation • interview • retelling • narration • summary • info gap • reasoning gap • opinion gap • jigsaw • problem solving • decision making • interactive DCT	Process focused: • observation with rubrics, checklists, anecdotal reports • self-reflection with journals, learning logs, think-alouds
Receptive	*Emergent*	*Full production or overall L2 performance*		

In the next section, I will describe four common approaches to grammar assessment.

The Discrete-Point Approach to Grammar Assessment

Probably the most common way of assessing grammar is to use SR tasks to isolate and measure discrete units of grammatical knowledge. The assumption underlying this approach is that learning involves the acquisition of a discrete and finite set of predictable patterns. Discrete-point tasks are capable of measuring a wide range of individual forms, are relatively practical to administer and easy to score, and can be used to provide fine-grained information on grammatical knowledge. These tasks are also notoriously difficult to construct well, even if they do not appear so. (See Chapter 52, Response Formats, for more information on writing items and tasks.)

SR tasks of grammatical knowledge present test input in the form of an item and are designed to measure recognition or recall (i.e., receptive knowledge), usually involving one area of knowledge. These tasks are traditionally scored right or wrong for accuracy, that is, dichotomous scoring. (For more information on scoring, see Chapter 51, Writing Scoring Criteria and Score Reports, and Chapter 58, Administration, Scoring, and Reporting Scores.) The following item aims to measure lexical form by means of a co-occurrence restriction between an adjective and its associated preposition:

Example 1: Grammatical form: lexical form (co-occurrence restriction)
I am interested _____ history.

a. at
*b. in
c. to
d. of (*correct response)

SR items can also be designed as "multitrak" items (Dávid, 2007), where examinees are presented with test input containing several potential choices for the context. In these items, examinees have to select the option that is *not* accurate, meaningful, appropriate, acceptable, natural, or conventional. The following multitrak item intends to measure the different meanings associated with modal auxiliaries (i.e., degrees of certainty). *Must* is the only option *not* semantically acceptable in this exchange.

Example 2: Semantic meaning: morphosyntactic meaning (degrees of certainty)
A: The evidence is still pretty unclear.
B: So then, it _____ be the butler or possibly someone else.

a. may
*b. must
c. might
d. could

SR items can also be designed to measure the overall semantic meaning of an utterance revolving around a specific form (Chang, 2004). The following item aims to measure the overall semantic meaning of an utterance containing a relative clause.

Example 3: Overall semantic meaning (relative clauses)
The woman in the corner who speaks Sicilian is my aunt.
My aunt speaks Sicilian.
*True
False

The obvious concern with discrete-point, SR tasks of grammatical knowledge is that knowledge of forms in isolation may not actually translate into the ability to use these forms meaningfully in communication; that is, these tasks fail to elicit responses capturing dynamic and complex understandings of the resources needed for communication. Nonetheless, this approach to grammar assessment is useful in situations where the goal is to observe the examinees' receptive knowledge of isolated language features.

In terms of research, several studies have examined the validity of using discrete-point, SR items as indicators of grammatical knowledge. Results from this research show that these tasks generally have high reliability and can be statistically plausible measures of grammatical knowledge. With regard to the effect of task format on the measurement of grammatical knowledge, Currie and Chiramanee (2010) examined the construct equivalence of using MC and CR tasks as measures of grammatical knowledge. They found that the MC format seems to elicit more format-related noise than the CR format, and that MC tasks do not reflect the same types of responses as those elicited in CR tasks. This study casts doubt on the validity of MC tasks as measures of grammatical form.

Finally, Purpura (2005) examined the convention of scoring MC grammar items dichotomously. He asked experienced teachers to judge the degree to which response options represented knowledge of grammatical form, meaning, or both. Teachers consistently agreed in their characterizations of how some options represented full knowledge, others represented some knowledge, and still others represented no knowledge of the targeted feature. These judgments were corroborated by student response data showing that the overall average scores of examinees selecting the different options corresponded to the expert judgments made by teachers regarding knowledge representation. Finally, when the responses were modeled in a partial credit statistical model, the number of thresholds observed for each item generally supported the results from the other two methods. Purpura concluded that there is seldom an empirical basis for scoring MC items dichotomously, and that doing so may underestimate the scores of those examinees who are still developing.

Another common way to assess grammar is by means of LP tasks designed to assess discrete units of grammatical knowledge. LP tasks present test input in the form of an item that requires examinees to produce a limited amount of language. LP tasks are based on the assumption that grammar learning transpires over time in developmental stages, represented by performance that is in variation on its

pathway to target-like proficiency. Discrete-point, LP tasks are also capable of measuring a wide range of individual forms. They are fairly easy to develop, relatively practical to administer, moderately easy to score, and can provide fine-grained, developmental information on grammatical knowledge—a major advantage over SR tasks.

The following LP item is designed to measure only one area of grammatical knowledge: morphosyntactic form of auxiliary verbs. Consequently, only one right response is possible. Scoring would be dichotomous, based on grammatical accuracy.

Example 4: Grammatical form: morphosyntactic form (auxiliary verbs)
If I (1) _____ known, I would (2) _____ done something.
Answers: (1) had; (2) have

The following LP item aims to measure more than one area of language knowledge, since the examinee needs to have knowledge of both grammatical form and lexical meaning in order to construct a correct response.

Example 5: Grammatical form and mean (future progressive)
Just think. This time next month, we _____ in the Mediterranean Sea.
Answer: will be swimming

If the examinee responds with *swimming*, this response would reflect knowledge of lexical meaning—that is, the verb "swim" for this context—but would show lack of knowledge of morphosyntactic form related to future progressives (i.e., the form dimension). Given the two dimensions, this item should probably be scored for semantic meaningfulness and grammatical accuracy. A score relating to only one dimension would underestimate the examinee's grammatical knowledge, and potentially lose important developmental information for providing corrective feedback.

The following LP item aims to measure the morphosyntactic form of relative clauses. Examinees are first asked to judge the accuracy of the target structure. If it is wrong, they are asked to correct it.

Example 6: Grammatical form and meaning: recognition/correction (relative clauses)
A: Do you have a computer I can borrow it?
Circle one: Correct? Incorrect?
Correction: _____
Answers: incorrect; a computer I can borrow

Like SR tasks, LP tasks have been used as viable indicators of grammatical knowledge. Despite their widespread use, surprisingly little research has been published on the LP format relating to grammar assessment.

The Performance-Assessment Approach to Grammar Assessment

Many L2 testers believe that the assessment of grammatical ability is best accomplished through *performance tasks*, where examinees are presented with input in

the form of a prompt and required to produce extended amounts of spoken or written data, of which the quality and quantity can vary considerably among test takers. Performance tasks, a kind of EP task, are best designed when they reflect the tasks learners might encounter in the TLU domain (for a more detailed discussion of performance assessment, see Chapter 37, Performance Assessment in the Classroom). Because of the amount of data produced by these tasks, assessment involves multiple areas of L2 knowledge depending on the assessment goal. Speaking performance tasks are thought to be good measures of the learners' implicit knowledge of grammar, given the online nature of performance (Ellis, 2001). In sum, performance tasks provide an excellent means of eliciting the ability to use grammatical resources to convey a range of meanings during task completion. However, it is often difficult to fully control the type of grammar that a performance assessment will naturally elicit.

The performance-assessment approach is characterized not only by EP tasks, but also by the process for scoring performance data. Before discussing scoring, consider the following example of a speaking performance task.

> *Example 7: L2 performance task (complaints)*
> Imagine you were just on a long-distance bus trip, and several things went wrong. When you call the bus company to complain, you are asked to leave a voice mail message. Describe what happened and express your feelings about the service. Include in your message at least three things you would like the bus company to do. You have one minute to plan your response. Be polite but firm.

Performance samples elicited from the task above are likely to provide multiple assessment opportunities. As the primary goal of this task is to communicate a meaningful complaint, we might begin by evaluating the response for *semantic meaningfulness*, that is, for a voice mail message with complete and valid information for the context. Then we might evaluate the degree to which the response displays *grammatical precision*. Precision refers to how grammatically accurate the response is (*accuracy*), how varied the forms are (i.e., *range*), how the response displays late-learned, sophisticated grammatical forms (e.g., past passive modals) and complex constructions involving coordination and subordination (i.e., *complexity*), and automatic and effortless delivery of the response (i.e., *fluency*, with a minimum of disfluencies). Beyond these features, responses might also need to display pragmatic knowledge, such as *appropriate* register (*sociolinguistic meanings*) and *appropriate* tone (*psychological meaning*s), or even *sensitivity* to the sociocultural conventions of complaining in a given culture (*sociocultural meanings*). In sum, performance assessments, if designed properly, elicit extremely rich grammatical (and pragmatic) data for assessment.

Finally, the performance-assessment approach is characterized by scoring procedures that involve human judges referring to a holistic or analytic rating scale. A *holistic* scoring rubric for the complaint task might minimally contain scaled descriptors characterizing the response's use of grammatical forms (the form dimension) to make a meaningful complaint (the meaning dimension). This would produce one overall score, perhaps on a scale from one (low performance) to five (high performance). An *analytic* scoring rubric might then contain two separate

components: one to characterize performance with respect to grammatical forms and the other with respect to the meaning dimension. This approach produces multiple scores that could be averaged or reported separately for formative feedback purposes.

Considerable research has been devoted to examining grammar performance by means of performance assessment. One early study performed by McNamara (1990) examined trained raters in the context of scoring the speaking section of the Operational English Test. Raters were asked to judge performance samples for resources of grammar and expression, fluency, intelligibility, appropriateness, comprehension, and overall task completion. The analyses showed that even though the raters had been trained to consider all components of speaking ability, they seemed to be making critical judgments about performance based on the resources of grammar and expression. McNamara concluded that the resources of grammar and expression seemed to provide the single best predictor of speaking ability in that test.

The L2 Production Features Approach to Grammar Assessment

Most SLA researchers and some testers maintain that the best way to understand what L2 resources learners have acquired is by asking them to engage in naturalistic (i.e., real-life) discussions, so that the features elicited by these discussions can be examined. However, these data are unrealistic for most assessment contexts. Therefore, a wide range of EP tasks have been successfully used to elicit production data containing many of the characteristics of naturalistic data.

In this approach, once performance is elicited, L2 knowledge can be inferred from the measurement of L2 production features thought to capture essential characteristics of speaking and writing performance, such as the percentage of error-free clauses or the length of the production. The claim underlying this approach is that if the linguistic characteristics of a learner's production are, in varying degrees, accurate, complex, fluent, meaningful, coherent, organized, conventional, and natural-sounding (to name a few), then this variability can be used to characterize and predict differences in speaking and writing proficiency. This approach differs from performance assessment in that it is concerned with characterizing performance in terms of production features in the data rather than judging specific L2 performance based on evidence in the data relating to a set of scaled descriptors.

While not necessarily framed this way, the L2 production features in these assessments revolve around the following knowledge components: (1) phonological, lexical, morphosyntactic, cohesive, and interactional forms and associated meanings (grammatical dimension); (2) propositions, topics, or idea units (semantic meaning dimension); and (3) markers of stance, coherence, and rhetorical or conversational organization (pragmatic dimension). In this section, I will describe three commonly examined features of L2 production (i.e., accuracy, complexity, and fluency) in this approach.

Wolfe-Quintero, Inagaki, and Kim (1998) defined *accuracy* as an error-free production unit (i.e., clause, t-unit). Several researchers (Skehan, 1998) have proposed measures of accuracy; some of the more common ones are the percentage of errors

per 100 words, the percentage of error-free clauses per total number of clauses, and the percentage of error-free t-units per total number of t-units.

Complexity is defined as the use of sophisticated forms (e.g., past passive modals), complex constructions (e.g., subordination), and various other late-learned production units. Ellis and Barkhuizen (2005) identified the following types of complexity depending on the feature being analyzed: (1) interactional (e.g., number of turns per speaker), (2) propositional (e.g., the density of the information unit), (3) functional (e.g., number of functions expressed), (4) grammatical (e.g., amount of subordination), and (5) lexical (e.g., number of academic words). Other complexity measures used to characterize L2 production include the total number of words uttered by a speaker per total number of speaker turns (*interactional complexity*); the frequency of major and minor propositions in a text (*propositional complexity*); the frequency of specific language functions (*functional complexity*); the number of words or clauses per t-unit (*grammatical complexity*); and the total number of different words used (type) per total number of words (token) (i.e., the type–token ratio) (*lexical complexity*).

Finally, *fluency* in oral production has been defined as the rapid production of language (Skehan, 1998) and operationalized by numerous measures. Ellis and Barkhuizen (2005) and Blake (2006) described fluency in terms of temporal variables (e.g., the number of syllables per second or minute on a task), hesitation variables (e.g., number of false starts, repetitions, reformulations, replacements, or other disfluencies), and the quantity of production (e.g., the response time or the number of syllables in a response).

While most of these measures come with serious caveats, many of the measures (or clusters of measures) have successfully predicted differences in L2 proficiency (Norris, 2006). As a result, serious research efforts are currently being devoted to understanding how these features relate to and even predict L2 oral and written proficiency, and what role these features might play in the development of automated scoring and feedback systems (Xi, 2010).

A growing body of research has been devoted to examining the grammatical features of L2 production in L2 assessments. Chapelle and Chung (2010) described how five automated scoring systems used measures of accuracy (e.g., agreement errors), complexity (e.g., average word length), fluency (e.g., essay length), topic relevance (e.g., topic-specific vocabulary usage), and diction (word length), to name a few, to examine relationships between these features and scores provided by human judges. Also, Ginther, Slobodanka, and Rui (2010) investigated how the automated scoring of 15 temporal measures of fluency (e.g., total response time, speech time) related to holistic ratings of speech quality. In the context of writing, Cumming et al. (2006) investigated the extent to which the features of L2 production for independent tasks differed from those for integrated tasks on the TOEFL writing exam. Examining lexical and syntactic complexity, grammatical accuracy, argument structure, orientations to evidence, and verbatim uses of source material, they found that in fact the discourse produced by examinees differed not only across tasks, but also across proficiency levels.

While these measures provide testers with a useful toolbox for characterizing L2 production within different assessment contexts, it remains unclear how these measures, individually or collectively, can be used to characterize what examinees

know or how the measures might be useful for characterizing performance for formative purposes.

The Developmental Approach to Grammar Assessment

Based on consistent findings in SLA that multiple structures seem to be acquired in a fixed developmental order and that the acquisition of single structures follows a fixed developmental sequence (see Ellis, 2008), some researchers (e.g., Pienemann, Johnston, & Brindley, 1988) have argued that grammatical assessments should be constructed, scored, and interpreted with developmental proficiency levels in mind. In fact, Ellis (2001) maintained that grammar test scores should be calculated to provide a measure of both target-like accuracy and acquisitional development; that is, a score linked to the different stages in the interlanguage continuum, so that information from these assessments could reflect both target-like and developmental criteria of specific grammatical forms.

Initial reactions to these intuitively appealing suggestions were strongly critical, arguing that the research relating to developmental orders and sequences was incomplete and at too early a stage to be used for assessment. Consequently, the use of development scores for anything more than research was discouraged.

Despite the caveats, Chang (2004) explored the degree to which scores on a relative clause test corresponded to scores on a developmental test designed to measure relative clause acquisition, based on Keenan and Comrie's (1977) accessibility hierarchy. The first section of his test included tasks aimed at measuring the forms, meanings, and pragmatic uses of relative clauses. The second consisted of two tasks developed to measure five types of relative clauses in the hierarchy. This section was designed to produce developmental scores. The first task in the developmental section asked examinees to indicate, on a scale of zero to five, how likely they were to use the targeted relative clauses in a dialogue. The responses were scored 1 for correct response, 0.5 for partially correct responses, and 0 for incorrect responses. The second task presented students with MC items designed to measure five types of relative clauses. Response options were based on the acquisitional characteristics of relative clauses and were scored as partial credit, similar to the scoring method Purpura (2005) used. Interestingly, Chang (2004) found that when form and meaning scores on a relative clause test were considered together, the observed order of difficulty for relative clauses strongly supported the noun phrase accessibility hierarchy, but when form alone was considered, the difficulty hierarchy was not fully supported.

More recently, Chapelle, Chung, Hegelheimer, Pendar, and Xu (2010) explored the potential of assessing productive ESL grammatical ability by targeting areas identified in SLA research, so that the items could be used on a computer-delivered and scored placement test. The test content was designed to measure structures on the morphosyntactic, syntactic, and functional levels (*forms and meanings*). The structures (rooted in SLA research) were putatively capable of predicting grammatical performance at the beginning, intermediate, and advanced levels. Examinees were presented with five LP tasks, where production ranged from a word to a sentence (as seen below), and one EP task, where they had to write a paragraph.

Example 8: Reorder jumbled word order (Level 3—subject–verb inversion with negative)
Complete the sentence using all the words given in the word list. Do NOT add more words or change the word forms.

seen, mess, a, they, have, such

Hardly ever _____
Answer: *have they seen such a mess*
(Chapelle et al., 2010, p. 455)

Responses were scored on a scale ranging from no evidence via partial evidence to evidence of knowledge of the targeted structure. The results showed that while the scores were indeed able to distinguish three proficiency levels, the LP tasks provided weak to moderate correlations with the EP task. Unfortunately, we have no information on whether the items themselves corresponded with the SLA level predictions. Finally, the scores from the entire productive grammar test produced expected moderate correlations with the TOEFL Internet-based test (iBT), suggesting that further research on the productive grammar test should be pursued.

Future Directions

Research and theory related to grammar assessment have made significant strides since the early 2000s, and this line of inquiry has become a vibrant area of scholarly endeavor and practical application. In the future, I believe that researchers will continue to explore the construct of grammatical ability and the resources that contribute to and predict the ability to convey meanings. Given the research in SLA on form–meaning connections and the recent research in L2 assessment on the role of meaning in grammatical knowledge, I believe that those interested in grammar assessment will move beyond the limitations of a uniquely syntacto-centric approach to grammar assessment, especially when the data clearly warrant the assessment of more than one dimension.

I also believe that grammar assessment, in both large-scale and classroom-based assessment contexts, will be significantly impacted by advances in information technologies. These technologies will remove many of the constraints of pencil and paper assessments by allowing for innovative test formats that use multimedia and flash technologies, multimodal assessment, interactivity in real time, and flexibility in test formats, so that examinees can be presented with discrete-point tasks or cognitively complex tasks, depending on the goal of assessment. Advances in test delivery systems will also allow us to assess a much wider array of grammar in a greater number of domains using a larger variety of tasks. Fortunately, these technologies will enable us to implement new and innovative ways of scoring that can provide stakeholders not only with summative information, but also with formative information. Learners will have information for closing grammar learning gaps. Already advances have been made to give learners immediate feedback on a number of grammatical features in writing and speaking. In the future, I see much greater efforts to provide learners with concrete feedback associated with individually tailored instruction and further grammar assessment.

I believe that researchers will continue to try to characterize grammatical ability at different proficiency levels and in different language use domains. We are still far from understanding what grammatical features constitute the ability to perform at different levels of L2 proficiency. I also think that corpus linguistics research will make contributions to this endeavor.

Finally, grammar is the fundamental linguistic resource of communicative language ability. We have seen this over and over again in the research. I believe that in the future L2 educators will recognize that there are many ways to define and measure grammatical ability, not just the traditional discrete-point approach. The bottom line is that all learners at times need feedback on their grammar performance. This feedback comes from assessment. I believe that in the future L2 educators will continue to recognize the importance of grammar assessment both in large-scale and classroom-based contexts.

References

Ameriks, Y. (2009). *Investigating validity across two test forms of the examination of proficiency in English (ECPE): A multi-group structural equation modeling approach* (Unpublished dissertation). Columbia University.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.

Bardovi-Harlig, K. (2000). Tense and aspect in second language acquisition: Form, meaning, and use. *Language Learning*, *50* (Supplement 1).

Blake, C. G. (2006). *The potential of text-based Internet chats for improving ESL oral fluency* (Unpublished doctoral dissertation). Purdue University.

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

Celce-Murcia, M., & Larsen-Freemen, D. (1999). *The grammar book: An ESL/EFL teacher's course* (2nd ed.). Boston, MA: Heinle.

Chang, J. (2004). *Examining models of second language knowledge with specific reference to relative clauses: A model-comparison approach* (Unpublished doctoral dissertation). Columbia University.

Chapelle, C., & Chung, Y.-R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, *27*(3), 301–15.

Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing*, *27*(4), 443–69.

Cumming, A., Kanto, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL* (TOEFL Monograph No. MS-30, ETS RM-05-13). Princeton, NJ: ETS.

Currie, M., & Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of language structure. *Language Testing*, *27*(4), 471–91.

Dakin, J. W. (2010). *Investigating the simultaneous growth of and relationship between grammatical knowledge and civics content knowledge of low-proficiency adult ESL learners* (Unpublished doctoral dissertation). Columbia University.

Dávid, G. (2007). Investigating the performance of alternative types of grammar items. *Language Testing*, *24*(1), 65–97.

Ellis, R. (2001). Some thoughts on testing grammar: An SLA approach. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 251–63). Cambridge, England: Cambridge University Press.

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

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford, England: Oxford University Press.

Ginther, A., Slobodanka, D., & Rui, Y. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379–99.

Grabowski, K. (2009). *Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking* (Unpublished doctoral dissertation). Columbia University.

Harris, D. P., & Palmer, L. A. (1986). *CELT Listening Form L-A, Structure Form S-A, Vocabulary Form V-A* (2nd ed.). New York, NY: McGraw-Hill.

Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*(2), 203–21.

Jaszczolt, K. M. (2002). *Semantics and pragmatics: Meaning in language and discourse*. London, England: Longman.

Keenan, E., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, *9*, 63–99.

Kim, H. J. (2009). *Investigating the effects of context and task type on second language speaking ability* (Unpublished dissertation). Teachers College, Columbia University.

Lado, R. (1961). *Language testing*. London, England: Longman.

Larsen-Freeman, D. (1991). Teaching grammar. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (pp. 279–96). Boston, MA: Heinle.

Liao, Y.-F. A. (2009). *Construct validation study of the GEPT reading and listening sections: Re-examining the models of L2 reading and listening abilities and their relations to lexico-grammatical knowledge* (Unpublished dissertation). Columbia University.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), 52–75.

Norris, J. (1996). *A validation study of the ACTFL guidelines and the German speaking test* (Unpublished MA thesis). University of Hawai'i.

Pienemann, M., Johnston, M., & Brindley, G. (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition Research*, *10*, 217–24.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge, England: Cambridge University Press.

Purpura, J. E. (2005). *Re-examining grammar assessment in multiple choice response format exams*. Paper presented at the Association of Language Testers of Europe Conference, Berlin, Germany.

Purpura, J. E. (2012). Assessment of grammar. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Malden, MA: Wiley-Blackwell.

Purpura, J. E., & Pinkley, D. (1991). *On target*. Glenview, IL: Scott Foresman.

Rea-Dickins, P. (1991). What makes grammar tests communicative? In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 112–35). New York, NY: HarperCollins.

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

Vafaee, P., Basheer, N., & Heitner, R. (2012). Application of confirmatory factor analysis in construct validity investigation: The case of the grammar sub-test of the CEP placement exam. *Iranian Journal of Language Testing*, *2*(1), 1–19.

VanPatten, B., Williams, J., Rott, S., & Overstreet, M. (2004). *Form–meaning connections in second language acquisition*. Mahwah, NJ: Erlbaum.

Wedgwood, D. (2005). *Shifting the focus: From static structures to the dynamics of interpretation*. Oxford, England: Elsevier.

Wolfe-Quintero, K., Inagaki, S., & Kim, H-Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity* (Technical report 17). Manoa, HI: University of Hawai'i Press.

Xi, X. (2010). Automated scoring and feedback systems. *Language Testing*, *27*(3), 291–300.