Advanced Review

Experimental methods in studying child language acquisition

Ben Ambridge* and Caroline F. Rowland

This article reviews the some of the most widely used methods used for studying children's language acquisition including (1) spontaneous/naturalistic, diary, parental report data, (2) production methods (elicited production, repetition/elicited imitation, syntactic priming/weird word order), (3) comprehension methods (act-out, pointing, intermodal preferential looking, looking while listening, conditioned head turn preference procedure, functional neuroimaging) and (4) judgment methods (grammaticality/acceptability judgments, yes-no/truth-value judgments). The review outlines the types of studies and age-groups to which each method is most suited, as well as the advantage and disadvantages of each. We conclude by summarising the particular methodological considerations that apply to each paradigm and to experimental design more generally. These include (1) choosing an age-appropriate task that makes communicative sense (2) motivating children to co-operate, (3) choosing a between-/within-subjects design, (4) the use of novel items (e.g., novel verbs), (5) fillers, (6) blocked, counterbalanced and random presentation, (7) the appropriate number of trials and participants, (8) drop-out rates (9) the importance of control conditions, (10) choosing a sensitive dependent measure (11) classification of responses, and (12) using an appropriate statistical test.

How to cite this article:
WIREs Cogn Sci 2013, 4:149–168. doi: 10.1002/wcs.1215

INTRODUCTION

Historically, much influential work in the domain of child language acquisition has consisted of what one might call paper-and-pencil theorizing. Indeed, it has been explicitly argued that linguistic performance (i.e., what children say and do) bears such an indirect relationship to linguistic competence (i.e., what children know) that the former is essentially useless as a diagnostic of the latter.[1] Currently, however, a consensus seems to be developing that all accounts of language acquisition phenomena, from whatever theoretical perspective, must be judged primarily on their ability to explain actual data collected from children and adults. Consequently it is useful for researchers to be familiar with as wide a range as possible of the experimental methods used in studying language acquisition, in order to enable them to select those that are most useful for studying particular phenomena.

While few modern researchers would argue that children's linguistic performance is largely irrelevant to their competence, it is true that it is impossible to investigate children's knowledge directly, and that this must be somehow inferred from observable behavior. To be sure, performance on any experimental task will be affected by both specific task demands (e.g., the ability to manipulate toys or to point to a particular video screen) and by general limitations such as attention and memory. The appropriate response to this situation is not to abandon experimental studies altogether but (1) to seek converging evidence from multiple paradigms and (2) to build into study-designs controls that allow particular task demands to be factored out. Suppose, for example, a study finds that children of a particular age are unable to produce a particular type of passive sentence. In order to rule out the possibility that children of this age are simply unable to produce sentences of this length in general (for example due to memory limitations), a suitable

Conflict of interest: The authors declare no conflict of interest.

*Correspondence to: Ben.Ambridge@Liverpool.ac.uk
University of Liverpool, Institute of Psychology, Health & Society, University of Liverpool, Liverpool, UK

control would be to include a condition where children attempt to produce an active sentence of the same length.

The aim of the present review is to provide a comprehensive outline of the major methods used in acquisition research, to evaluate their strengths and weakness, and to highlight a number of current methodological controversies, particularly in the areas of study design and statistical analysis. The methods are grouped into *naturalistic*, *production, comprehension*, and *judgment* paradigms, although many studies combine elements of each. For example, researchers might use a naturalistic corpus analysis to select items for a production study[2,3] or a parental-report measure to select participants for a comprehension study.[4,5] Although the example studies given are mostly from the domain of syntax and morphology, the majority of studies in many other domains of language acquisition research (e.g., word-learning, phonology, pragmatics) use one or other of the paradigms discussed here (broadly defined). We have not included a separate section on standardized tests such as the Clinical Evaluation of Language Fundamentals[6] the Test of Early Language Development[7] or the Peabody Picture Vocabulary test[8] because all use some variant of the methods outlined here (e.g., elicited production, comprehension, etc.). However, it is worth bearing in mind that they may be useful for selecting children to participate in a study, or for matching groups in between-subjects designs.

SPONTANEOUS NATURALISTIC SPEECH, DIARY & PARENTAL REPORT DATA

Perhaps the simplest method for studying language development is simply to record children's spontaneous speech, in conversation with caregivers. Indeed, the analysis of spontaneous speech data is often contrasted with 'experimental' methods. The paradigm is included here because the collection and analysis of such data requires many of the same decisions regarding design and controls as other paradigms. Indeed, asking caregivers to spend long periods devoted solely to conversation with their children is far from 'naturalistic' (though some studies have attempted to address this problem by sampling at random throughout the day[9]).

Naturalistic speech analysis can be very cost-effective in the long-term. Although data transcription and collection is initially very time-consuming, the resulting dataset can be used multiple times both by the original researchers and by others if made publicly available (see, e.g., http//:childes.psy.cmu.edu). However, analyses must be conducted with care in order to take account of sampling constraints. Typically recording regimes are no more frequent than around 1 h/week,[10] and often considerably less frequent,[11,12] although one research group is currently engaged in a project to transcribe all the conversation taking place within a household.[12] Thus far from providing a direct window on what children say and hear, spontaneous-speech methods provide a tiny sample, and failing to account for this fact can result in erroneous conclusions. For example, consider the ongoing debate over whether 2-year-old children show full productivity with determiners (which, if so, is argued to constitute support for generativist over constructivist accounts of acquisition).[13–16] The question is, essentially, whether every noun that occurs with *a* in a particular child's speech also occurs with *the*. It is easy to see how thinly sampled data could underestimate productivity, for example by recording *the* + *ball* but missing uses of *a* + *ball*. Somewhat less intuitively, thin sampling can also overestimate productivity,[17] in this case by capturing frequent nouns that the child uses with both *the* and *a*, but missing less frequent nouns that occur only with one or other of the determiners.

One way to address this issue is to use the caregiver's speech as a control, randomly deleting utterances so that the corpus is matched in size to that of the child.[18] If, with this control in place, adults (who can be assumed to have a fully productive system) appear to show no greater productivity than children, this would suggest that apparent effects of lexical-specificity are in fact caused entirely by thin sampling. Another potential solution is to calculate how often particular combinations (e.g., *the* + *ball*) would be expected to occur, given the independent frequencies of the individual items (e.g., *the* and *ball*[19]). This is particularly informative when evaluating theoretical accounts that predict that particular combinations (e.g., *him* + *plays*) will be entirely absent.[20] A third approach is to attempt to address the sampling issue at the design stage. For many theoretical questions, it may be that a sampling regime of, for example, 2 h/day for 13 consecutive days is much more informative than a sample of 1 h/fortnight spread over the course of a year. Of course, for phenomena with an important developmental aspect, the reverse will be the case. The point is that, ideally, samples should be collected with a particular theoretical question in mind, which allows the recording regime to be tailored to that purpose. A fourth alternative is to eschew sampling in favor of a diary method that attempts to capture all instances of a particular structure, as has been

attempted for questions,[21] basic argument structure constructions[22] and individual verbs.[23,24] It is important to bear in mind, however, that even full sampling—or the most assiduous diarist—does not record everything that the child *could* say, only everything that she *does* say. Thus, if the goal is to investigate acquisition of less frequent (or novel) items and structures, then additional, less naturalistic paradigms will be needed.

Another measure based on children's spontaneous speech is the parental report questionnaire, the most popular of which is the MacArthur-Bates Communicative Development Inventory.[25] This test consists of two sections: words and gestures (8–16 months) and words and sentences (16 months and older). Although primarily a measure of vocabulary, it also includes some questions relating to basic syntax and morphology (e.g., adding *–s* to mark noun plurals). The MCDI correlates well with lab-based measures of children's performance,[26–28] and is very useful for obtaining a global measure of children's linguistic abilities (e.g., for selecting participants for a study or matching between-subjects experimental groups). However, since there is a significant degree of variability between parents with regard to reporting accuracy, there is debate regarding its suitability for (1) comparing groups of children from different socioeconomic groups, (2) profiling individual children (e.g., to identify those at risk for language impairment) and (3) evaluating the effectiveness of interventions.[29] Measures such as the MCDI also tend to underestimate children's verb vocabularies (as compared to their noun vocabularies), presumably because it is easier for parents to determine whether children know the word for a concrete object than a more abstract activity or state.[30] Finally, it is important to remember that, like all standardized measures, communicative development inventories must be normed individually for different languages and dialects (e.g., American versus British English).

EXPERIMENTAL PRODUCTION METHODS

Experimental Production methods lie along a continuum from *elicited production*, where children essentially choose what to say, albeit with some strong hints from an experimenter, to *repetition/elicited imitation*, where children are asked to repeat a particular string. In between these more loosely- and more highly structured extremes lie *syntactic priming* and *weird word order* paradigms, in which an experimenter models a particular sentence structure, though not necessarily the particular lexical items. This section is structured around these experiment subtypes, although it should be borne in mind that one often blends into another in particular studies.

Elicited Production

If the analysis of spontaneous speech data is the 'default' paradigm in language acquisition research then a family of methods known as 'elicited production' runs a close second. This situation has presumably arisen because recording children's speech (with or without specific prompts from an experimenter) enjoys the most face validity. Although the notion of a hierarchy of methods is almost certainly misguided, with the best method depending on the particular question being asked, there are indeed many cases in which some form of elicited production will be the most appropriate paradigm.

Elicited production methods - like experimental production methods more generally - lie along a continuum from least to most structured. At one end, children are simply shown a picture, animation, video or live enactment and asked a neutral question such as 'What's happening?'.[24,31] When it is desirable to exert more control over the constructions that the child could use for her answer, more constraining questions are often used. For example, an agent-focussed question such as *What's Ernie doing?* pulls for an active transitive response (e.g., *[he's] kicking the ball*), while patient-focussed questions such as *What happened to the ball*? or *What's happening with the ball?* pull for a passive (e.g., *it got kicked*) or intransitive (e.g., *it's rolling*).[32,33] Yet more control can be exerted by the use of a sentence-/stem-completion technique. For example, in most studies of the English past-tense,[34–36] the experimenter elicits forms using a script such as the following: *The bunny likes to run. Look, there he is running. Every day he runs. So yesterday he…*

An important consideration when designing elicited production studies is ensuring that the task makes communicative sense. Children are unlikely to respond willingly or appropriately to questions that they consider unnecessary or nonsensical. They are particularly sensitive to the difference between test questions and those that constitute a genuine request for information,[37] and even children aged as young as 2 years 6 months are able to take the speaker's prior knowledge into consideration when responding to questions and requests.[38,39] One solution is to set up a test context in the form of a game where the child is describing the scene not to an experimenter who has herself witnessed it, but to a third party (e.g., parent, puppet, or experimenter) who cannot see the scene

and has been explicitly given the task of noting down what is depicted in the scene[40] or of searching for a matching scene on a sticker or card,[41] which is then handed to the child to enable her to complete a sticker book or bingo grid[42] (such modifications also help to ensure that children remain motivated to cooperate). Conversely, if the aim is to elicit a question, it is important that the child herself is ignorant of the answer and is addressing a third party who knows the answer (e.g., because the question relates to an action that took place behind a screen, but in view of the question-answerer).[43–45] Similarly, a target utterance that includes a relative clause such as *The boy [who is washing the elephant] is happy* is felicitous only if it would be ambiguous without the relative clause (i.e., the picture to be described must show one boy who is washing an elephant, and another who is not).[46–49]

Because elicited production tasks vary so widely, it is difficult to predict in advance which age groups will be able to complete a particular task. The lower limit is probably around $M = 2$ years, at which age children have successfully completed studies investigating simple intransitives and transitives.[50] It is, however, often very difficult to elicit utterances from 2-year-olds, so other methods are probably more appropriate for children under three years of age.

Repetition/Elicited Imitation

A repetition (or elicited imitation) paradigm is generally chosen over elicited production when it is necessary to constrain the target utterance more precisely (e.g., when children must attempt to use a particular pronoun instead of a full lexical NP[51]), or when the target structure is sufficiently rare or complex that children are unlikely to attempt to produce it spontaneously, regardless of the discourse context. For example, this method has been successfully used to investigate children's production of wh-questions,[52] tag questions,[53] relative clause constructions[54,55] and constructions involving anaphora and control.[56]

Imitation is a valuable index of production ability because, provided that the target utterance is too long or complex to be stored in some kind of temporary phonological buffer, speakers instead appear to extract the semantic message of the target utterance, and then regenerate this utterance from scratch using standard production mechanisms.[56–58] Consequently, the paradigm is suitable for use with children as young as $M = 2$ years 6 months[3] and can be extended to older children and even adults, provided that participants are asked to perform a concurrent task that disrupts phonological storage such as backwards counting or simply waiting in silence.[59,60] Although few studies have addressed this issue directly, it would seem likely that a given elicited imitation study will yield a lower overall error rate than an equivalent elicited production study, on the assumption that children will retain some residual memory of the presented sentence. Thus in studies where the aim is to compare error rates across conditions, an imitation task will often be less sensitive (i.e., less likely to uncover differences that exist between conditions) than an equivalent production task. Indeed, this was the pattern observed in one recent study where the same children completed these two tasks with the same stimuli.[61] One way to increase the sensitivity of this method (or indeed elicited production[62]) is to use as the dependent variable not only a categorical measure (e.g., incorrect/correct) but a continuous measure such as reaction time (i.e., latency to begin the repetition)[63] or the duration of the repeated utterance (with the rationale being that easier utterances are produced more fluently, and hence are shorter in duration[3,64]).

Syntactic Priming

The term *syntactic priming* refers to the phenomenon whereby a speaker who has recently heard or produced a particular construction (e.g., the double-object dative, as in *The lifeguard tossed the child a rope*) is more likely to subsequently use this construction (e.g., *John read Sue a book*) than an alternative (e.g., the prepositional-object dative, as in *John read a book to Sue*), even when the *prime* and *target* utterance exhibit no lexical or semantic overlap.[65,66] The experimental paradigm of syntactic priming capitalizes on this phenomenon in order to elicit attempts at a particular construction. Like elicited imitation, this paradigm is particularly useful for eliciting attempts at rare or complex constructions such as the passive,[67–73] with the youngest age group in these studies generally around $M = 3$ years 2 months. This method is excellent for investigating which factors facilitate the use of otherwise difficult constructions and for tracking the processing of syntactic structure across development, since the same method can be used, with little or no modification, across children and adults.[42] However, a difficulty arises in that both the mechanisms that underlie syntactic priming and the role that lexical items play in priming tasks are still poorly understood.[42,74–77] Consequently, it is premature to assume either (1) that lexically dependent priming effects provide evidence for constructivist-style lexical frames or (2) that successful priming in the absence of lexical overlap indicates adultlike abstract

knowledge of the primed construction; assumptions shared by most studies of priming in children (though see Ref 29).

In *weird word order* studies (a subtype of priming studies), children are taught *novel* word order constructions, of which they—by definition—have no prior knowledge. For example, in the first such study,[78] the experimenter used a non-English word order construction with a novel verb (SOV, e.g., *Elmo the car gopping* or VSO, e.g., *Tamming Elmo the car*) to describe a new action. Children were then asked to describe a scene in which the same action was performed by different characters, in order to investigate whether they would imitate the experimenter's weird word order (the aim of the study was to test the claim that children's initial knowledge of word order is tied to individual verbs[79]). Follow-up studies have extended this paradigm to investigate verb frequency effects (using real verbs[80]), other languages,[81] and single-argument constructions,[82] which allow the paradigm to be used with children as young as $M = 2$ years 4 months (as opposed to $M = 2$ years 8 months in the original study). When designing a weird-word-order study, it is important to include a control condition with familiar high-frequency verbs, in order to obtain a baseline measure of how often children will use weird word orders that they know to be incorrect, simply to please the experimenter by 'playing the game'.[80]

General Considerations when Using Production Paradigms

Most of the considerations that arise regarding production paradigms are also shared by many comprehension and judgment paradigms, and will be discussed subsequently. Four considerations apply particularly (though not exclusively) to production paradigms. The first is that it is important to categorize children's responses not as simply correct/incorrect, but as to the particular type of error made. For example, when attempting to produce *wh-* questions, children's errors include non-inversion (e.g., **What she can eat?*), auxiliary-omission (e.g., **What she eat?*) and auxiliary-doubling (e.g., **What can she can't eat?*), all of which have different theoretical implications. Even correct responses can vary in a manner that may have theoretical consequences. For example, in a study where the correct response is some form of transitive utterance, if children consistently use a particular subject+auxiliary+morpheme+pronoun combination (e.g., *He's _ing it*) rather than producing more varied utterances (e.g., *Elmo's rolling the ball; Minnie's eating the cake*), this may suggest evidence of a particular lexically specific pattern in children's grammars.[83] In many studies, children produce a large number of null, irrelevant or 'other' responses. Here, it is important to distinguish between those that constitute (1) a clear attempt at the target, (2) clear avoidance of the target, and (3) data missing at random (e.g., because the child is distracted). For example, if the DV is the proportion/percentage as opposed to raw number of errors of a particular type—that is, whenever there are significant missing data—it would seem important when calculating the denominator to exclude (3) but include (1). The appropriate treatment of apparent avoidance responses (2) is less clear. One possible solution is to run both a raw-data analysis in which they are included, and a proportional analysis in which they are excluded. Note that proportionalized data should be transformed before analysis (e.g., arcsine-square-root transformation), or avoided altogether (e.g., by the use of mixed-effects models).[84]

The second consideration is that certain errors may be a consequence of particular methods or prompts used in the experiment. For example, rates of *wh*-question non-inversion error (e.g., **What she can eat?*) are presumably boosted by experimenter prompts such as 'I wonder *what she can eat*. Ask the dog *what she can eat*'. The solution is to check whether the same errors are observed when using different prompts[85] and in spontaneous production data.[86]

The third consideration relates to the use of fillers to avoid participants becoming aware that they are being asked to produce repeated instances of the same construction. In general, production tasks are used to investigate whether or not children are capable of producing a particular target utterance (e.g., with novel verbs). Thus it is of no particular concern if children become aware that this is what they are being asked to do; indeed it may benefit the study by increasing children's attempts at the target construction. Furthermore, since the overall number of trials is limited by children's attention span and willingness to continue, few designs will have trials to spare for fillers. An exception is some priming studies where children's ability to produce the constructions is not at issue and the question is rather which of two possible constructions they will produce. Here fillers would seem advisable,[42] particularly if the same study is to be run with adults, who are likely to notice repeated use of particular constructions. In studies that do not use a priming design, self-priming can be a problem. For example, experimental studies of past-tense production find higher rates of overregularization errors (e.g., **sitted*)

than naturalistic studies, and one study demonstrated empirically that this is due to a self-priming effect.[87] Few researchers will wish to adopt the drastic solution adopted in this study—presenting just one trial per participant—but if it is not possible to use fillers, it is at least important to present all trials in a counterbalanced or randomized order so that any self-priming effect is cancelled out across items, and to remember that absolute rates of error/correct production may be unreliable.

A final consideration to bear in mind when designing production studies (and, indeed, many comprehension and judgment studies) is whether to use novel items (typically nouns or verbs) or familiar, real items. The advantage of using novel items is that they allow the researcher to investigate whether or not children have acquired a generalization (or 'rule') that is independent of particular lexical items; for example that English uses [SUBJECT] [VERB] [OBJECT] word order, and forms (many) past-tenses using the morpheme *-ed* (e.g., *play+ed*). If familiar items are used instead, it is impossible to tell whether children have formed a generalization, or simply acquired a low-level rule relating to a particular lexical frame (e.g., *I'm kicking X*), or indeed a fully-specified lexical form (e.g., *played*). For this reason, the use of novel verbs and nouns is generally standard in investigations of syntax (see[88] for a review) and morphology (e.g.,.[36,89,90] Where novel items are used it is important to check that they are not only plausible but also reasonably prototypical as words in the relevant language (unless they are deliberately intended to be non-prototypical). For example, when using novel verb stems to investigate English past-tense production (e.g., *blick*), it is important to design stems that are phonologically typical for English.[89] When using novel verbs with novel meanings, it is important to check that the target language encodes the relevant linguistic features (i.e., that a verb with this meaning could in principle exist, but simply happens not to), for example by obtaining plausibility ratings from adults..[91] A second advantage of novel items is that they avoid 'garden path' effects whereby children use or interpret familiar (particularly high frequency) items in the way that they are most commonly used, whether or not this is appropriate given the linguistic context in which they are presented in the task.[61,92]

The disadvantage of using novel words is that they increase the difficulty of the experimental task, by imposing the additional requirement of remembering the novel form (and, sometimes, also its meaning). Thus when there is nothing particular to be gained by the use of novel items, it would seem sensible to avoid them. For example, in a study comparing children's performance at forming questions with different *wh-* words and auxiliaries (e.g., *what does…, why is…*), there would seem to be no reason to use novel as opposed to familiar main verbs. Furthermore, for some study designs there is a positive reason to use familiar items. For example, if the goal is to investigate graded effects of verb frequency and phonology on inflection,,[35,93] it is practicable to do so only by selecting (say) 50 familiar items that vary continuously along these dimensions (as opposed to generating 50 novel verbs and presenting them with different frequencies during training). Another disadvantage of using novel words (particularly verbs) is that it is difficult to be sure that participants have acquired the precise intended meaning (where this is important). Finally, studies with novel words lack a certain face validity, though this problem is probably more perceived than actual, as it is hard to see any in-principle difference between real words that the child has yet to encounter and novel words created for the purposes of an experiment (provided that they are phonologically and semantically plausible). As ever, the best solution to these methodological problems is triangulation: Where it would make sense to do so, findings with novel items should be replicated with real items and vice versa.

COMPREHENSION METHODS

Comprehension methods do not require children to produce any language, but simply to respond to a presented sentence in some manner such as pointing—or simply visually attending to—a picture, animation or audio stimulus such as a sentence. Such methods are extremely valuable when children are too young to complete a production task. In addition, they are more suitable than production tasks for addressing some questions, even with older children and adults. For example, adults find passive sentences easier to process with some verbs than others (as measured by latency to choose a matching picture), even though they would presumably be able to produce all correctly.[94] Note also that (full) comprehension does not always precede production, particularly for complex constructions such as those involving relative clauses[95] or anaphora.[40] Also included in this section on comprehension tasks are studies where participants respond to sentences on the basis of their phonological properties (e.g., whether a sentence sound more English-like or Russian-like), even though this does not necessarily entail comprehension in the sense of extracting meaning.

Act-Out

Act-out tasks are generally used to assess children's knowledge of syntax; for example whether English-speaking children know that, given a NOUN VERB NOUN string, the first noun (usually) denotes the AGENT and the second the PATIENT. Children are given appropriate toys and asked - for example - to 'Make Ernie kick Big Bird'; the dependent measure being simply whether they correctly produce an enactment with the first-named character as agent. In some studies[24,83,96,97] novel verbs are used to rule out the possibility that children are succeeding on the basis of verb-specific templates (e.g., KICKER kick KICKEE) as opposed to some kind of abstract construction (e.g., AGENT ACTION PATIENT or SUBJECT VERB OBJECT). Others use familiar verbs, but pit word-order against other cues to agenthood such as case-marking, animacy (or plausibility),[98] prosody,[99] and verb/construction semantics.[100,101]

As with production tasks, it is important to ensure that the sentence that the child hears makes communicative sense given the discourse context. For example, early act-out studies of relative clause sentences (e.g., *The dog that jumps over the pig bumps into the lion*) yielded inconsistent results, apparently because when given only one of each animal toy, the appropriate response is somewhat unclear.[102] If children are first given a lead-in such as *This* [other] *dog jumps over the pig. This dog pushes a cow*, the relative-clause sentence *The dog that jumps over the pig bumps into the lion* now makes communicative sense, and performance improves dramatically.[103]

A disadvantage of act-out tasks is that they can be (from an adult perspective) surprisingly difficult. For example, one study of simple transitives[24] found that only 4/10 children aged 2 years 8 months to 2 years 10 months performed at above-chance levels in a novel-verb task. This may be because children have difficulty keeping the sentence in mind while planning the enactment or because, having picked up one character at random (or because they prefer it to the other) they automatically assign it the dominant role of agent, ignoring cues such as word order and case marking. Thus for children of this age and younger, the following less demanding tasks may be more appropriate.

Pointing

In classic picture-choice studies, children hear a sentence (e.g., *The cat was pushed by the dog*) and are asked to point to the matching picture; typically, one correct and one relevant foil (e.g., a cat pushing a dog vs a dog pushing a cat). This method is particularly common amongst studies investigating children's acquisition of later-acquired constructions such as the passive,[73,104–106] and sentences with quantification (e.g., *Every alligator is in a bathtub*).[107,108] However, it is possible that the use of still pictures to illustrate dynamic actions may not only depress overall levels of performance, but also change the pattern of results. For example, the studies cited above find that children have particular difficulty with non-actional passives (e.g., *The cat was heard* [cf. *pushed*] *by the dog*), but this may be due, at least in part, to the difficulty of illustrating non-actional verbs pictorially.

More recent studies investigating comprehension of the dative[109] and in/transitive[110] constructions have used live action or cartoon movies displayed side-by-side on a computer screen (or asked children to point to one particular character),[111] and it would seem likely that most pointing studies can only benefit from the use of this methodology. Typically the movies freeze on completion in such a way that the path of action remains clear, meaning that the method has all the advantages of picture-choice, but avoids the major disadvantage. It is perhaps for this reason that pointing, using movies, has become the paradigm of choice for recent studies[112–114] in which children are taught novel constructions (e.g. a [SUBJECT] [LOCATION] [NOVEL-VERB+o] construction with the meaning of appearance;*The sailor the pond neebo-ed* means 'the sailor sailed onto the pond from out of sight').

The major advantage of pointing over other comprehension paradigms is that it produces an unambiguous response on virtually every trial with very few missing datapoints. It is also less demanding than an act-out task. For example 2-year-olds demonstrate comprehension of the transitive construction in pointing[110] but not act-out.[17] The main disadvantage is that the paradigm is generally unsuitable for children younger than 2 years, who have difficulty understanding the instructions of the task.

Intermodal Preferential Looking

For children younger than 2 years, the intermodal preferential-looking (IPL) paradigm can be used instead. The method is similar to pointing, except that the child makes no explicit response. Instead, the dependent measure is simply the time spent looking at a particular picture or movie (or an alternative measure such as the single longest look). The method exploits listeners' natural tendency to spontaneously look to a word or picture that matches the language that they are hearing, though the audio passage typically begins with an additional prompt

such as *find* or *where's*. Studies focussing on the transitive construction have demonstrated knowledge of canonical linking (i.e., that the first noun is the agent and the second the patient) as young as 1 year 9 months. For example, in a novel verb study, children at this age associate *The boy is glorping the girl* with a video in which a boy is performing an action on a girl as opposed to the reverse (and vice versa).[5] Although there is controversy over the finding that success is apparently contingent upon training with familiar verbs,[115,116] the important point here is simply that the method works with one-year-olds (though this caveat should certainly be borne in mind).

Indeed, studies of word-learning (as opposed to syntax) have demonstrated that preferential looking tasks can be used successfully with children as young as 1 year for novel nouns,[117] and 6 months for very familiar words such as Mummy and Daddy.[118] The method may, however, be less reliable with older children and adults, whose visual response to the audio stimuli may be more fleeting than that of younger children. For developmental comparisons, alternative measures such as pointing or looking while listening may be preferable. There is also the problem that looking-time, unlike pointing, does not constitute an unambiguous measure of comprehension. For example, a child could in principle recognize one video as matching the sentence, but spend longer looking at the non-matching video (possibly because its discrepancy with the audio makes it more interesting). Consequently, there is controversy over how data from preferential-looking studies should be interpreted (see Box 1).

Looking While Listening

For some studies, the most appropriate dependent measure is not simply which picture/animation is chosen, but the time taken to make this choice (i.e., *latency* or *reaction time*) or the time-course of looking (e.g., the order in which elements of the scene are inspected). Until recently, most studies of this type have measured looking by video-recording the child's eyes for subsequent frame-by-frame coding.[120,121] Relevant here is a particularly innovative method that combines three paradigms: looking while listening, priming and act-out.[122,123] The child hears either a DO-dative (e.g., *Send the frog the gift*) or a PO-dative (e.g., *Send the gift to the frog*) prime and then acts out a sentence with a temporary ambiguity (e.g., Show the hor*se the book [DO]* or Show the hor*n to the dog [PO]*). The time-course of the child's looking between the four toys (*horse, dog, book, horn*) is recorded, to investigate whether children show priming

BOX 1

INTERPRETING PREFERENTIAL-LOOKING DATA

Preferential-looking data can be analyzed in many different ways, though each suffers from at least one disadvantage.

Raw looking times. The simplest method compares total looking time to the target and foil. *Disadvantage:* Because the two measures are mutually exclusive, this violates the statistical assumption of independence (like testing whether a coin is fair by comparing 60 heads vs 40 tails).

Comparison to chance. In the coin-toss scenario, the correct test compares observed performance (60 heads) against chance (50 heads). Looking time to the target screen can also be compared to chance (50% of total looking time) in this way. *Disadvantage:* This assumes that chance responding entails looking at the two scenes equally; in fact children may show a baseline preference for one scene.

Comparison to baseline. Before the test trial, children complete a baseline trial with the same scenes and uninformative audio (e.g., 'look they're different now'), in order to uncover any baseline preference for a particular scene. The dependent measure is then increased looking to the target screen (i.e., looking time to target in test trial minus looking time to target in baseline trial). *Disadvantage:* Children may switch their preference due to boredom with the scene to which they looked longer during the baseline trial. This can be addressed by using a different group of children in the baseline condition (though at the expense of statistical power).

Number of children looking longer at matching than non-matching screen. A problem with all the above measures is that the data are usually analyzed using parametric tests that treat looking time as if it can be meaningfully interpreted as a continuous measure (i.e., as if a child who looks at the target for twice as long for Sentence A than Sentence B understands sentence as 'twice as well'). This method compares the number of children who looked for longer at the target than foil screen to chance using a binomal test (e.g., 15/20 participants vs 10/20, $p = 0.02$). *Disadvantage:* Collapsing across all trials in this way (so that each child provides only one datapoint) results in decreased power, and does not control for baseline preferences.

Mixed effects models. Both of these problems can be addressed by coding each baseline and test trial for whether the child looked longer at the foil or target scene and analyzing these data using binary logistic mixed-effects models,[119] to which we return in the discussion.

in comprehension; for example whether a DO-prime (*Send the frog the gift*) causes children to look first to the horse (e.g., *Show the horse the book* [DO]) rather than the horn (e.g., *Show the horn to the dog* [PO]).

Recently, more researchers have begun to use fully automated eye-trackers (as opposed to simply video-recording children's eyes); a trend that seems likely to continue as the equipment becomes increasingly affordable and child-friendly. For example, one recent study[124] found that 3–5-year old children use both order-of-mention and pronoun gender to guide their interpretation of an otherwise-ambiguous referent (e.g., *Puppy* [wearing boys' clothes] *is having lunch with Froggy* [wearing girls' clothes]. *She wants some milk*).

Conditioned Head-Turn Preference Procedure

A close relative of preferential-looking is preferential-listening (or—to give it its full title—the conditioned head-turn preference procedure). The difference is that instead of choosing how long to look at a particular screen, children choose how long to listen to a particular audio stimulus (i.e., a word, phrase, sentence, or passage). In most cases, the paradigm uses two loudspeakers—placed to the left and right of a central fixation point—with a light above each. In a training phase, the child learns that when a light above a speaker is flashing, audio (here, music or speech other than the test stimuli) will continuously play from that speaker for as long as she fixates visually on the light. In the test phase, different audio (e.g., different languages or familiar versus novel words from a training session) is available from each speaker,[125–129] though usually only one or other speaker per trial. As infants control the duration of playback, this allows the experimenter to investigate which audio stream (e.g., native vs non-native language; words vs non-words) infants prefer. It is not always easy to predict in advance whether children will show a novelty- or familiarity preference (though attempts have been made to do so[130,131]), but any consistent preference across children demonstrates discrimination between the two types of stimuli.

One variant of this design uses a single speaker plus light, with only one of the two audio passages available on each trial.[132] Two other variants, *non-nutritive sucking* (on a dummy/pacifier) and *kicking*, allow this paradigm to be used with newborn[133] and even unborn[134] infants. Any of these paradigms can be used with a habituation–dishabituation design, which capitalizes on infants' tendency to become bored with repeated presentations of identical or similar stimuli and hence to decrease rates of looking, sucking or kicking across consecutive trials. If a sufficiently different stimulus is then presented, dishabituation occurs, and the rate recovers. For example, in one study,[135] newborns were habituated to sentences in one language before being presented with either (1) further sentences in the same language produced by a different speaker (control condition) or sentences in a different language (switch condition). Infants demonstrated greater recovery of sucking in the switch than control condition, and hence discrimination between the two languages (here, French and Russian).

A caveat is in order with regard to the interpretation of listening-time data. Whenever a difference between conditions is observed, this demonstrates simply that children perceive some difference between the audio stimuli, and not that they necessarily understand its implications for language learning. For example, a number of studies[127,136] have shown that infants discriminate between otherwise-identical strings with different pauses and pitch contours, which are potential cues to NP/VP phrase boundaries (e.g., *Many different kinds of animals/live in the zoo*). However, such findings do not demonstrate that children know that these *are* potential cues to phrase boundaries, let alone that they actually use these cues to segment speech into phrases.

Functional Neuroimaging Methods

We conclude our discussion of comprehension methods with a brief survey of the use of functional (as opposed to structural) neuroimaging methods with children. Neuroimaging can also be combined with a production or judgment task, but the simplest and most commonly-used method - particularly when studying children - involves simply presenting participants with linguistic stimuli and measuring brain responses. *Functional Magnetic Resonance Imaging (fMRI)* uses a scanner to measure local changes in cerebral blood oxygenation. This method has been used to investigate sentence processing in children as young as 0;3, with studies demonstrating that, even at this age, certain neural regions show greater activation to normal than backwards speech.[137,138] A

similar finding was observed in a study[139] using the technique of *Near Infra Red Spectroscopy (NIRS)*. Like fMRI, NIRS works by measuring changes in blood oxygenation, in this case by passing near-infrared light through brain tissue, and monitoring levels of absorption, which vary according to hemoglobin oxygen levels. With regard to both techniques, it must be borne in mind, however, that there exists a degree of individual variability with regard to brain structure, which much be controlled out using a process of spatial normalization. Consequently, it is extremely difficult to compare findings across individual participants or indeed across different studies.[140]

A more common technique, particularly in studies with children, is *Electroencephalography (EEG)*, in which electrodes placed on the scalp are used to measure voltage fluctuations that occur as a result of neuronal firing. Compared with fMRI, EEG has much poorer spatial resolution (because each electrode covers a wide area), but much better temporal resolution (because voltage changes are much quicker than changes in blood flow/oxygenation). Hence EEG is used to measure how patterns of brain activation change in real-time as a sentence is processed. In studies with adults, two ERP signatures (i.e., changes in activation patterns) have received particular attention.[141] A negative wave peaking at around 400 ms (N400) is associated with semantic processes, and is of larger amplitude in cases of semantic incongruity (e.g., *He ate a **car***). Syntactic violations and sentences requiring syntactic reanalysis (e.g., *The horse raced past the barn fell*) are associated with a late positive wave (P600), preceded by an earlier negative wave. These patterns have also been investigated in children. For example, one study[142] found that 14-month-old children showed an N400 for semantic mismatches (e.g., hearing 'house' paired with a picture of a dog). Interestingly one pair of studies[143,144] found that syntactic violations (e.g., *My uncle will **watching**...*) seem to elicit a P600 in 3–4 year olds, with the effect failing to reach significance for a younger group aged 2 years 6 months. Of course, it would be a gross overinterpretation to claim on this basis that only 3–4 year olds have any knowledge of morphosyntax; this example illustrates the importance of exercising caution when interpreting such findings.

General Considerations when Using Comprehension Paradigms

Many of the same general considerations that apply to production paradigms apply equally to comprehension paradigms (e.g., ensuring that the task makes communicative sense; defining and excluding 'missing' trials). Another consideration that applies in both cases, but perhaps especially to comprehension, is ensuring a sufficient number of trials and participants. Above we argued that the outcome of preferential-looking/pointing studies should be a binary measure for each trial (looked longer at target than foil, or pointed to target) for subsequent analysis using binary logistic mixed-effects models. Because binary measures are less powerful than continuous measures, this increases the importance of including as many trials as possible, given the concentration span of the participants. Since participants in these studies are generally aged 2 years or younger, this will generally mean including only a handful of trials. One way to compensate for this reduced power at the item level is to increase power at the participant level, by recruiting a large number of participants.

When choosing an experimental paradigm, it is important to be aware of the typical drop-out rate for each possible method, and to consider whether or not it is unacceptably high for the claim under investigation. For example, if the claim is that children of a particular age 'succeed' at a particular task, it is important that the drop-out rate be relatively low (otherwise the claim does not generalize to the relevant age-group population as a whole). High drop-out rates are less problematic in experiments looking for differences between experimental conditions (as long as they are roughly equal 'at-random' dropouts from all conditions). Even for such designs, though, it remains possible in principle that any between-condition differences observed would have disappeared (or even reversed) had it been possible to include all children (particularly if the drop-out rate is greater than 50%). Drop-out rates are a particularly important consideration for the comprehension paradigms discussed here as (1) these paradigms are often used to make claims regarding the general abilities of children at a particular age and (b) drop-out rates seem to be particularly high (though we are aware of no formal comparisons between paradigms), perhaps because they tend to be used with younger children. For example, one recent pointing study which drew general conclusions regarding 'argument structure in the third year of life '[145] reported a drop-out rate of 50%; the majority due to either a side bias or 'fussing out' (a term which is a little unclear to the present [British] authors).

Another consideration that applies to all experimental study designs, but perhaps especially comprehension methods, is counterbalancing. Counterbalancing is particularly crucial in preferential-looking/pointing studies with a choice between two

screens (L/R) or two characters in a single scene. It is crucial that counterbalancing (or matching) removes any potential extraneous cues to targethood such as side (e.g, target always on the left), direction in which the action unfolds (e.g., target always L -> R), first movement (e.g, agent always before patient), total amount of movement (e.g., agent moves more than the patient), identity/size/animacy of agent (e.g., agent is always a particular character or bigger/more animate than the patient). Children often show unpredictable preferences for particular scenes, which can be controlled for by having each scene the target for one counterbalance condition and the foil for another, and/or by using a baseline control group (ideally between-subjects).

The issue of counterbalancing relates to a final consideration that is important for both comprehension and production methods; whether trials are to be presented in (1) blocked, (2) random, or (3) counterbalanced order. Blocked presentation (e.g., five active sentences followed by five passive sentences) presumably incurs the lowest task demands, and hence is generally used when the aim is to investigate whether or not children display some particular phenomenon in the limit (i.e., when the experiment is deliberately set up to help them). For this reason, most child priming studies use blocked designs, in order to allow priming to build up over the course of the block,[67,70] but see.[42] Thus a blocked design is the least suitable for obtaining an objective measure of performance (e.g., error rate at a particular age) that is generalizable to other contexts. Neither is such a design suitable for investigating differences between different trials/items (e.g., different verbs), unless these are additionally randomized within the block. Studies with these goals therefore generally use a random or counterbalanced order of presentation. Full counterbalancing, whereby each item is followed by and preceded by every other item an equal number of times, is the best option, as it allows serial position effects to be investigated and—if present—controlled out statistically. In practice, this is rarely done, as with more than a handful of items, an extremely large number of trials are needed to include all possible orders for each participant (some studies[3] are fully counterbalanced across all participants combined, but not within each participant individually). Randomization is therefore more commonly used. Although this method generally neutralizes serial-position effects (given a sufficient number of trials), it does not actually allow for specific detailed investigation of them (though if mixed-effects models are used, trial number can be included as a fixed effect to control out order effects at a general level). Other options are to use a Latin Square design, which provides some—but not full—counterbalancing, or to combine ordering strategies (e.g., using full counterbalancing at the block level, but randomization within each block).

JUDGMENT METHODS

Judgment methods are somewhat similar to comprehension methods, in that participants do not produce language, but—in this case—judge the grammatical acceptability or truth value of a particular sentence (or sentence interpretation). Although this obviously requires comprehension of the sentence, comprehension and judgments methods are used to answer different questions. For example, considering an ungrammatical sentence such as '*The zebra goes the lion', different theoretical questions are addressed by a comprehension task[100,101] (Is the preferred interpretation that the zebra *makes* the lion go or goes *to* the lion?) and a judgment task (assuming the first interpretation, how acceptable do participants consider this sentence to be?). Thus the primary reason for selecting a judgment task over an alternative is not that it is easier for children to complete (though this may sometimes be the case), but that it maps more directly onto a particular theoretical claim.

Grammaticality/Acceptability Judgments

When a theoretical account makes the prediction that some factor such as age or verb frequency will affect the unacceptability of errors,[146] this prediction can be most directly tested using a grammaticality/acceptability judgment task (most studies use the latter term in participant instructions to avoid the intrusion of prescriptive rules of grammar). The child hears a sentence, generally spoken by a puppet who, the child has been told, sometimes makes mistakes or 'says things a bit silly'. It is important to present each sentence with an accompanying animation, video or live enactment, in order to ensure that the child understands the intended meaning. This is particularly important for ungrammatical sentences. For example, in one study with no visual materials,[147] causativization errors with *laugh* (e.g., **The funny clown laughed Lisa* meaning 'the funny clown made Lisa laugh') were misinterpreted as errors involving the omission of *at* ('the funny clown laughed at Lisa'). This problem was addressed by the inclusion of animated cartoons in a subsequent study.[148,149]

In a binary judgment study, the child's task is simply to tell the puppet whether he 'said it right' or whether it 'sounded a bit silly', either by moving him to a tick or a cross[147] or by presenting him

with a reward or punishment, as in the truth-value judgment studies discussed below. The advantage of a binary task is that it can be completed by children as young as 3–4 years.[150] Below this age, children have a tendency to base their judgments on truth value or other factors (e.g., whether a particular action is 'good' or 'naughty'), rather than grammaticality.[151] The disadvantage of a binary task is that it cannot be used to test theories that predict graded judgments. For example, some theoretical proposals predict a continuous negative relationship between verb frequency and the acceptability of errors with that verb[146] (e.g., **The clown laughed/giggled/chuckled Lisa* [='made Lisa laugh/giggle/chuckle']), even though all would presumably be rated as unacceptable (at least by adults) in a binary judgment task.

Thus when studying adults and older children, the use of a graded judgment scale is preferable. One paradigm suitable for use with children aged 4 years 0 months and upwards[149] involves children first selecting either a red or green counter to denote unacceptable/acceptable, and then placing this counter on a five-point 'smiley face' scale to denote the degree of perceived (un)acceptability. This paradigm is suitable for investigating morphological errors at the single-word level[36,152] (e.g., **sitted; *unbreak*), as well as the sentence-level argument-structure overgeneralization errors discussed above (e.g., **The funny clown laughed Lisa*),[153,154] and can be used with novel verbs.[36,91,94,148,149,155]

It should be emphasized that judgment tasks of this type are really the only suitable way to investigate whether a child considers a particular utterance to be ungrammatical. Production tasks are not suitable because, while a child may not produce an utterance such as **The magician disappeared the rabbit* (even if placed under discourse pressure to do so), this does not constitute evidence that she considers it to be ungrammatical. Even the production of an alternative formulation (e.g., *The magician MADE the rabbit disappear*) does not constitute strong evidence that the erroneous form is ruled out by the child's grammar; this form may simply be dispreferred. Comprehension measures are unsuitable because utterances such as **The magician disappeared the rabbit* are easily interpretable, even though they are ungrammatical. Thus a child who considered such an utterance to be ungrammatical would still be likely to 'succeed' at an act-out or preferential looking/pointing task.

Yes/No and Truth-Value Judgments

In some cases, the theoretical question under investigation is not whether or not a particular utterance is grammatically acceptable, but whether or not a certain reading is possible. For example, *Mama Bear is washing her* is a perfectly acceptable sentence, but it can mean only 'Mama Bear is washing someone else' not 'Mama Bear is washing herself'. In *yes/no judgment* tasks, children are shown a picture (e.g., Mama Bear washing herself) and asked a *yes/no* question (e.g., *This is Goldilocks. This is Mama Bear. Is Mama Bear washing her?*). If children understand the relevant aspect of syntax ('Principle B'), they will answer 'no', since the picture shows Mama Bear washing *herself* whereas *her*, in this context, can refer only to *Goldilocks*.[156]

A problem with this task is that children may be reluctant to 'contradict' an adult (particularly an unfamiliar experimenter) by answering 'no'. One way to address this problem is to have the sentences spoken by an animal puppet, which the child can reward or punish, for example by giving it a cookie or a rag,[157] thus avoiding the need for any verbal response. This method is sometimes referred to as the *truth value judgment paradigm* to differentiate it from the *yes/no* judgement paradigm discussed above; though the terms are often used interchangeably.

A truth-value (or yes/no) judgment task is the most appropriate method when the theoretical question under investigation is which possible sentence interpretations children will allow (e.g., whether *her* may refer to *Mamma Bear* in the sentence *Mamma Bear is washing herself*). Thus this task is the paradigm of choice when studying other cases of (im)possible interpretations such as 'Principle C' sentences (e.g., *She$_i$ washes Sarah$_i$*[158]) and sentences with universal quantification[159] (e.g., *Every farmer is feeding a donkey*).

As with production and judgment tasks, it is important to ensure that the yes/no or truth-value judgment task makes communicative sense to children. Speakers do not typically ask questions about truth value (e.g., *Is every farmer feeding a donkey?*) when (1) the answer is already known to both the speaker and the listener, (2) the answer is not relevant to any ongoing discussion, or (3) one particular possibility was never under consideration. Thus it is important to have a lead-in story that raises the alternative possibilities and situates them in an ongoing discussion (e.g., Farmer Jack was telling the other farmers that he may be away from his farm and hence unable to feed his donkey) and a character who does not know the outcome (e.g., a puppet who is guessing or misrembering the story[159]).

Even with these controls in place, *yes/no* and truth value judgment tasks suffer from two important

shortcomings. The first is that, as for grammaticality judgment tasks, children may not always base their decisions on the intended factors. For example, in one study,[40] some children rejected accurate picture descriptions such as *Mama Bear is washing her* (where *her* = Goldilocks) and, when questioned, objected to the use of pronouns in this context (i.e., they wanted to hear *Mama Bear is washing Goldilocks*). The second is that, even when a puppet is used, some form of *yes*-bias still remains. In fact, the problem is not so much a general *yes*-bias, which can be corrected for by the calculation of a Receiver Operating Characteristic score from signal detection theory[40] (though this does not seem to be done as a matter of course). The problem is that children display a readiness to accept picture descriptions that are 'good enough' even if they are not perfect,[160] and are not the descriptions that children would themselves produce. For example, in one study most children who incorrectly accepted violations such *Mama Bear is washing her* in place of *Mama Bear is washing herself* actually produced the correct version when asked to describe the pictures themselves.[40] One approach is to avoid this method altogether. Most truth-value judgment tasks can be converted into an equivalent binary pointing task (e.g., one screen would show Mama bear washing Goldilocks, the other Mama Bear washing herself).[107,108] The disadvantage here is that this method reveals only participants' preferred interpretation, not whether or not they would accept the other. Another option is to conduct a graded version of the judgment task, such as—for example—asking children to reward the speaker with a 'small, big or huge strawberry'.[160] Importantly, this study found that children who accepted 'good enough' (but underinformative) descriptions in a binary version of the task did not deem such sentences as deserving of the largest reward.

General Considerations when Using Judgment Paradigms

Training and practice trials are important for all experimental designs, but particularly so for judgment studies where it is necessary to establish the basis on which judgments are to be made (e.g., acceptability vs truth value) and—for graded tasks—the appropriate use of the scale. It is necessary to strike a balance by providing practice with the broad classes of errors that will be encountered in the main study (e.g., overgeneralizations of verb morphology) but avoiding the actual error (e.g., overgeneralizations of the past-tense *-ed* marker), which could unduly influence judgments.

When designing stimuli, it is important to ensure that there is an equal split between trials where the correct response is acceptable/'yes' and those where it is unacceptable/'no', and that this holds across all the cells of the design. For example, it should not be the case that all intransitive sentences are grammatical and all transitive sentences are grammatical. This is important for two reasons. First, it precludes the possibility of task-dependent strategies (e.g., giving low acceptability ratings for all trials of a particular type). Second, it allows ratings for ungrammatical items to be converted into difference scores that reflect the degree of dispreference for ungrammatical versus grammatical uses. This is important as it allows any general dispreferences that children may show for particular items (e.g., verbs that denote 'naughty' actions) to be controlled out.

CONCLUSION

The conclusions of this review are summarized in a checklist for researchers at the experimental-design stage. Our aim is not to preach or patronize—we would certainly not claim to have run a study that ticks every box—but a checklist seems to us to be the most concise and helpful way of summarizing the issues covered here.

- Is a *naturalistic, production, comprehension or judgment paradigm* most appropriate for answering the particular theoretical question under investigation? Can more than one be used to provide converging evidence?
- Is the task *suitable for children at the age under investigation* (or is it too easy or difficult?). Should a *parental report or standardized test* be used to select children of a suitable age/stage of development? If different age groups are included, is the task *equally suitable for all age groups*, or are modifications for older or younger participants required?
- Does the experimental task make *communicative sense*, or does it place children in an unusual or infelicitous communicative context (e.g., providing an answer that the questioner already knows)?
- Relatedly, is there sufficient *motivation* for children to continue to participate? Is the game enjoyable? Is some kind of reward for each trial (e.g., sticker) appropriate (though if the game is 'too fun' or the reward too desirable, children can lose focus on the task)?

- Is a *between-subjects, within-subjects or mixed design* most appropriate? Within-subjects designs have more power to detect experimental effects but can suffer from carry-over effects. Where this is a potential concern (e.g., particularly for baseline control conditions) a between-subjects manipulation may be more appropriate.
- Is there an advantage to be gained by the use of *novel items* (verbs, nouns, etc.)?
- Are *fillers* necessary?
- Is it better to use (1) *blocked lists*, (2) *randomization*, or (c) *counterbalanced lists* for factors such as trial order, location of target screen (L/R, for pointing/IPL studies), sentence type, real versus novel verbs etc.
- Are there a *sufficient number of trials and participants* to observe a significant effect if one exists (is a formal power-analysis necessary)?
- Is the anticipated *drop-out rate* sufficiently low, given the claim under investigation?
- Is there a *control condition* to rule out alternative simpler explanations of possible findings of interest; possibly one that allows for exclusion of children who do not understand the task?
- Is the *dependent measure* as sensitive as possible? Is it possible to obtain an appropriate continuous measure (e.g., reaction time, duration, graded judgment) in addition to a binary measure (correct/incorrect, grammatical/ungrammatical)? Does it avoid potential ambiguities (e.g., pointing vs looking time measures) and biases (e.g., *yes*-bias)? Is the measure suitable for the most appropriate and most powerful statistical analysis?
- How will *responses be classified?* What types of responses will be classified as missing? Will the remaining data be proporptionalized (and, if so, transformed) or analyzed using statistical methods that are robust to missing data (e.g., mixed-effects models)?
- Could some particular responses (particular errors) be a *function of the task set-up* and unrepresentative of children's everyday language?

One final consideration relates to a new emerging standard in psychology and linguistics research. Historically, child language researchers have paid little attention to the need to conduct statistical analyses by items as well as by subjects (which has long been the prevailing standard in adult psycholinguists[161]). For example, a study comparing children's performance with high-frequency and low-frequency verbs might include just three of each (e.g., *fall, laugh, disappear* vs *tumble, giggle, vanish*). The problem is that, with so few items in each cell of the design, any observed difference may reflect different performance with the specific items *fall, laugh*, and *disappear* versus *tumble, giggle* and *vanish,* rather than between high- and low-frequency verbs in general. Nobody would perform the equivalent by-subjects extrapolation (e.g., from a study comparing three boys and three girls to the entire populations of boys and girls in general), but for some reason we have not applied the same standard to items. Following the recent publication of a number of influential papers,[84,119,162,163] many leading journals have begun to require not only separate by-items and by-subjects analyses but also mixed-effects models that include both subjects and items as random effects. What this means at the study-design stage is that researchers should include as many different items as possible in each cell of the design given constraints such as children's attention span and the availability of suitable items. A drawback is that mixed effects models are not yet widely understood in the field, and debate continues as to the most appropriate analysis procedures.[164] Until this situation is resolved, we suggest that some researchers may prefer to use simpler statistical techniques (e.g., ANOVA or regression) in the main analysis, and use mixed effects models—perhaps reported in a footnote—to check that their findings generalize across participants and items.

A Final Thought

Browsing through child language articles from 30 or 40 years ago, it is striking, first, how many are based on relatively informal observations and interviews, and, second, how many experimental studies include no formal statistical analyses and few controls for potential confounds. The increasing sophistication of experimental design and analysis in the studies summarized in this review is a welcome development, and one that promises to make an important contribution to our understanding of the processes underlying children's language acquisition and development.

REFERENCES

1. Chomsky N. *Aspects of the Theory of Syntax*. Cambridge: MIT Press; 1965.
2. Matthews D, Bannard C. Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. *Cogn Sci* 2010, 34: 465–488.
3. Bannard C, Matthews DE. Stored word sequences in language learning: the effect of familarity on children's repetition of four-word combinations. *Psychol Sci* 2008, 19:241–248.
4. Naigles L. Children use syntax to learn verb meanings. *J Child Lang* 1990, 17:357–374.
5. Gertner Y, Fisher C, Eisengart J. Learning words and rules: abstract knowledge of word order in early sentence comprehension. *Psychol Sci* 2006, 17: 684–691.
6. Semel EM, Wiig EH, Secord WA. *Clinical evaluation of language fundamentals, fourth edition (CELF-4)*. Toronto, Canada: Psychological Corporation/A Harcourt Assessment Company; 2003.
7. Hresko WP, Reid DK, Hammill DD. *Test of Early Language Development*. 3rd ed. TX: Pro-Austin; 1999.
8. Dunn LM, Dunn DM. *Picture Vocabulary Test*, (PPVT™-4). 4th ed. Peabody: Pearson; 2007.
9. Wells G. *Learning through Interaction: The Study of Language Development*. Cambridge: Cambridge University Press; 1981.
10. Theakston AL, Lieven EV, Pine JM, Rowland CF. The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *J Child Lang* 2001, 28:127–152.
11. Gathercole VM, Sebastian E, Soto P. The early acquisition of spanish verbal morphology: across-the-board or piecemeal knowledge? *Int J Bilingual* 1999, 2 and 3:133–182.
12. Vosoughi S, Roy DK. A longitudinal study of prosodic exaggeration in child-directed speech. *Proceedings of the 6th International Conference of Speech Prosody, Shanghai, China*; 2012.
13. Valian V, Solt S, Stewart J. Abstract categories or limited-scope formulae? The case of children's determiners. *J Child Lang* 2009, 36:743–778.
14. Yang C, Who's Afraid of George Kingsley Zipf? Unpublished manuscript.
15. Pine JM, Lieven EVM. Slot and frame patterns and the development of the determiner category. *Appl Psycholing* 1997, 18:123–138.
16. Pine JM, Freudenthal D Krajewski G. Do children have adult-like syntactic categories? The case of the determiner. *Cognition*. Submitted for publication.
17. Rowland CF, Fletcher SL. The effect of sampling on estimates of lexical specificity and error rates. *J Child Lang* 2006, 33:859–877.
18. Aguado Orea JJ. The acquisition of morpho-syntax in Spanish: implications for current theories of development, Theses P, University of Nottingham, 2004.
19. Pine JM, Rowland CF, Lieven EVM, Theakston AL. Testing the agreement/tense omission model: why the data on children's use of non-nominative 3psg subjects count against the ATOM. *J Child Lang* 2005, 32:269–289.
20. Wexler K. Very early parameter setting and the unique checking constraint: a new explanation of the optional infinitive stage. *Lingua* 1998, 106:23–79.
21. Rowland C, Pine JM, Lieven EVM, Theakston AL. The incidence of error in young children's wh-questions. *J Speech Lang Hearing Res* 2005, 48:384–404.
22. Tomasello M. *First Verbs: A Case Study of Early Grammatical Development*. Cambridge: Cambridge University Press; 1992.
23. Naigles LR, Hoff E, Vear D. *Flexibility in Early Verb Use: Evidence from a Multiple-N Diary Study*. Boston, Mass., Oxford: Wiley-Blackwell; 2009.
24. Akhtar N, Tomasello M. Young children's productivity with word order and verb morphology. *Dev Psychol* 1997, 33:952–965.
25. Fenson L, Marchman V, Thal D, Reznick S, Bates E. *The MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual*. 2nd ed. Baltimore, MD: Paul H. Brookes; 2007.
26. Dale PS. The validity of a parent report measure of vocabulary and syntax at 24 months. *J Speech Lang Hearning Res* 1991, 34:565–571.
27. Dale PS, Bates E, Reznick JS, Morisset C. The validity of a parent report instrument of child language at 20 months. *J Child Lang* 1989, 16:239–249.
28. Feldman HM, Dale PS, Campbell TF, Colborn DK, Kurs-Lasky M, Rockette HE, Paradise JL. Concurrent and predictive validity of parent reports of child language at ages 2 and 3 years. *Child Dev* 2005, 76:856–868.
29. Feldman HM, Dollaghan CA, Campbell TF, Kurs-Lasky M, Janosky JE, Paradise JL. Measurement properties of the MacArthur communicative development inventories at ages one and two years. *Child Dev* 2000, 71:310–322.
30. Pine JM, Lieven EVM, Rowland C. Observational and checklist measures of vocabulary composition: what do they mean? *J Child Lang* 1996, 23:573–590.
31. Brooks PJ, Tomasello M. How children constrain their argument structure constructions. *Language* 1999b, 75:720–738.
32. Brooks PJ, Tomasello M. Young children learn to produce passives with nonce verbs. *Dev Psychol* 1999a, 35:29–44.

33. Brooks PJ, Tomasello M, Dodson K, Lewis LB. Young children's overgeneralizations with fixed transitivity verbs. *Child Dev* 1999, 70:1325–1337.
34. Berko J. The child's learning of English morphology. *Word* 1958, 14:150–177.
35. Marchman VA. Children's productivity in the English past tense: The role of frequency, phonology, and neighborhood structure. *Cogn Sci* 1997, 21:283–303.
36. Ambridge B. Children's judgments of regular and irregular novel past-tense forms: new data on the English past-tense debate. *Dev Psychol* 2010, 46:1497–1504.
37. Grosse G, Tomasello M. Two-year-old children differentiate test questions from genuine questions. *J Child Lang* 2012, 39:192–204.
38. Wittek A, Tomasello M. Young children's sensitivity to listener knowledge and perceptual context in choosing referring expressions. *Appl Psycholing* 2005, 26:541–558.
39. Matthews D, Lieven E, Theakston A, Tomasello M. The effect of perceptual availability and prior discourse on young children's use of referring expressions. *Appl Psycholing* 2006, 27:403–422.
40. Matthews D, Lieven E, Theakston A, Tomasello M. Pronoun co-referencing errors: challenges for generativist and usage-based accounts. *Cogn Ling* 2009, 20:599–626.
41. Matthews D, Lieven E, Tomasello M. How toddlers and preschoolers learn to uniquely identify referents for others: a training study. *Child Dev* 2007, 78:1744–1759.
42. Rowland CF, Chang F, Ambridge B, Pine JM, Lieven EVM. The development of abstract syntax: Evidence from structural priming and the lexical boost. *Cognition* 2012, 125:49–63.
43. Valian V, Casey L. Young children's acquisition of wh-questions: the role of structured input. *J Child Lang* 2003, 30:117–143.
44. Ambridge B, Rowland CF. Predicting children's errors with negative questions: testing a schema-combination account. *Cogn Ling* 2009, 20:225–266.
45. Ambridge B, Rowland CF, Theakston AL, Tomasello M. Comparing different accounts of inversion errors in children's non-subject wh-questions: 'What experimental data can tell us?' *J Child Lang* 2006, 33: 519–557.
46. Ambridge B, Rowland CF, Pine JM. Is structure dependence an innate constraint? New experimental evidence from children's complex-question production. *Cogn Sci* 2008, 32:222–255.
47. Crain S, Nakayama M. Structure dependence in grammar formation. *Language* 1987, 63:522–543.
48. Zukowski A. Elicited production of relative clauses in children with Williams syndrome. *Lang Cogn Process* 2009, 24:1–43.
49. Thornton R. Elicited production. In: McDaniel D, McKee C, Smith-Cairns H, eds. *Methods for Assessing Children's Syntax*. Cambridge, MA: MIT Press; 1996, 77–102.
50. Olguin R, Tomasello M. 25-month-old children do not have a grammatical category of verb. *Cogn Dev* 1993, 8:245–272.
51. Ambridge B, Pine JM. Testing the agreement/tense omission model using an elicited imitation paradigm. *J Child Lang* 2006, 33:879–898.
52. Santelmann L, Berk S, Austin J, Somashekar S, Lust B. Continuity and development in the acquisition of inversion in yes/no questions: dissociating movement and inflection. *J Child Lang* 2002, 29:813–842.
53. Weeks LA. Preschoolers' production of tag questions and adherence to the polarity-contrast principle. *J Psycholing Res* 1992, 21:31–40.
54. Diessel H, Tomasello M. A new look at the acquisition of relative clauses. *Language* 2005, 81:882–906.
55. Kidd E, Brandt S, Lieven E, Tomasello M. Object relatives made easy: a cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. *Lang Cogn Process* 2007, 22:860–897.
56. Lust B, Flynn S, Foley C. What children know about what they say: elicited imitation as a research method for assessing children's syntax. In: McDaniel D, McKee C, Cairns H, eds. *Methods for Assessing Children's Syntax*. Cammbridge, MA: MIT Press; 1996.
57. Bernstein-Ratner N. Elicited imitation and other methods for the analysis of trade-offs between speech and language skills in children. In: Menn L, Bernstein-Ratner N, eds. *Methods for Studying Language Production*. Hillsdale, NJ: Lawrence Erlbaum Associates; 2000.
58. Vinther T. Elicited imitation:a brief overview. *Int J Appl Ling* 2002, 12:54–73.
59. Baddeley AD, Hitch GJ, Allen RJ. Working memory and binding in sentence recall. *J Memory Lang* 2009, 61:438–456.
60. McDade HL, Simpson MA, Lamb DE. The use of elicited imitation as a measure of expressive grammar: a question of validity. *J Speech Hearing Disord* 1982, 47:19–24.
61. Rasanen S, Ambridge B, Pine JM, Infinitives or bare stems? Are English-speaking children defaulting to the highest frequency form? *J Child Lang*. Submitted for publication.
62. Ferreira F. Effects of length and syntactic complexity on initiation times for prepared utterances. *J Mem Lang* 1991, 30:210–233.
63. Hudgins JC, Cullinan WL. Effects of sentence structure on sentence elicited imitation responses. *J Speech Hearing Res* 1978, 21:809–819.

64. Silverman SW, Ratner NB. Syntactic complexity, fluency, and accuracy of sentence imitation in adolescents. *J Speech Lang Hearing Res* 1997, 40:95–106.
65. Bock JK. Syntactic persistence in language production. *Cogn Psychol* 1986, 18:355–387.
66. Pickering MJ, Ferreira VS. Structural priming: a critical review. *Psychol Bull* 2008, 134:427–459.
67. Savage C, Lieven E, Theakston A, Tomasello M. Testing the abstractness of children's linguistic representations: lexical and structural priming of syntactic constructions in young children. *Dev Sci* 2003, 6:557–567.
68. Huttenlocher J, Vasilyeva M, Shimpi P. Syntactic priming in young children. *J Memory Lang* 2004, 50:182–195.
69. Shimpi PM, Gamez PB, Huttenlocher J, Vasilyeva M. Syntactic priming in 3-and 4-year-old children: evidence for abstract representations of transitive and dative forms. *Dev Psychol* 2007, 43:1334–1346.
70. Bencini GML, Valian VV. Abstract sentence representations in 3 year-olds: evidence from language production and comprehension. *J Mem Lang* 2008, 59:97–113.
71. Messenger KM, Branigan HP, McLean JF. Evidence for (shared) abstract structure underlying children's short and full passives. *Cognition* 2011, 121:268–274.
72. Messenger KM, Branigan HP, McLean JF. Is children's acquisition of the passive a staged process? Evidence from six- and nine-year-olds' production of passives. *J Child Lang* 2012, 39:991–1016.
73. Messenger KM, Branigan HP, McLean JF, Sorace A,C. 2012. Is young children's passive syntax semantically constrained? Evidence from syntactic priming *J Memory Lang*.
74. Bernolet S, Hartsuiker RJ. Does verb bias modulate syntactic priming?. *Cognition* 2010, 114:455–461.
75. Coyle JM, Kaschak MP. Patterns of experience with verbs affect long-term cumulative structural priming. *Psych Bull Rev* 2008, 15:967–970.
76. Kaschak M, Borreggine KL. Is long-term structural priming affected by patterns of experience with individual verbs? *J Memory Lang* 2008, 58:862–878.
77. Jaeger TF, Snider N. Implicit learning and syntactic persistence: Surprisal and cumulativity. In: Wolter L, Thorson J, eds. *University of Rochester Working Papers in the Language Sciences*. 2007, 3:26–44.
78. Akhtar N. Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *J Child Lang* 1999, 26:339–356.
79. Tomasello M. *First Verbs: A Case Study of Early Grammatical Development*. New York: Cambridge University Press; 1992.
80. Matthews D, Lieven E, Theakston A, Tomasello M. The role of frequency in the acquisition of English word order. *Cogn Develop* 2005, 20:121–136.
81. Matthews D, Lieven E, Theakston A, Tomasello M. French children's use and correction of weird word orders: a constructivist account. *J Child Lang* 2007, 34:381–409.
82. Abbot-Smith K, Lieven E, Tomasello M. What preschool children do and do not do with ungrammatical word orders. *Cogn Dev* 2001, 16:679–692.
83. Childers JB, Tomasello M. The role of pronouns in young children's acquisition of the English transitive construction. *Dev Psychol* 2001, 37:739–748.
84. Jaeger TF. Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J Mem Lang* 2008, 59:434–446.
85. Pozzan L, Quirk E. Asking questions in L2 English: an elicited production study. *Applied Psycholinguistics*. Submitted for publication.
86. Hartshorne JK, Ullman M. Why girls say ''holded'' more than boys. *Dev Sci* 2006, 9:21–32.
87. Ramscar M. The role of meaning in inflection: why the past tense does not require a rule. *Cogn Psychol* 2002, 45:45–94.
88. Tomasello M. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press; 2003.
89. Albright A, Hayes B. Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition* 2003, 90:119–161.
90. Prasada S, Pinker S. Generalization of regular and irregular morphological patterns. *Lang Cogn Process* 1993, 8:1–56.
91. Ambridge B, Pine JM, Rowland CF, Chang F. The roles of verb semantics, entrenchment, and morphophonology in the retreat from dative argument-structure overgeneralization errors. *Language* 2012, 88:45–81.
92. Kidd E, Bavin EL, Rhodes B. Two-year olds' knowledge of verbs and argument structure. In: Almgren M, ed. *Research on Child Language Acquisition*. Somerville, MA: Cascadilla Press; 2001.
93. Marchman VA, Wulfeck B, Weismer SE. Morphological productivity in children with normal language and SLI: a study of the English past tense. *J Speech Lang Hearing Res* 1999, 42:206–219.
94. Ambridge B, Pine JM, Rowland CF, Bidgood A. Passive resistance: why can't some verbs be (easily) passized? Unpublished manuscript.
95. Arnon I. Rethinking child difficulty: the effect of NP type on children's processing of relative clauses in Hebrew. *J Child Lang* 2010, 37:27–57.
96. Dittmar M, Abbot-Smith K, Lieven E, Tomasello M. German children's comprehension of word order and case marking in causative sentences. *Child Dev* 2008, 79:1152–1167.
97. Chan A, Lieven E, Tomasello M. Children's understanding of the agent-patient relations in the transitive construction: cross-linguistic comparisons between

Cantonese, German, and English. *Cogn Ling* 2009, 20:267–300.

98. MacWhinney B, Bates E. *The Cross-Linguistic Study of Sentence Processing*. New York: Cambridge University Press; 1989.
99. Grunloh T, Lieven EVM, Tomasello M. German Children's Use of Prosodic Cues in Resolving Participant Roles in Transitive Constructions. *Poster Presented at the Boston University Conference on Language Development* 34, Boston MA, 2009.
100. Naigles L, Fowler A, Helm A. Developmental shifts in the construction of verb meanings. *Cogn Dev* 1992, 7:403–427.
101. Naigles L, Gleitman H, Gleitman LR. Syntactic bootstrapping and verb acquisition. In: Dromi E, ed. *Language and Cognition: A Developmental Perspective*. NJ: Ablex; 1993.
102. Hamburger H, Crain S. Relative acquisition. In: Kuczaj SA, ed. *Language Development: Syntax and Semantics*. Hillsdale, NJ: Erlbaum; 1982.
103. Correa A. An alternative assessment of children's comprehension of relative clauses. *J Psycholing Res* 1995, 24:183–203.
104. Horgan D. Development of full passive. *J Child Lang* 1978, 5:65–80.
105. Sudhalter V, Braine MD. How does comprehension of passive develop? A comparison of actional and experiential verbs. *J Child Lang* 1985, 12:455–470.
106. Maratsos M, Fox DE, Becker JA, Chalkley MA. Semantic restrictions on children's passives. *Cognition* 1985, 19:167–191.
107. Brooks PJ, Sekerina IA. Shortcuts to quantifer interpretation in children and adults. *Lang Acquisit* 2006, 13:177–206.
108. Street JA, Dabrowska E. More individual differences in language attainment: how much do adult native speakers of English know about passives and quantifiers? *Lingua* 2010, 120:2080–2094.
109. Rowland CF, Noble C. Knowledge of verb argument structure in early sentence comprehension: evidence from the dative. *Lang Learn Dev* 2011, 7:55–75.
110. Noble CH, Rowland CF, Pine JM. Comprehension of argument structure and semantic roles: evidence from english-learning children and the forced-choice pointing paradigm. *Cogn Sci* 2011, 35:963–982.
111. Fisher C. Structural limits on verb mapping: the role of analogy in children's interpretations of sentences. *Cogn Psychol* 1996, 31:41–81.
112. Casenhiser D, Goldberg AE. Fast mapping between a phrasal form and meaning. *Dev Sci* 2005, 8:500–508.
113. Boyd JK, Gottschalk E, Goldberg AE. Linking rule acquisition in novel phrasal constructions. *Lang Learn* 2009, 93:418–429.
114. Wonnacott E, Boyd JK, Thomson J, Goldberg AE. Input effects on the acquisition of a novel phrasal construction in 5 year olds. *J Mem Lang* 2012, 66:458–478.
115. Chan A, Meints K, Lieven E, Tomasello M. Young children's comprehension of English SVO word order revisited: testing the same children in act-out and intermodal preferential looking tasks. *Cogn Dev* 2010, 25:30–45.
116. Dittmar M, Abbot-Smith K, Lieven E, Tomasello M. Young German children's early syntactic competence: a preferential looking study. *Dev Sci* 2008b, 11:575–582.
117. Smith L, Yu C. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition* 2008, 106:1558–1568.
118. Tincoff R, Jusczyk PW. Some beginnings of word comprehension in 6-month-olds. *Psychol Sci* 1999, 10:172–175.
119. Barr DJ. Analyzing 'visual world' eyetracking data using multilevel logistic regression. *J Memory Lang* 2008, 59:457–474.
120. Fernald A, Zangl R, Portillo AL, Marchman VA. Looking while listening: using eye movements to monitor spoken language comprehension by infants and young children. In: Sekerina IA, Fernandez EM, Clahsen H, eds., *Developmental Psycholinguistics: On-line Methods in Children's Language Processing*. Amsterdam: John Benjamins; 2008.
121. Pyykkonen P, Matthews D, Jarvikivi J. Three-year-olds are sensitive to semantic prominence during online language comprehension: a visual world study of pronoun resolution. *Lang Cogn Process* 2010, 25:115–129.
122. Thothathiri M, Snedeker J. Give and take: syntactic priming during spoken language comprehension. *Cognition* 2008, 108:51–68.
123. Thothathiri M, Snedeker J. Syntactic priming during language comprehension in three- and four-year-old children. *J Memory Lang* 2008, 58:188–213.
124. Arnold JE, Brown-Schmidt S, Trueswell JC. Children's use of gender and order-of-mention during pronoun comprehension. *Lang Cogn Process* 2007, 24:527–565.
125. Jusczyk P. Infants acquisition of the sound structure of their native language. *Int J Psychol* 1992, 27:59–59.
126. Jusczyk PW, Cutler A, Redanz NJ. Infants preference for the predominant stress patterns of English words. *Child Dev* 1993, 64:675–687.
127. Jusczyk PW, Hirsh-Pasek K, Nelson DGK, Kennedy LJ, Woodward A, Piwoz J. Perception of acoustic correlates of major phrasal units by young infants. *Cogn Psychol* 1992, 24:252–293.
128. Gerken LB, Thomas G. Linguistic Intuitions are the Result of interactions between perceptual processes and Linguistics Universals. *Cogn Sci* 1986, 10:457–476.

129. Saffran JR, Aslin RN, Newport EL. Statistical learning by 8-month-old infants. *Science* 1996, 274: 1926–1928.
130. Hunter M, Ames AEW. A multifactor model of infant preferences for novel and familiar stimuli. *Adv Infancy Res* 1988, 5:69–95.
131. Houston-Price C, Nakai S. Distinguishing novelty and familiarity effects in infant preference procedures. *Infant Child Dev* 2004, 13:341–348.
132. Soderstrom M, Seidl A, Kemler-Nelson DG, Jusczyk PW. The prosodic bootstrapping of phrases: evidence from prelinguistic infants. *J Memory Lang* 2003, 49:249–267.
133. Moon C, Cooper RP, Fifer WP. 2-Day-olds prefer their native language. *Infant Behav Dev* 1993, 16:495–500.
134. Hepper PG, Scott D, Shahidullah S. Newborn and fetal response to maternal voice. *J Reprod Infant Psychol* 1993, 11:147–153.
135. Mehler J, Jusczyk P, Lambertz G, Halsted N, Bertoncini J, Amiel-Tison C. A precursor of language acquisition in young infants. *Cognition* 1988, 29:143–178.
136. Gerken L, Jusczyk PW, Mandel DR. When prosody fails to cue syntactic structure: 9-month-olds' sensitivity to phonological versus syntactic phrases. *Cognition* 1994, 51:237–265.
137. Dehaene-Lambertz G, Dehaene S, Hertz-Pannier L. Functional neuroimaging of speech perception in infants. *Science* 2002, 298:2013–2015.
138. Dehaene-Lambertz G, Hertz-Pannier L, Dubois J. Nature and nurture in language acquisition: anatomical and functional brain-imaging studies in infants. *Trends Neurosci* 2006, 29:367–373.
139. Pena M, Maki A, Kovacic D, Dehaene-Lambertz G, Koizumi H, Bouquet F, Mehler J. Sounds and silence: an optical topography study of language recognition at birth. *Proc Natl Acad Sci U S A* 2003, 100:11702–11705.
140. Brett M, Johnsrude IS, Owen AM. The problem of functional localization in the human brain. *Nature Rev Neurosci* 2002, 3:243–249.
141. Friederici AD. The neural basis of language development and its impairment. *Neuron* 2006, 52:941–952.
142. Friedrich M, Friederici AD. Lexical priming and semantic integration reflected in the event-related potential of 14-month-olds. *Neuroreport* 2005, 16:653–656.
143. Silva-Pereyra J, Rivera-Gaxiola M, Kuhl PK. An event-related brain potential study of sentence comprehension in preschoolers: semantic and morphosyntactic processing. *Cogn Brain Res* 2005, 23: 247–258.
144. Silva-Pereyra J, Klarman L, Lin LJF, Kuhl PK. Sentence processing in 30-month-old children: an event-related potential study. *Neuroreport* 2005, 16:645–648.
145. Fernandes KJ, Marcus GF, Di Nubila JA, Vouloumanos A. From semantics to syntax and back again: argument structure in the third year of life. *Cognition* 2006, 100:B10–B20.
146. Braine MDS, Brooks PJ. Verb argument strucure and the problem of avoiding an overgeneral grammar. In: Tomasello M, Merriman WE, eds. *Beyond Names for Things: Young Children's Acquisition of Verbs*. Hillsdale, NJ: Erlbaum; 1995, 352–376.
147. Theakston AL. The role of entrenchment in children's and adults' performance on grammaticality judgement tasks. *Cogn Dev* 2004, 19:15–34.
148. Ambridge B, Pine JM, Rowland CF, Young CR. The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition* 2008, 106:87–129.
149. Ambridge B. Paradigms for assessing children's knowledge of syntax and morphology. In: Hoff E, ed. *Guide to Research Methods in Child Language*. London: Blackwell-Wiley; 2011, 113–132.
150. Rice ML, Wexler K, Redmond SM. Grammaticality judgments of an extended optional infinitive grammar: Evidence from English-speaking children with specific language impairment. *J Speech Lang Hearing Res* 1999, 42:943–961.
151. de Villiers PA, Hosler B, Miller K, Whalen M, Wong J. Language, Theory of Mind, and Reading Other People's Emotions: A Study of Oral Deaf Children. 1997.
152. Ambridge B., How do children restrict their linguistic generalizations?: An (un-) grammaticality judgment study. *Cogn Sci*. In press.
153. Ambridge B, Pine JM, Rowland CF. Semantics versus statistics in the retreat from locative overgeneralization errors. *Cognition* 2012, 123:260–279.
154. Ambridge B, Pine JM, Rowland CF, Jones RL, Clark V. A semantics-based approach to the "no negative evidence" problem. *Cogn Sci* 2009, 33:1301–1316.
155. Ambridge B, Pine JM, Rowland CF. Children use verb semantics to retreat from overgeneralization errors: a novel verb grammaticality judgment study. *Cogn Ling* 2011, 22:303–323.
156. Chien Y, Wexler K. Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Lang Acquisit* 1990, 1:225–295.
157. McKee C. A comparison of pronouns and anaphors in Italian and English. *Lang Acquisit* 1992, 1:21–55.
158. Lust B, Eisele J, Mazuka R. The binding theory module: evidence from first language acquisition for Principle-C. *Language* 1992, 68:333–358.
159. Crain S, Thornton R, Boster C, Conway L, Lillo-Martin D, Woodams E. Quantification without qualification. *Lang Acquisit* 1996, 5:83–153.

160. Katsos N, Bishop DVM. Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition* 2011, 120:67–81.

161. Clark HH. The Language-as-fixed-effect falacy: A crititque of language statistics in psychological research. *J Verbal Learn Verbal Behav* 1973, 12: 335–359.

162. Baayen H, Davidson DJ, Bates DM. Mixed-effects modelling with crossed random effects for subjects and items. *J Memory Lang* 2008, 59: 390–419.

163. Quene H, van den Bergh H. Examples of mixed-effects modeling with crossed random effects and with binomial data. *J Memory Lang* 2008, 59: 413–425.

164. Barr DJ, Levy R, Scheepers C, Tily HJ, Random effects structure in mixed-effects models: keep it maximal. Submitted for publication.

FURTHER READING

All but a handfull of the studies cited in the present review are discussed in the following textbook, which also contains a brief summary of all the methods covered here (p. 6–12).

Ambridge B, Lieven EVM. *Child Language Acquisition: Contrasting Theoretical Approaches*. Cambridge: Cambridge University Press; 2011.

The following an excellent resource for methods in child language acquisition research: Hoff E. *Guide to Research Methods in Child Language*. London: Blackwell-Wiley; 2012.