8 Assessing Grammatical Knowledge (with Special Reference to the Graded Grammaticality Judgment Paradigm)

Ben Ambridge

Summary

This chapter briefly summarizes some of the most widely used experimental paradigms in the domain of grammatical development (elicited production, repetition, weird word order, priming, act-out, and preferential looking and pointing tasks) before focusing in more detail on a relatively new grammaticality judgment paradigm. This new paradigm allows children to provide graded acceptability judgments for sentences (e.g., **The magician disappeared the rabbit*) and individual lexical forms of both familiar (e.g., *unlock*, **unsqueeze*) and novel verbs (e.g., *rifed* and *rofe* as the past-tense form of *rife*). The paradigm is suitable for use with young children (M = 4:6 for the youngest group tested so far) and also with older children and adults (where it can be used to assess the relative unacceptability of errors that these speakers would not usually produce). The paradigm yields unambiguous numerical data that do not require scoring, recoding, or reliability checking, and that are suitable for most commonly used statistical analyses (e.g., ANOVA, regression). It is well suited to research questions for which competing theoretical accounts make quantitative predictions regarding the relative (un)acceptability of particular forms (including, for example, the retreat from argument structure overgeneralization and the English past-tense debate).

Research Methods in Child Language: A Practical Guide, First Edition.
Edited by Erika Hoff.

Many different experimental paradigms have been used to assess children's knowledge of grammar (see especially McKercher and Jaswal, Chapter 10 this volume; Vasilyeva, Waterfall, and Gómez, Chapter 11 this volume). This chapter has two aims. The first is to briefly outline the most commonly used paradigms, along with their advantages and disadvantages, directing interested researchers to relevant articles (or other chapters in this volume). The second is to discuss in more detail grammaticality judgment paradigms that are suitable for use with children and, in particular, a new paradigm that my colleagues and I developed to obtained graded (as opposed to binary) judgments (Ambridge *et al*., 2008).

Production and Comprehension Paradigms

Experimental paradigms for assessing children's knowledge of grammar can be broadly divided into three types: production, comprehension, and judgment. *Judgment paradigms* are discussed extensively later in this chapter, and we will say no more about them here. *Production paradigms* use various techniques to "persuade" children to attempt to produce particular sentence types (or individual word forms), often in the hope of eliciting a particular error that is of theoretical interest. In *comprehension paradigms*, children are not required to produce language. Instead, children demonstrate their comprehension of a sentence that is verbally presented to them by choosing a matching picture from a selection (either explicitly by pointing or implicitly by looking).

Elicited Production

Probably the most commonly used paradigm is *elicited production*, whereby the experimenter aims to elicit an attempt at a particular structure by placing the child in a discourse scenario in which the target response is particularly appropriate. There are three contexts (not mutually exclusive) in which elicited production studies of this type are particularly useful.

The first is where a researcher wishes to investigate whether children have abstract knowledge of a particular structure. For example, there is a debate in the syntax acquisition literature as to whether young children are in possession of an abstract SUBJECT VERB OBJECT construction that can be used with any verb, or a set of verb-specific templates (e.g., *KICKER kick THING-KICKED*; see Tomasello, 2000, for a review). Akhtar and Tomasello (1997) investigated this issue by teaching children a novel verb ("This is called *chamming*") to describe a particular novel action (e.g., one character bouncing another on a rope). At test, the experimenter used toys to enact a scenario such as Ernie chamming Big Bird and asked the child, "What's happening (with Ernie/Big Bird)?" Since the verb is novel, a response such as *Ernie's chamming him* (produced by 80% of 3-year-olds, but only 20% of 2-year-olds) constitutes evidence that the child has some type of verb-general

knowledge. In addition to "live action" scenarios, children can also be asked to describe videos, animations, or still pictures (see Tomasello, 2000, and Ambridge and Lieven, 2011, for a summary of elicited production studies of this type).

A second scenario in which elicited production paradigms are particularly useful is when a researcher wishes to investigate children's acquisition of a structure that they rarely produce spontaneously, such as a complex question (e.g., *Is the boy who is smoking crazy?*) or the past-tense form of a low frequency verb (e.g., *rang*). One useful technique can be to engage children in a dialogue with a puppet or talking toy (who produces responses by means of a loudspeaker connected to a computer or mp3 player with pre-recorded responses). For example, Ambridge, Rowland, and Pine (2008) elicited attempts at complex questions (e.g., *Is the boy who is smoking crazy?*) by having children put questions to a talking dog toy who could "see" a picture illustrating the answer (hidden from view of the child). In some cases a "fill in the blank" technique is used. For example, in many past-tense studies (e.g., Marchman, 1997) children are presented with prompts such as, "Every day John likes to sing. Today he is singing. Yesterday he. …" As these examples illustrate, the elicited production paradigm is really a family of related techniques that may differ in detail, but are united in their aim to persuade children to attempt to produce a particular utterance.

Finally, elicited production paradigms are useful for investigating the effect of one particular variable, whilst holding other factors constant. For example, one study of question acquisition (Ambridge *et al*., 2006) used the talking dog procedure outlined above to investigate whether children produce fewer errors for questions with higher frequency auxiliaries (e.g., *can*) than lower frequency auxiliaries (e.g., should), whilst holding other aspects of the question constant (e.g., *What can/should Mickey eat?*).

The main advantage of elicited production studies is that the experimenter can exert a reasonable degree of control over what children are likely to say (though, of course, some children will not produce the intended utterances), and hence manipulate the variable(s) of interest. The main disadvantage is that elicited production tasks are probably the most difficult for children to complete. Hence children may fail not because they lack the required knowledge, but because they do not understand the nature of the task, or because one or more of the various task components (e.g., interpreting the scenario to be described, choosing the right words, planning the utterance) interferes with their ability to produce the correct form.

Repetition or Elicited Imitation

Repetition or *elicited imitation* tasks are useful when it is difficult to conceive of a discourse scenario that would restrict children to the particular structure of interest, or when this structure is sufficiently infrequent or complex that children will rarely produce it spontaneously in an elicited production task. For example, Kidd, Lieven, and Tomasello (2006) used a repetition task to assess children's ability to produce sentential complement clause constructions (e.g., *I hope she is making a chocolate cake*). The procedure is simply that the experimenter (or a puppet or cartoon character) produces an utterance, which the child is then asked to repeat. It may

seem that this task is trivially easy, and that even young children would make few errors. In fact, errors (such as substituting *think* for *hope* in the study of Kidd, Lieven, and Tomasello, 2006) are relatively common (Ambridge and Pine, 2006, identified a number of children who consistently repeated such simple sentences as *She is playing football* as **Her is playing football*). It seems that such errors occur because, rather than storing the incoming sentence verbatim, children encode the "message" of the sentence and then construct a "new" sentence using their own grammar (Lust, Flynn, and Foley, 1996). Even when children do not make errors, the time taken to repeat a sentence can be used as a measure of the relative familiarity of particular strings (e.g., Bannard and Matthews, 2008). The main advantage of the paradigm is the high degree of control that it affords over the precise form and wording of the target utterance. The main disadvantage is that it cannot be used with older children, who – at some stage – will be able to repeat a sentence verbatim using a pure "parroting" strategy, whether or not they could produce it spontaneously.

Weird Word Order and Syntactic Priming

Somewhere in between the elicited production and imitation paradigms lies the *weird word order* paradigm (Akhtar, 1999). The experimenter and child take turns describing video clips (or live actions performed by puppets), often using novel verbs that describe novel actions. For some verbs, the experimenter uses conventional word order (e.g., *Fox meeked Bear*). For others, she uses a weird word order not found in the language (e.g., *Fox Bear tammed*). The aim (as in elicited production studies such as that of Akhtar and Tomasello, 1997) is to investigate whether children have verb-general knowledge of word order. If so, when asked to describe a new video using the novel verb presented in a weird word order, they should correct to the word order that is conventional for their language (e.g., *Duck tammed Snake*). If, on the other hand, children learn individual constructions for each verb (e.g., *TAMMER THING-TAMMED tam*) they will use this construction to produce a weird word order sentence such as *Duck snake tammed* (in fact, the 2-year-old children studied by Akhtar, 1999, produced both types of response at similar rates, suggesting some verb-general and some verb-specific knowledge). This paradigm has also been used to investigate verb frequency effects (Matthews *et al*., 2004) and the intransitive construction (Abbot-Smith, Lieven, and Tomasello, 2001), and to compare word order acquisition crosslinguistically (Matthews *et al*., 2007). The weird word order paradigm shares with the elicited production/imitation paradigms to which it is related the advantage of a high degree of control over the target structure. A disadvantage is that children (particularly older children) may mimic word orders that they know to be incorrect, either "for fun" or because they assume that this is what is required of them (though it is usually possible to control out this confound by using real verbs to estimate rates of deliberate weird word order responses). Like all other production paradigms, it is suitable for use only with children old enough to be able to produce the relevant sentence types (see below).

As the *syntactic priming* paradigm is discussed in detail in Vasilyeva, Waterfall, and Gómez (Chapter 11 this volume), I mention it here simply to point out that the

findings of weird word order studies make the interpretation of syntactic priming studies less straightforward than is generally assumed. Syntactic priming refers to the phenomenon whereby hearing a particular construction (e.g., *The digger pushed the bricks*) increases the likelihood that the child will use the same construction (e.g., *The hammer broke the vase*) than a possible alternative (e.g., *The vase was broken by the hammer*) to describe a subsequently presented scene. Such findings are generally taken as evidence for prior knowledge of the construction (for this example, the SUBJECT VERB OBJECT transitive construction). The caveat from weird word order studies is that identical priming effects (though they are not usually described as such) are sometimes observed for constructions of which children cannot possibly have had prior knowledge (i.e., weird word order constructions). Thus care must be taken when interpreting syntactic priming as evidence for prior knowledge of a construction.

Comprehension Paradigms: Act-Out Tasks and Preferential Looking/Pointing

A problem shared by all production paradigms is that children may in principle have knowledge of a particular structure that is not sufficient to support production (which may be interrupted by the demands involved in utterance planning and formulation), but that is sufficient for comprehension. Comprehension tasks are used to investigate this possibility.

Act-out studies are primarily used to investigate children's knowledge of word order. As in the elicited production studies outlined above (e.g., Akhtar and Tomasello, 1997) children are taught a novel verb (e.g., *chamming*) to describe a novel action. Instead of describing an enactment performed by an experimenter, however, children are given a sentence and asked to enact it themselves (e.g., show me *Ernie chamming Big Bird*). As with the elicited production equivalent, the rationale is that if children can correctly enact the sentence (i.e., with Ernie as SUBJECT and Big Bird as OBJECT as opposed to vice versa), they must be in possession of some knowledge of word order that is verb general (SUBJECT VERB OBJECT). Act-out studies can also be used to investigate children's sensitivity to the different cues to SUBJECT (or AGENT) found crosslinguistically such as case marking (e.g., MacWhinney and Bates, 1989). In principle, the advantage of act-out studies is that they can be used with younger children than equivalent production studies (e.g., children who are not yet capable of producing three-word utterances with a novel verb). In practice, however, act-out tasks appear to be surprisingly demanding for young children: the study of Akhtar and Tomasello (1997) also included an act-out task, for which most children aged 2:10 showed at-chance performance.

Preferential looking/pointing paradigms (e.g., Naigles, 1990; Gertner, Fisher, and Eisengart, 2006) reduce task demands further (and hence generally show verb-general knowledge in younger children than act-out or production tasks). Children again hear a sentence such as *Ernie is chamming Big Bird* but, instead of enacting the sentence with toys, must "choose" from two video displays: one showing the scenario described, one with the roles reversed (e.g., Big Bird chamming Ernie). When a pointing task is used, children are taught to explicitly select the matching scene.

Preferential looking tasks make use of the fact that children generally spontaneously look for longer to the matching than the nonmatching image to infer comprehension.

The main advantage of the preferential looking paradigm (discussed in detail in Piotroski and Naigles, Chapter 2 this volume; and see also Golinkoff and Hirsh-Pasek, Chapter 5 this volume) is that it can be used with very young children (i.e., children who are too young to make any explicit response). Indeed, studies using the paradigm have demonstrated apparent verb-general knowledge in children aged as young as 1:9 (Gertner, Fisher, and Eisengart, 2006). The disadvantage is that, since children's looking behavior is not an unambiguous measure of their comprehension, the most appropriate interpretation of any given set of findings is not always clear, and is often controversial (see Ambridge and Lieven, 2011, Chapter 3; Chan *et al*., 2010; Dittmar *et al*., 2008). The pointing version of the paradigm produces unambiguous data, but presumably is suitable for use only with slightly older children (the youngest group studied so far had a mean age of 2:3; Noble, Rowland, and Pine, in press).

Grammaticality Judgment Paradigms

As we have already seen, there are many areas of investigation for which production and comprehension measures can be used to assess children's grammatical knowledge (indeed, for many research questions, these paradigms are more suitable than a judgment task). As we will see, however, the main advantage of the grammaticality judgment paradigm is that it allows the researcher to answer questions that cannot be directly addressed using production or comprehension measures, by investigating children's knowledge of grammar (both syntax and morphology) in a relatively explicit manner. The graded grammaticality judgment paradigm to be introduced here provides unambiguous, numerical data that do not require scoring, recoding, or checking for interrater reliability, and that are suitable for most commonly used statistical analyses (e.g., ANOVA, regression). As for many of the paradigms discussed above and elsewhere in this volume, novel items (usually verbs) can be created for use in the study, in order to test children's general syntactic or morphological knowledge independent of their knowledge of particular lexical items. The paradigm is relatively demanding, and hence is most suitable for use with relatively old children (we have not yet attempted to test children younger than 4). Generally speaking, grammaticality judgment tasks are also suitable for children with specific language impairment (e.g., Rice, Wexler, and Redmond, 1999; and see McGregor, Chapter 21 this volume) and second language learners (e.g., Mandell, 1999), though, of course, this may raise the minimum age further.

Research Aim

My own interest in developing a graded grammaticality judgment paradigm for use with children stems from my research on a topic that has become known as Baker's paradox (or the "no negative evidence" problem). Suppose that a child hears a

particular verb (e.g., *break*) in both an intransitive sentence (e.g., *The stick broke*) and a transitive causative sentence (e.g., *The man broke the stick*). Through repeated encounters with other pairs that fit this pattern (e.g., for *roll* and *open*), the child will set up some kind of generalization or "rule" that (informally speaking) generates transitive causative sentences for verbs that have appeared only in the intransitive:

Intransitive sentence		*Transitive causative sentence*
[The stick] [broke]		[The man] [broke] [the stick]
[The ball] [rolled]		[John] [rolled] [the ball]
[The door] [opened]		[Louise] [opened] [the door]
Rule: [NP1] [VERB]	→	[NP2] [VERB] [NP1]

Suppose, for example, that the child hears *The cup smashed*. The child can use this rule to generate a sentence such as *Mummy smashed the cup*, even if no sentence of this type has been encountered in the input.

How do we know that children are forming generalizations of this type? One answer is simply that they must be, otherwise language would consist of nothing more than a set of rote-learned sentences, which is clearly not the case (Chomsky, 1959). A better answer is that many experimental studies (see Tomasello, 2000, for a review) have shown that, when taught a novel verb in intransitive sentences only (e.g., *The ball is tamming*), most children aged 3:0 and older are able to use this verb in a transitive causative sentence (e.g., *The mouse is tamming the ball*). Another source of evidence comes from children's overgeneralization errors. Many researchers (most notably Bowerman, 1988) have found that children produce utterances such as **The magician disappeared the rabbit*. Such utterances cannot have been learned by rote from the input (as adults do not produce them), and hence must have come from the application of a generalization process of the type outlined above. Errors of this type are termed *argument structure overgeneralization errors*, because a verb (*disappear*) has been used in an *argument structure construction* (sentence frame) in which it is not permitted in the adult grammar (here the transitive causative), through the *over*-application of a *general* rule.

Explaining how children learn not to make these errors turns out to be a very difficult problem. It cannot be simply that children avoid using verbs in sentence constructions in which they have not appeared in the input, or they would never make such errors in the first place (or be able to produce novel utterances such as *The mouse is tamming the ball*). Whilst implicit or explicit correction by parents and caregivers is no doubt useful (e.g., Chouinard and Clark, 2003), this cannot be the whole story, as adult speakers are able to reject as ungrammatical errors that they are extremely unlikely to have produced – and subsequently had corrected – during childhood (e.g., **The clown chuckled the man*).

The goal of the research program for which my colleagues and I developed the graded grammaticality judgment paradigm was to test various proposals for how, having begun to produce overgeneralization errors such as **The magician disappeared the rabbit*, children "retreat" from these errors. For example, one proposal, Braine and Brooks's (1995) *entrenchment hypothesis*, states that repeated presentation of a verb in particular constructions (e.g., *The rabbit disappeared*)

gradually causes the child to probabilistically infer that the verb cannot be used in nonattested constructions (e.g., **The magician disappeared the rabbit*). Intuitively, the idea is that the child (not consciously of course) forms an "inference from absence" along the lines of "if *disappear* could be used in this way, surely I would have encountered it by now." The prediction from this account is that overgeneralization errors should be deemed more unacceptable for high frequency verbs than for semantically matched lower frequency verbs (e.g., **The magician disappeared/vanished the rabbit*), as this inference from absence is stronger for the former.

Choosing a Suitable Paradigm

In order to test this prediction, we need to obtain from children a measure of the relative (un)acceptability of different overgeneralization errors (and, as a control, correctly formed utterances). In fact, experimental tasks other than the grammaticality judgment paradigm do not provide a direct measure of the relative unacceptability of particular utterances.

An act-out, preferential looking/pointing comprehension task would provide information about the relative interpretability of a number of utterances, but there does not necessarily exist any correlation between interpretability and grammatical acceptability. Intuitively, it would seem that had we asked children to enact, for example, **The magician disappeared the rabbit* and **The magician vanished the rabbit*, they would have had little difficulty with either.

An elicited production task, in which the experimenter attempts to elicit each sentence from children, is more suitable (such a study was conducted by Brooks and Tomasello, 1999). Again, however, the paradigm does not provide a direct measure of grammatical acceptability. A child might produce an utterance that she considers to be ungrammatical (e.g., **He disappeared the rabbit*) if placed in a discourse scenario where such a response seems to be expected (e.g., *What did the magician do?*), particularly if she has not yet learned a suitable alternative formulation (e.g., *He made the rabbit disappear*). Conversely, the child's failure to produce a particular utterance does not constitute strong evidence that she considers it to be ungrammatical.

Consequently, any attempt to infer the relative unacceptability of two or more erroneous utterances from the relative rates at which they are produced is problematic. Suppose, for example, that a particular child produces five overgeneralization errors with *vanish* (e.g., **He vanished the rabbit*) and only two with *disappear* (e.g., **He disappeared the rabbit*). Is the correct conclusion (1) that the child deems the latter to be less acceptable or (2) that, having produced both utterances, the child considers both to be acceptable? After all, the normal assumption (assuming an idealized scenario with no pure "production errors") is that speakers' utterances reflect their grammars: if a speaker produces an utterance, she considers it to be grammatical.

It is also difficult to see how an elicited production task could be used to ask which of two alternative sentence constructions with the same verb children deem to be more grammatical. For example, if one wishes to test whether children know that

The rabbit disappeared is more acceptable than **The magician disappeared the rabbit*, one cannot simply compare the rates at which children produce each sentence in an elicited production task, as the sentences are not matched for difficulty. The second is longer and includes more participants (placing a higher load on memory) and is hence presumably more difficult for a child to produce, even if she considers it to be perfectly grammatically acceptable.

The best way to obtain a measure of the relative (un)acceptability of particular utterances is, of course, to ask children directly, using a grammaticality judgment task (though, in fairness, some of the children studied by Brooks and Tomasello, 1999, were probably too young for this to be feasible). We are by no means the first researchers to come to this conclusion. For example, Theakston (2004) investigated the entrenchment hypothesis using a *binary grammaticality judgment task*. Under this paradigm (discussed in more detail in McKercher and Jaswal, Chapter 10 this volume), children are asked simply to indicate whether or not each sentence is acceptable, as opposed to providing a graded judgment of the degree of (un)acceptability of a particular sentence. In this study, sentences containing overgeneralization errors (e.g., **I'm gonna disappear it*) were read aloud by an experimenter. The child's task was to help a toy animal decide whether each sentence was "OK" or "a bit silly" by moving the animal to a card showing a red cross or a green tick.

The advantage of a binary judgment task is that it can be performed by young children (Theakston's youngest group had a mean age of 5:9, though the task has been used with children as young as 4:1, e.g., Rice, Wexler, and Redmond, 1999). The disadvantage is that, for each child and each sentence, the task produces only a binary outcome measure (grammatical or ungrammatical). This means that to compare the judged grammaticality of two sentences (e.g., **I'm gonna disappear/vanish it*) it is possible to compare only the *number of children* who judged each sentence to be ungrammatical. One consequence of this is that it is impossible to analyze the data using parametric statistical tests (e.g., ANOVA) which can be used to look for interactions between variables, and which can be run within subjects, hence increasing the power of the analysis (maximizing the likelihood of finding any effect that is present). A more serious problem is that, beyond a certain age, it will no longer be possible to compare the *relative* ungrammaticality of two ungrammatical sentences (e.g., **I'm gonna disappear/vanish it*), as both will be classified as ungrammatical by close to 100% of children.

It is for this reason that Theakston (2004) used a *graded grammaticality judgment task* with her adult participants. In a graded grammaticality judgment task, participants are asked to judge the relative (un)acceptability of utterances using a graded scale – in this case a seven-point Likert-type scale – ranging (for example) from "completely unacceptable" to "completely acceptable" (the precise wording varies between studies). Grammaticality judgment studies with adults often use more sophisticated measurements such as a visual analog scale, which is not divided into discrete ratings (participants indicate their judgment by making a mark on a continuous line), or magnitude estimation, in which participants' ratings are not confined to a particular scale (e.g., Bard, Robertson, and Sorace, 1996). Our goal, however, was to develop a graded grammaticality judgment paradigm that could be used in exactly the same format with adults and children.

Procedure

Smiley-Face Scale

Under the graded grammaticality judgment paradigm (Ambridge *et al*., 2008), participants indicate their judgments using the five-point "smiley-face" scale shown in Plate 5 (reproduced with permission from Ambridge *et al*., 2008, p. 105).

The scale consists of five cartoon faces and has a midpoint denoted by a neutral face, two "more acceptable" levels denoted by smiling green faces, and two "less acceptable" levels denoted by frowning red faces (the neutral face is split into red and green halves). The child has two counters – one red and one green – and indicates her judgment, first, by choosing either the red or the green counter (to indicate unacceptable/acceptable) and, second, by placing her chosen counter on one of the faces to indicate the *degree* of (un)acceptability (either counter can be placed on the middle face). We have never encountered a child who placed a red counter on a green face or vice versa. The goal of this "two-step" procedure is to ensure that any children who are unable to provide a graded judgment (by using the faces scale) still provide a binary judgment (by choosing the red or green counter). However, we have not yet found an age at which children are able to use the counters but not the scale (though we have only tested children aged 4 years and older). Testing can be conducted using either (1) a booklet with one scale for each test item (in which case the experimenter ticks or circles the relevant face after the child has made her selection) or (2) a single scale which is reused for each trial (in which case the experimenter notes down each judgment on a separate sheet). Note, however, that older children and adults generally prefer to mark their choice directly on the scale, necessitating option 1.

Training (Warm-Up) Procedure

Children are introduced to the use of the scale through a carefully constructed training procedure. First the experimenter explains the nature of the game: the "talking dog" (a soft toy containing a loudspeaker connected to a laptop computer or mp3 player) is "learning to speak English but, because he's only a dog, sometimes gets it wrong and says things a bit silly." The child's task is to help him by letting him know whether he "said it right" or "a bit silly." The use of a talking toy is designed to overcome any reluctance a child may have with regard to "correcting" an adult, and also to make the task more enjoyable for children. (Although most enjoy hearing the dog speak, very occasionally we encounter children who are too frightened to continue; and according to Core, Chapter 6 this volume, the talking dog is frightening to most 2-year-olds).

The experimenter then provides (via the dog) an example of a maximally acceptable sentence (e.g., *The cat drank the milk*) and places the green counter on the happiest face, explaining "when he gets it right, we're going to choose the green counter and

put it here." Next, the experimenter provides an example of a maximally unacceptable sentence (e.g., **The dog the ball played with*) and places the red counter on the saddest face, explaining "when he says it wrong, we're going to choose the red counter and put it here. Don't worry about these other faces [*indicates the middle three faces*] for now." The child then completes two practice trials designed to provide further examples of maximally acceptable and unacceptable sentences (e.g., *The frog caught the fly*; **His teeth man the brushed*).

Taking the green counter, the experimenter then explains that "Sometimes he [indicates dog] says it right but it's not perfect. If it's good but not perfect, you can put the counter here [indicates second happiest face]. If it's a little bit right and a little bit wrong, or somewhere in between, you can put it here [indicates middle face]." Taking the red counter, the experimenter continues, "Sometimes he says it wrong but it's not really terrible. If it's wrong but not terrible, you can put the counter here [indicates second saddest face]. If it's a little bit wrong and a little bit right, or somewhere in between, you can put it here [indicates middle face]." The child then completes three further training trials designed to illustrate intermediate degrees of (un)grammaticality.

The sentences for these training trials need to be carefully chosen for the relevant study to ensure – on the one hand – that they exemplify the general type of error that will be judged in the main part of the study (e.g., argument structure overgeneralization errors as opposed to past-tense *-ed* overgeneralization errors) and – on the other – that they are not of exactly the same specific type (e.g., transitive causative overgeneralizations of intransitive verbs), to avoid providing hints that could affect responses in the main part of the study. For our study of transitive causative overgeneralization errors, the three intermediate training items involved overgeneralizations of prepositional-dative-only verbs into the double-object dative construction: **The woman said the man a funny story* (intended rating 2/5), **The girl telephoned her friend the news* (3/5 or 4/5) and **The man whispered his friend the joke* (4/5). By way of comparison, a study of the acceptability of various past-tense forms (Ambridge, 2010) used incorrect regular and irregular noun plurals as training items. Children's ratings are generally broadly in line with these target ratings but, if not, the experimenter can re-explain the procedure and give feedback. The child then moves on to the main part of the study, which proceeds in the same way (though with trials presented in random order).

Animations

For all training and test trials, a cartoon animation depicting the event being described by the dog is shown on a laptop screen, which both the child and the dog are "watching." This ensures that the truth value of the dog's description is never in doubt, and that the child is judging the sentence purely on the basis of grammatical acceptability. This also guards against misinterpretation of the sentences (for example, some of Theakston's, 2004, adult participants seemed to interpret the sentences **Don't laugh/giggle me* as *Don't laugh/giggle at me* rather than, as intended *Don't make me laugh/giggle*).

Control Sentences

Another important feature of the design is that, for every ungrammatical sentence (e.g., **The magician disappeared the rabbit*), a grammatical control sentence (e.g., *The rabbit disappeared* or *The magician made the rabbit disappear*) is included. This allows the researcher to control statistically for any general (dis)preferences that may exist for particular items by calculating *preference-for-grammatical-use* (or *difference*) scores (discussed in more detail below).

It is also prudent to avoid a scenario where every utterance of a particular type (e.g., transitive causative) is ungrammatical, whilst every utterance of another type (e.g., intransitive) is ungrammatical, to guard against the possibility of children developing a task-dependent strategy such as rating all transitive causative sentences as ungrammatical. Whilst this precaution was not followed in the study of Ambridge *et al*. (2008), subsequent studies that have included this control have yielded a similar pattern of findings (Ambridge *et al*., submitted a; submitted b; Ambridge, 2010).

Because the task is relatively demanding and time consuming (young children are reluctant to complete more than about 40 trials, even if this is split over several sessions) we do not generally include any "filler" trials (i.e., trials where children rate unrelated sentence types). However, if particular study designs have trials "to spare," the inclusion of filler trials can only be beneficial.

Another difficult issue relates to the number of items per "cell" of the design. If a complex design with several variables is used, it may be difficult to include more than one or two trials per cell, whilst keeping the overall number of trials manageably low. For example, Ambridge *et al*. (2008) included only one transitive causative sentence with each verb (e.g., **The magician disappeared the rabbit*), whereas ideally one would take an average rating across several (e.g., **The witch disappeared the frog*, **The conjurer disappeared the card*, etc.). An approach followed in subsequent studies (e.g., Ambridge, 2010) is to have two (or more) versions of "the same" experiment with different items (e.g., half of the children would rate **The magician disappeared the rabbit* and half **The witch disappeared the frog*). This allows the number of items per cell to be doubled (or trebled, quadrupled, etc.) without increasing the time taken for an individual child to complete the study.

Data

As previously mentioned, an advantage of the graded grammaticality judgment paradigm is that it yields numerical data that can be analyzed using techniques such as ANOVA or regression: specifically a rating between 1 and 5 for each item (e.g., sentence) from each participant (where 5 represents the happiest face, i.e., the most acceptable). Technically, one might object that the data are not true interval-scale data (a requirement of parametric tests such as ANOVA) as we have no way of knowing whether an increase from (say) 2/5 to 3/5 on the scale represents the same increase in perceived grammaticality as (say) an increase from 4/5 to 5/5. However, the treatment of rating-scale data as interval data is so commonplace in psychology

Table 8.1 Some examples of children's judgments of grammatical and ungrammatical sentences on the five-point smiley-face scale (*5* = happiest face = most acceptable)

	4–5 (N = 20)		*5–6 (N = 27)*		*6–7 (N = 24)*		*Adults (N = 42)*	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Intransitive: *Bart disappeared*	3.15	0.39	4.63	0.14	4.92	0.06	5.00	0.00
Transitive: **The magician disappeared Bart*	2.25	0.31	3.26	0.26	2.92	0.23	2.60	0.14
Difference (intransitive minus transitive)	0.90	0.55	1.37	0.26	2.00	0.24	2.41	0.14
Intransitive: *Bart vanished*	4.25	0.23	4.70	0.12	4.92	0.06	4.95	0.03
Transitive: **The magician vanished Bart*	3.45	0.30	4.19	0.24	3.78	0.23	3.10	0.15
Difference (intransitive minus transitive)	0.80	0.34	0.52	0.25	1.13	0.26	1.86	0.15
Intransitive: *Bart blicked*	4.05	0.23	3.48	0.27	4.75	0.11	4.31	0.21
Transitive: **The magician blicked Bart*	3.70	0.34	3.48	0.30	4.00	0.22	3.67	0.18
Difference (intransitive minus transitive)	0.35	0.32	0.00	0.33	0.75	0.25	0.64	0.22

that, in practice, one will rarely encounter such an objection outside statistics textbooks (and, in many cases, a alternative nonparametric test is available). It is important to bear in mind, however, that the *absolute* values are almost certainly not particularly meaningful. Participants tend to rate the acceptability of one item with reference to another, meaning that the same sentence could receive very different absolute mean ratings in two studies with different items. The more meaningful comparison is between different items in the same study.

As an example of the type of data that the graded grammaticality judgment paradigm yields, Table 8.1 shows the mean scores for **The magician disappeared/vanished/blicked Bart* (where *blick* denotes a novel type of disappearing action) and the control sentences *Bart disappeared/vanished/blicked* (for novel verbs, the claim is that children should be able to use the semantics of these verbs to determine the constructions in which they can and cannot appear; see Pinker, 1989). Note that this table shows both the raw scores and, for each grammatical/ungrammatical pair, the difference (preference-for-grammatical-use) score, calculated by subtracting the rating for the ungrammatical sentence from the rating for the grammatical sentence (on a pair-by-pair and child-by-child basis). Data for the three older groups are taken from Ambridge

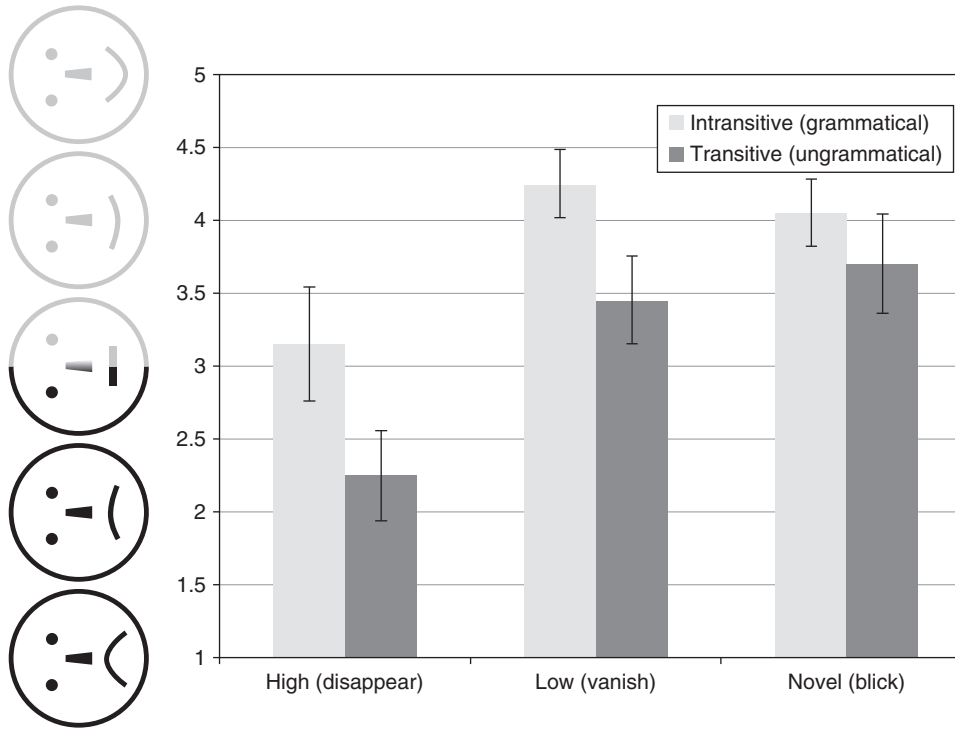

Figure 8.1 Four-year-olds' ratings for grammatical intransitive sentences (light bars) and ungrammatical transitive sentences (dark bars) for (from left to right) a high frequency, a low frequency, and a novel verb (*disappear/vanish/blick*). Error bars show standard error.

et al. (2008), and those for the younger group from a recent pilot study with 20 children aged 4:1–5:0 (M = 4:6). As an example of how data collected using this paradigm can be presented graphically, the scores for the youngest group only are also shown in Figure 8.1.

The data from the three older groups are analyzed in Ambridge *et al.* (2008), and hence will not be discussed in detail here. It will suffice to note that children aged 5–6 are clearly capable of completing the task, and give a pattern of judgments very similar to that shown by older children and adults.

For the younger children, there are two points to note. First, for the English verbs *vanish* and (marginally) *disappear*, 4–5-year-olds rated grammatical intransitive uses as significantly more acceptable than ungrammatical transitive causative uses (*vanish*, $t_{19} = 2.37$, $p = 0.014$; *disappear*, $t_{19} = 1.63$, $p = 0.058$, one-tailed test; for means see Table 8.1 and Figure 8.1). This finding is important, as it demonstrates, for the first time, that children aged 4–5 are able to use the scale to rate sentences appropriately (though the high standard error scores reflect considerable variation in this ability). Like the 5–6-year-olds, the youngest group do not appear to be able to use the semantics of the novel *disappearing* verb (or a novel *laughing* or *falling* verb, data for which are not shown) to determine the constructions in which it may and may not appear (though 5–6-year-olds can do so for a novel *laughing* verb).

Whether this is because the youngest children have yet to acquire the relevant semantics–syntax links or because the introduction of novel verbs makes the judgment task too difficult is unclear at this stage.

The second point relates to the importance of analyzing difference (preference-for-grammatical-use) scores in addition to raw scores. The entrenchment hypothesis predicts that ungrammatical transitive sentences should be rated as more acceptable for the low frequency verb (e.g., *vanish*) than for the high frequency verb (e.g., *disappear*). Looking again at the youngest group, if one compares the *raw* ratings for **The magician vanished Bart* ($M = 3.45$, $SE = 0.30$) and **The magician disappeared Bart* ($M = 2.25$, $SE = 0.31$), this prediction appears to be supported ($t_{19} = 2.60$, $p = 0.018$). However, this is misleading, because this difference is presumably a consequence – at least in part – of the fact that (for whatever reason) these children give higher ratings to sentences containing *vanish* than *disappear*, even when they are grammatical (*Bart vanished*, $M = 4.25$, $SE = 0.23$; vs *Bart disappeared*, $M = 3.15$, $SE = 0.39$). When one controls for this baseline preference by comparing *difference* scores, as opposed to raw scores, the preference for grammatical over ungrammatical uses (i.e., the dispreference for ungrammatical uses) is no longer significantly smaller for *vanish* ($M = 0.80$, $SE = 0.34$) than *disappear* ($M = 0.90$, $SE = 0.55$; $t_{19} = 0.15$, $p = 0.88$, n.s.).

Further Applications

Although the graded grammaticality judgment paradigm was initially developed to obtain ratings of verb argument structure overgeneralization errors (Ambridge *et al.*, 2008; 2009b; submitted a; submitted b), in subsequent work we have obtained judgments of past-tense forms of novel verbs (e.g., *rife* → *rifed*; *rife* → *rofe*; see Ambridge, 2010) and of grammatical and ungrammatical *un-* prefixed forms (e.g., *unlock*, *unwrap*; **unsqueeze*, **unfill*; see Ambridge *et al.*, 2009a; Ambridge, submitted). Beyond grammaticality, the smiley-face scale could also potentially be used to obtain judgments of familiarity (e.g., Ibbotson *et al.*, submitted), truth value, semantic plausibility, and so forth.

Conclusion

We end by summarizing the advantages and disadvantages of the graded grammaticality judgment paradigm introduced in this chapter. The primary advantage is that the paradigm can be used to address questions on which comprehension and production data bear only indirectly. For any domain in which the predictions of the competing theoretical accounts relate to the relative (un)acceptability of particular forms, a judgment task is – all other things being equal – more appropriate than a comprehension or production task. A related advantage is that the paradigm can be used with older speakers and adults to

obtain ratings of the relative unacceptability of errors that these speakers would not produce themselves. For example, whilst adult speakers rate **The magician disappeared Bart* as less acceptable than **The magician vanished Bart*, it would presumably be impossible to tap into the knowledge that underlies these judgments using a production task, as adults would likely produce neither. Another advantage of this paradigm over many comprehension and production measures is that it produces an unambiguous response that does not require interpretation, coding, or reliability checking. The paradigm yields numerical data that can be analyzed directly using common statistical techniques such as ANOVA and regression. An advantage that the paradigm shares with most of the comprehension and production techniques discussed in this volume is that novel verbs (or nouns, etc.) can be used in order to test whether children are in possession of item-general knowledge (as opposed to lexically specific knowledge). Finally, the paradigm can be used to obtain acceptability judgments both for whole sentences and for individual lexical items, and the "smiley-face" procedure can potentially be extended into domains where graded judgments of factors other than grammatical acceptability are required.

One disadvantage of the paradigm is that it is presumably unsuitable for use on children much younger than 4. Although we have not attempted to test children younger than 4:6 (mean age), the considerable variation in performance observed at this age (which would be considered relatively old for many domains of acquisition) means that the paradigm is unlikely to work well for younger children. That said, it may well be that younger children are able to complete a binary version of the task. Clearly this is a question that requires future research. Another concern is that, compared to many comprehension or production tasks (and particularly naturalistic data collection), the paradigm is relatively artificial, in that children are being asked to do something that is far removed from their everyday experience and use of language. There is little that can be done to address this concern, except to seek to corroborate findings from judgment tasks using comprehension, production, and naturalistic data studies, where this would be appropriate.

Finally, it is important to note that there are many research questions for which a judgment task would be either altogether inappropriate, or considerably less appropriate than a comprehension or production task. For example, when the question relates to the age at which children have abstract item-general knowledge of a particular structure (e.g., the active SVO transitive), an elicited production (e.g., Akhtar and Tomasello, 1997), repetition (e.g., Kidd, Lieven, and Tomasello, 2006), weird word order (e.g., Akhtar, 1999), priming (e.g., Savage *et al.*, 2003), act-out (e.g., Akhtar and Tomasello, 1997), preferential looking (e.g., Gertner, Fisher, and Eisengart, 2006), or pointing (e.g., Rowland and Noble, 2011) task is more appropriate. Indeed, many of our own studies use an elicited production (e.g., Ambridge *et al.*, 2006; Ambridge, Rowland, and Pine, 2008; Ambridge and Rowland, 2009) or repetition paradigm (e.g., Ambridge and Pine, 2006) for precisely this reason (though always with the "talking dog," as an additional incentive for children to respond). However, for questions where the competing theories make predictions regarding the relative unacceptability of particular forms (as opposed to error rates,

rates of correct production, etc.), some kind of judgment paradigm is clearly the most appropriate. We hope that the paradigm outlined here will therefore inspire future research into such questions.

Acknowledgments

Thanks are due to Julian Pine and Caroline Rowland for their help both with this chapter and with the development of the judgment paradigm described herein. Thanks to Glen Goodliffe-Davies for collecting the pilot study data used in Table 8.1. This research was supported by grants RES-062-23-0931, RES-000-22-1540 from the Economic and Social Research Council.

Key Terms

Binary grammaticality judgment paradigm A grammaticality/acceptability judgment paradigm in which participants are asked to indicate simply whether a particular form is acceptable or unacceptable (see McKercher and Jaswal, Chapter 10 this volume).

Comprehension paradigm Any paradigm in which children are required not to produce language, but to demonstrate their comprehension (understanding) of a utterance produced by another speaker. Children can demonstrate comprehension via the ability to enact a sentence using toys (*act-out* task), or to "choose" a picture that matches the sentence, either implicitly by looking for longer at the target than a distracter (*preferential looking*) or explicitly by *pointing*.

Difference score A score calculated by subtracting the acceptability rating for one form (e.g., **The magician disappeared Bart*) from the acceptability rating for a related form (e.g., *Bart disappeared*), in order to control for any baseline preference that may exist, regardless of grammaticality (for this example, the extent to which participants "like" sentences that contain the noun *Bart* and the verb form *disappeared*). If the difference score is calculated by subtracting the rating for an ungrammatical form from the rating for a grammatical form (as in the above example), it may also be referred to as a *preference-for-grammatical-use score*. In some cases, it may be more appropriate to calculate the difference score by consistently subtracting the rating for one particular sentence type (e.g., irregular past-tense form) from the rating for another sentence type (e.g., regular past-tense form), regardless of which form is predicted to be more acceptable (e.g., rating for *rifed* minus rating for *rofe*).

Graded grammaticality judgment paradigm A grammaticality/acceptability judgment paradigm in which participants are asked to indicate *the extent to which* a particular form is acceptable or unacceptable, using some kind of linear (graded) scale (e.g., Likert scale, visual analog scale, or, as in the studies discussed here, smiley-face scale).

Grammaticality judgment, acceptability judgment A rating (either *binary* or *graded*) of the acceptability of a particular form. Although the terms have, on the whole, been used interchangeably here, the second, more general term is probably more appropriate when an individual word form (e.g., **Unsqueeze*, *rifed*, *rofe*) as opposed to a sentence (e.g., **The magician disappeared Bart*) is being judged. This is because, for individual word forms, it is debatable whether it is *grammatical* acceptability (as opposed to morphological or phonological acceptability) that is being rated. Whatever the domain, our written instructions to adult participants usually do not mention "grammaticality," in order to avoid participants basing their ratings on prescriptive rules.

Judgment paradigm Any paradigm in which children rate the acceptability of a sentence or an individual word form (a *grammaticality/acceptability judgment* task), the truth value of an utterance (a truth value or *yes/no* judgment task), their confidence that a form has been previously encountered, etc.

Production paradigm Any paradigm in which children are required to produce language. Commonly used production paradigms include *elicited production* (where the child describes or asks questions about a scene, often to a puppet or toy), *repetition* (where the child repeats an utterance produced by an experimenter, puppet, or toy), and *priming* (where the child and experimenter take turns to describe scenes, with the experimenter sometimes using a *weird word order* for some verbs).

Smiley-face scale A five-point pictorial scale that can be used by children to give graded judgments of grammatical acceptability (or sentence familiarity, etc.) (see Figure 8.1).

References

Abbot-Smith, K., Lieven, E., and Tomasello, M. (2001) What preschool children do and do not do with ungrammatical word orders. *Cognitive Development*, 16 (2), 679–692.

Akhtar, N. (1999) Acquiring basic word order: evidence for data-driven learning of syntactic structure. *Journal of Child Language*, 26, 339–356.

Akhtar, N., and Tomasello, M. (1997) Young children's productivity with word order and verb morphology. *Developmental Psychology*, 33 (6), 952–965.

Ambridge, B. (2010) Children's judgments of regular and irregular novel past tense forms: new data on dual- versus single-route debate. *Developmental Psychology*, 46 (6), 1497–1504.

Ambridge, B. (submitted) Testing a probabilistic semantic account of the formation and restriction of linguistic generalizations: a grammaticality judgment study.

Ambridge, B., Freudenthal, D., Pine, J.M., *et al.* (2009a) Un-learning un-prefixation errors. Paper presented at the International Conference on Cognitive Modelling 2009, Manchester, UK.

Ambridge, B., and Lieven, E.V.M. (2011) *Child language acquisition: contrasting theoretical approaches*. Cambridge: Cambridge University Press.

Ambridge, B., and Pine, J.M. (2006) Testing the agreement/tense omission model using an elicited imitation paradigm. *Journal of Child Language*, 33 (4), 879–898.

Ambridge, B., Pine, J.M., Rowland, C.F., and Clark, V. (submitted a) The retreat from argument-structure overgeneralization errors: verb semantics, entrenchment or both? *Cognitive Linguistics*.

Ambridge, B., Pine, J.M., Rowland, C.F., and Clark, V. (submitted b) Restricting dative argument-structure overgeneralizations: a grammaticality-judgment study with adults and children. *Language*.

Ambridge, B., Pine, J.M., Rowland, C.F., and Young, C.R. (2008) The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition*, 106 (1), 87–129.

Ambridge, B., Pine, J.M., Rowland, C.F., *et al.* (2009b) A semantics-based approach to the "no negative evidence" problem. *Cognitive Science*, 33 (7), 1301–1316.

Ambridge, B., and Rowland, C.F. (2009) Predicting children's errors with negative questions: testing a schema-combination account. *Cognitive Linguistics*, 20 (2), 225–266.

Ambridge, B., Rowland, C.F., and Pine, J.M. (2008) Is structure dependence an innate constraint? New experimental evidence from children's complex question production. *Cognitive Science*, 32 (1): 222–255.

Ambridge, B., Rowland, C.F., Theakston, A.L., and Tomasello, M. (2006) Comparing different accounts of inversion errors in children's non-subject wh-questions: "What experimental data can tell us?" *Journal of Child Language*, 33 (3), 519–557.

Bannard, C., and Matthews, D. (2008) Stored word sequences in language learning: the effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, 19 (3), 241–248.

Bard, E.G., Robertson, D., and Sorace, A. (1996) Magnitude estimation of linguistic acceptability. *Language*, 72 (1), 32–68.

Bowerman, M. (1988) The "no negative evidence" problem: how do children avoid constructing an overly general grammar? In J.A. Hawkins (ed.), *Explaining language universals* (pp. 73–101). Oxford: Blackwell.

Braine, M.D.S., and Brooks, P.J. (1995) Verb argument structure and the problem of avoiding an overgeneral grammar. In M. Tomasello and W.E. Merriman (eds), *Beyond names for things: young children's acquisition of verbs* (pp. 352–376). Hillsdale, NJ: Erlbaum.

Brooks, P.J., and Tomasello, M. (1999) How children constrain their argument structure constructions. *Language*, 75 (4), 720–738.

Chan, A., Meints, K., Lieven, E.V.M., and Tomasello, M. (2010) Young children's comprehension of English word order in act-out and intermodal preferential looking tasks. *Cognitive Development*, 25, 30–45.

Chomsky, N. (1959) A review of B.F. Skinner's *Verbal Behavior. Language*, 35 (1), 26–58.

Chouinard, M.M., and Clark, E.V. (2003) Adult reformulations of child errors as negative evidence. *Journal of Child Language*, 30 (3), 637–669.

Dittmar, M., Abbot-Smith, K., Lieven, E., and Tomasello, M. (2008) Young German children's early syntactic competence: a preferential looking study. *Developmental Science*, 11 (4), 575–582.

Gertner, Y., Fisher, C., and Eisengart, J. (2006) Learning words and rules: abstract knowledge of word order in early sentence comprehension. *Psychological Science*, 17 (8), 684–691.

Ibbotson, P., Theakston, A., Lieven, E.V.M, and Tomasello, M. (submitted) Prototypical semantics of the transitive construction: developmental comparisons.

Kidd, E., Lieven, E., and Tomasello, M. (2006) Examining the role of lexical frequency in the acquisition and processing of sentential complements. *Cognitive Development*, 21 (2), 93–107.

Lust, B., Flynn, S., and Foley, C. (1996) What children know about what they say: elicited imitation as a research method for assessing children's syntax. In D. McDaniel, C. McKee, and H. Cairns (eds), *Methods for assessing children's syntax*. Cambridge, MA: MIT Press.

MacWhinney, B., and Bates, E. (eds) (1989) *The cross-linguistic study of sentence processing*. New York: Cambridge University Press.

Mandell, P.B. (1999) On the reliability of grammaticality judgement tests in second language acquisition research. *Second Language Research*, 15 (1), 73–99.

Marchman, V.A. (1997) Children's productivity in the English past tense: the role of frequency, phonology and neighborhood structure. *Cognitive Science*, 21 (3), 283–304.

Matthews, D., Lieven, E., Theakston, A.L., and Tomasello, M. (2004) The role of frequency in the acquisition of English word order. *Cognitive Development*, 20, 121–136.

Matthews, D., Lieven, E., Theakston, A., and Tomasello, M. (2007) French children's use and correction of weird word orders: a constructivist account. *Journal of Child Language*, 34 (2), 381–409.

McDaniel, D., and Cairns, H. (1996) Eliciting judgments of grammaticality and reference. In D. McDaniel, C. McKee, and H. Cairns (eds), *Methods for assessing children's syntax*. Cambridge, MA: MIT Press.

Naigles, L. (1990) Children use syntax to learn verb meanings. *Journal of Child Language*, 17 (2), 357–374.

Noble, C.H., Rowland, C.F., and Pine, J.M. (in press) Comprehension of argument structure and semantic roles: evidence from infants and the forced-choice pointing paradigm. *Cognitive Science*.

Pinker, S. (1989) *Learnability and cognition: the acquisition of argument structure*. Cambridge, MA: MIT Press.

Rice, M.L., Wexler, K., and Redmond, S.M. (1999) Grammaticality judgments of an extended optional infinitive grammar: evidence from English-speaking children with specific language impairment. *Journal of Speech, Language and Hearing Research*, 42, 943–961.

Rowland, C.F., and Noble, C.H. (2011) The role of syntactic structure in children's sentence comprehension: evidence from the dative. *Language Learning and Development*, 7 (1), 55–75.

Savage, C., Lieven, E., Theakston, A., and Tomasello, M. (2003) Testing the abstractness of children's linguistic representations: lexical and structural priming of syntactic constructions in young children. *Developmental Science*, 6 (5), 557–567.

Theakston, A.L. (2004) The role of entrenchment in children's and adults' performance on grammaticality judgement tasks. *Cognitive Development*, 19 (1), 15–34.

Tomasello, M. (2000) Do young children have adult syntactic competence? *Cognition*, 74 (3), 209–253.

Further Reading and Resources

Because so little research has been conducted using this new paradigm, there is very little further reading to recommend. The paper that sets out the paradigm in detail (Ambridge *et al.*, 2008) is available from my website (http://pcwww.liv.ac.uk/~ambridge/). Theakston (2004) is a good example of a study that uses a binary judgment paradigm, whilst McDaniel and Cairns (1996) provide an interesting discussion of methodological considerations in child judgment studies. A comprehensive discussion of studies that have investigated children's grammatical knowledge using elicited production, repetition, weird word order, priming, act-out, and preferential looking and pointing tasks can be found in Ambridge and Lieven (in press, Chapters 5–7).

The smiley-face scale is reproduced here as Plate 5. We have reproduced the scale and cut-out counters in color, with the intention that readers can photocopy the scale for use in their own studies.

For the studies discussed here, animations were produced using either Adobe Flash Professional (http://www.adobe.com/uk/products/flash/), an educational version of which is available at a large discount, or (in most cases) Anime Studio (http://anime.smithmicro.com/). Sound files were recorded using the freeware Audacity program (http://audacity.sourceforge.net/). Animations created using these programs (with or without embedded sound files) can be played in most internet browsers and media software including VLC (http://www.videolan.org/vlc/), QuickTime (http://www.apple.com/quicktime/download/), and (for Flash animations) SwfMax (http://www.swfmax.com/).