Oral language elicitation tasks in applied linguistics research

Faidra Faitaki, Victoria A. Murphy
Published online on: 02 Jan 2020


Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Oral language elicitation tasks in applied linguistics research

Faidra Faitaki and Victoria A. Murphy

Introduction
We can only ever infer competence from performance (Chomsky, 1965). In carrying out research investigating a learner’s language production skills, we always have to measure participants’ verbal behaviour (cf. Skinner, 1957). The nature of that behaviour (or performance) can vary significantly, ranging from the language used in a naturalistic conversation to performance on a timed lexical decision task carried out on a computer. Given most applied linguists are behaviourists (of a kind) and all research findings are methodological artefacts, it behoves us to understand as much as is possible the relative strengths and weaknesses of different methodologies that lead us to our inferences and conclusions. The focus of this chapter is to delve into some of the issues surrounding one specific methodology: namely, language elicitation tasks. The focus of this chapter is to delve into some of the issues surrounding one specific methodology: language elicitation tasks. These are crucial and oft-used tasks in applied linguistics research, utilised whenever a researcher attempts to measure a participant’s linguistic production (be it oral or written). Our working definition then of language elicitation tasks is simply that – a task that requires a participant to produce some form of language, in either oral (speaking) or written form. However, we focus on oral tasks in this chapter.

Language elicitation tasks come in many forms. On the one hand, they can be very loosely structured, such as recording and analysing naturalistic speech produced by a learner. In these tasks, researchers typically analyse participants’ speech based on the pre-determined characteristics associated with the focus of the study: for example, pauses and self-repairs if measuring fluency (e.g., Segalowitz, 2010), or the number of inflectional morphemes if the focus is on aspects of morphosyntax (Brown, 1973) and so on. Language elicitation tasks can, on the other hand, be highly constrained, such as computer-based naming tasks where words must be read out loud as quickly and accurately as possible. In these tasks, the researcher typically measures both reaction time and accuracy and is often interested in differences in processing speed across different items (e.g. Murphy & Hayes, 2010).

Strengths and weaknesses of language elicitation tasks have been discussed at length (e.g. Menn & Bernstein Ratner, 2000; Miller, 1981; Underhill, 1987). Summaries about what we have learned as a field by using these tasks are also readily available (e.g. Butterworth,
Oral language elicitation tasks

1980; Kormos, 2006; Xavier-Alario, Costa, Ferreira, & Pickering, 2006). In this chapter we begin by examining the relative advantages and disadvantages of using standardised versus non-standardised (researcher-made) tasks aimed at eliciting speech. Both have their uses, but similarly can be limited and be used ineffectually and inappropriately. We then turn to a few key non-standardised language production tasks and discuss their uses.

**On standardised tasks**

Standardised tasks are those that are typically norm-referenced against a given population, and importantly have undergone rigorous scrutiny regarding their reliability and validity. At its most basic level, a standardised test is one where every test-taker answers the questions of the test under the same conditions, and of course, answers the same questions. Similarly, the way the responses on the test are coded and graded follows a previously determined procedure. A standardised test typically also has been norm-referenced against a large population of respondents, enabling us to understand a given individual’s score in reference to that larger population. A typical example might be a standardised test which measures reading ability in primary school students. Standardised tests such as the York Assessment of Reading Comprehension allow us to identify whether a child’s reading skill is age-appropriate (Stothard, Hulme, Clarke, Barmby, & Snowling, 2010) and many such tests offer a ‘reading age’ which highlights the comparison of an individual test taker against the population against which it has been referenced. Due to processes and procedures involved in norm-referencing a test against a population, we also tend to know quite a bit about a given standardised test’s reliability and validity. This information enables researchers to make informed decisions about whether the test they are considering is appropriate for their use.

We have comparatively few standardised tests of language production for L2 learners or bilingual participants. The Test of English for Foreign Language and International English Language Testing System are primary examples of oft-used high-stakes tests which include a productive component for spoken language; for child participants, there is the Cambridge Young Learners test. Apart from these we generally have very few standardised and norm-referenced tests which measure linguistic production skills in L2 speakers.

Given the general lack of standardised assessments in this domain, tests which have been norm-referenced against a largely monolingual population are often used to measure skills in a bilingual sample. For example, the York Assessment of Reading Comprehension (Stothard et al., 2010) has been used by a number of researchers to measure reading skills in children with English as an additional language (e.g. Smith & Murphy, 2015). Often researchers opt for this route because the standardised test offers information about how reliable and valid the test is at measuring the construct under investigation. However, if the test has not been norm-referenced with a bilingual population, age norms are no longer relevant, and the reliability and validity metrics may no longer hold. Furthermore, when monolingual norm-referenced tests are used with a bilingual population, it is no surprise that the bilingual sample under-performs relative to the monolingual benchmarks (e.g. Burgoyne, Whiteley, & Hutchinson, 2013). In other words, using such standardised tests disadvantages the bilingual or L2 learner. There is a delicate balance to be struck, therefore, between using a well-known test that has clear reliability and validity metrics but which has been norm-referenced and validated on a different (non-bilingual) sample. Without more standardised tests that offer us measures of language elicitation in L2 and bilingual samples, we are left with making a choice between standardised tests that may not have been developed with L2 speakers in mind or researcher-developed tests.

Another issue related to standardised tests in the context of research is how sensitive they are to small, incremental changes in knowledge or performance. Standardised tests often are
not sensitive to small changes in development. This issue is illustrated clearly in educational intervention studies where key skills are typically measured at baseline (pre-test) and then again at post-test, following from some form of intervention. Researchers typically use both standardised and non-standardised assessments in their intervention studies, and interestingly, benefits of intervention are often found with non-standardised assessment but more rarely so with standardised assessments. Fricke and Millard (2016), for example, carried out an oral language intervention study on young children with English as an additional language and found positive effects of their intervention on the researcher-made target vocabulary test, but not on any of the standardised tests that were administered at post-test. The lack of sensitivity of standardised tests to small growth has also been clearly demonstrated in Rogde (2017) who carried out a meta-analysis of intervention studies aimed at improving children’s linguistic comprehension and found that typically the effect of interventions could be observed on researcher-developed measures but were less likely to be manifest on standardised tests. Therefore, while we might know much about a standardised test’s reliability and validity statistics, if we want to use the test to measure growth or change we might best use a researcher-developed test.

On researcher-developed measures

Numerous measures have been used within applied linguistics research to measure language production. We have chosen a small handful here to demonstrate some of the issues which we believe are important to consider in deciding which task to use. The tasks we have chosen range from the less to the more controlled and prescriptive, each of which have their own advantages and disadvantages.

Spontaneous speech data analysis

The use of spontaneous speech data became widespread following the publication of seminal longitudinal studies on children’s L1 acquisition (e.g. Braine, 1976; Brown, 1973) and such data have also been valuable in bilingual language (2L1) acquisition research. It is through the analysis of naturalistic speech that researchers were able to explore bilingual children’s 2L1 development (e.g. Dopke, 1998); debate the nature and organisation of the bilingual brain (Genesee, Nicoladis, & Paradis, 1995; Volterra & Taeschner, 1978); and decipher whether, how, and why children’s two languages interact during 2L1 acquisition (Hulk & Müller, 2000; Paradis & Genesee, 1996; Paradis & Navarro, 2003; Serratrice, Sorace, & Paoli, 2004; Yip & Matthews, 2000).

The use of spontaneous speech data, however, is less widespread in child L2 acquisition research (but see the CHILDES database for a few, small corpora) as children tend to learn their L2(s) at school and many child L2 acquisition studies tend to be conducted within classroom contexts. In turn, classroom contexts are irregular and variable, which hinders the collection of naturalistic data (Dimroth, Rast, Starren, & Watorek, 2013). Nonetheless, a recent project entitled ‘Varieties of Initial Learners in Language Acquisition’ (VILLA) attempts to deal with this problem. VILLA involves the collection of video recordings from classrooms in which groups of ten-year-old children with five different L1s (French, Italian, German, Dutch, and English) receive 14 hours of exposure to an L2 they have no prior experience with, Polish, by a trained L1 Polish instructor. Through the analysis of the video recordings, the researchers can capture different aspects of child L2 acquisition (e.g. for morphology, Latos, 2014; Watorek, Durand, & Starosciak, 2016; for phonology, Shoemaker, 2014). This design makes VILLA the first project to systematically investigate L2 acquisition under controlled input conditions, neutralising the differences between varied classroom contexts.
The use of spontaneous speech data has been valuable across many areas of research in applied linguistics. The usefulness of the method arises from two factors: its cost-effectiveness for researchers and its naturalness for participants. Spontaneous speech data – which often derives from longitudinal investigations of small numbers of participants – is substantial but also rich in content and it can be (re)analysed for different linguistic and developmental aspects over the years. Moreover, participants in spontaneous speech studies are never asked to produce target linguistic structures or comment on aspects of their linguistic development; rather, they are recorded whilst being engaged in conversation or play with their caretakers or the researchers. This inobtrusive and undemanding method allows for participants’ speech to be natural, giving rise to data that accurately reflects their use of language.

Nevertheless, such approaches to language elicitation have some limitations. First, most spontaneous speech studies are not reproducible, as the data are specific to the context (where context is defined as a constellation of factors including the participants, the linguistic combination, and the environment) in which data collection takes place. The context sensitivity of the data, combined with the fact that they are typically drawn from small samples, makes spontaneous speech studies difficult to replicate. Second, spontaneous speech data reflects participants’ unbiased use of language; thus, the data are bound to be uninformative for researchers who wish to investigate the acquisition of infrequent or imperceptible linguistic structures. If the structure under investigation is rare (e.g. cleft sentences, like ‘it is the girl who sings the song’) it might never appear during the recording session(s). Likewise, if the structure is too subtle to perceive (e.g. third person singular -s, like ‘she sings a song’), it might go unnoticed by the researcher. Again, a balance needs to be found between attempting to characterise the language a learner might use in spontaneous conversations against specificity and precision of measurement of key linguistic features of interest.

Experimental language production tasks

With the increased frequency of cross-sectional studies in applied linguistics research, the need for quantification, reproducibility, and replicability foregrounded the use of experimental language production methods. These methods give rise to carefully controlled experiments and statistically analysable data. Furthermore, according to Ambridge and Rowland’s (2013) thorough overview, experimental language production methods form a continuum from most to least structured (that is, controlled by the researcher). We focus our discussion of experimental language production methods on three tasks that capture this continuum: narration (in which output is not controlled), question-and-answer (which allow the researcher some control over output), and sentence-completion (which allow the researcher high control).

Narration

In narration tasks, the participants are shown a visual stimulus, such as a wordless picture book, and are asked to describe it. In this task the researcher exerts limited control, as participants can communicate the content of the story in their own words. The simplicity of narration tasks makes them good candidates for testing different aspects of linguistic development. In the context of L1 acquisition, narration tasks have been used to investigate the acquisition of morphology, syntax, semantics, and pragmatics (e.g. Bamberg & Reilly, 1996; Berman & Slobin, 1994; Chen & Pan, 2009; Colozzo & Whitely, 2014; Mimeau, Plourde, Ouellet, & Dionne, 2015; Wong & Johnston, 2004). Narration tasks have also been used to investigate (extra)linguistic aspects of 2L1 and L2 development (e.g. Hoang, Nicoladis, Smithson, & Furman, 2016;
Larrañaga & Guijarro-Fuentes, 2013; Pinto, 2013). For example, in Hoang et al. (2016), a narration task was used to measure the use (and shift) of tense in 12 French-English bilingual children between eight and nine years of age. In this study, children watched two videos and were asked to describe what happened. Bilingual children were tested twice, once in French and once in English, and their performance was compared to that of two age-matched monolingual control groups. Following testing, all narratives were coded for the use of different tenses (e.g. past, present) at different, discourse-relevant points (e.g. throughout the narration, at the climax of the story). While monolingual children used the present tense only 10% of the time, bilingual children tended to produce their entire narratives in the present tense. However, both groups seemed to disregard the discourse function of shifting tenses at the key points of the two stories. Despite its methodological caveats (i.e. a small sample size overall and per group and the fact that bilingual participants watched and narrated the same videos twice), the study is a good example of a narration task can provide rich and varied data in the context of 2L1 acquisition research.

A further advantage of narration tasks are their efficiency and simplicity for young participants. As children are accustomed to hearing and narrating stories with their parents at home and their teachers at school, they are familiar with a narration task’s instructions and able to perform in it without hesitation. Thus, the task can be used with young children, as well as with older children and adults. Moreover, narration tasks are natural and able to show participant’s real use of language.

A feature that the researcher has to manipulate effectively in narration tasks is communicative sense or authenticity. Narratives should be directed to an interlocutor who has not seen (or at least pretend to not have seen) the visual stimulus before, otherwise participants might feel confused and/or unwilling to narrate the story in detail. Furthermore, in the context of 2L1 and L2 acquisition, participants ideally would not be exposed to the same visual stimulus in both their languages; having encountered the stimulus before, they would be likely to produce similar structures for the languages and perhaps perform better on the second round of testing. A simple way of dealing with this issue without creating two different sets of materials for the two languages, is to counterbalance the order in which each language is tested: half of the children could be tested in their L1 first, while the other half could be tested in their L2. It would then be preferable to leave a gap between the first and the second testing session so as to prevent children from remembering the stimulus (and their description of it) from their first exposure.

Furthermore, to get useful data, the researcher must ensure that the visual stimulus creates sufficient obligatory contexts in which the linguistic structure(s) under investigation can be produced; in other words, the stimulus must be valid and reliable. Yet even if the stimulus is well-chosen, narration data face the issue that participants might avoid using a less frequent and complex target linguistic structure in favour of a more frequent and simple structure that has the same function. Thus, narration data are best used for the investigation of structures which are ubiquitous and alternate systematically in natural speech, such as tenses and referential terms, as these structures are bound to appear in their various forms over the course of the task.

**Question and answer**

Like narration tasks, question-and-answer tasks involve participants being exposed to a stimulus before being asked to answer a series of questions about it. These questions can be either open or restraining, but in either case they should guide the respondent to produce the linguistic structure(s) under investigation. In L1 acquisition research, question-and-answer tasks have been used to investigate word-level features (e.g. Matthews, Lieven, Theakston, & Tomasello,
Oral language elicitation tasks

2006; Räsänen, Ambridge, & Pine, 2016) as well as sentence-level structures (e.g. Matthews, Lieven, Theakston, & Tomasello, 2005; Rowland, Chang, Ambridge, Pine, & Lieven, 2012). Question-and-answer tasks have also been extensively used in 2L1 and L2 acquisition studies (e.g. Argyri & Sorace, 2007; Blom, Chondrogianni, Marinis, & Vasić, 2016; Daskalaki, Chondrogianni, Blom, Argyri, & Paradis, 2019; Hervé, Serratrice, & Corley, 2016; Nicoladis, 2012; Nicoladis, Rose, & Foursa-Stevenson, 2010). Building on previous research regarding the use of referential expressions, Argyri and Sorace (2007) used a question-and-answer paradigm to test 32 eight-year-old English-Greek bilingual children. Participants saw pictures with cartoon characters who were engaged in various actions and answered questions like ‘what did Maria do to the dog?’ or ‘what doesn’t the grandma remember?’ Analysing children’s answers, the authors expected to find whether children’s use of the target structures (null and overt subjects in preverbal and postverbal position, and wh-embedded interrogatives) in Greek was influenced by their English and there was some influence of English on Greek as far as both structures were concerned.

It should be noted that the authors also used an acceptability judgement task to test whether their participants were sensitive to (un)grammatical uses of the target structures. In their study, the acceptability judgement task, which is a comprehension measure, was administered after the elicited production tasks in order to avoid influencing children. In their analysis, the acceptability judgement task complemented their elicited production data. This study is a good example of using comprehension and production data in tandem to provide a more accurate picture of children’s linguistic competence.

As this study illustrates, question-and-answer tasks can be used to test broad and simple structures like referential expressions but also rare and complex ones like wh-embedded interrogatives. Like narration tasks, question-and-answer tasks are straightforward and engaging even for younger participants. Thus, they can be used with children of (almost) all ages, provided that the level of word knowledge and world knowledge that the task involves is compatible with the children’s ages. Of course, they can also be used with adult participants. The task’s conceptual and linguistic level is modulated through the question-prompt: a simple question that urges the respondent to describe the depicted characters or events (e.g. ‘what happened?’ or ‘who did it?’) might be more suitable for young children, while a complex question that requires the participant to retain or integrate (extra)linguistic information to provide an answer (e.g. ‘how did it happen?’ ‘who said they did it?’) might work only with older children or adults.

Another important point is that, like all language elicitation methods, question-and-answer tasks should as natural as possible. Participants are sensitive to what their interlocutors see and say and, therefore, they should not be asked to describe a stimulus to an interlocutor who has ostensibly seen it (unless this manipulation forms part of the design). In addition, to ensure that participants’ answers are not due to chance and that their performance reflects their linguistic competence, researchers must include control and filler items. When dispersed among the stimuli of the task, these items should allow researchers to identify baseline performance and whether the participants are focused on the task. Moreover, the stimuli (including control and filler items where appropriate) should be presented in a counterbalanced order, so as to ensure that participants’ performance is not due to a cumulative effect of the task.

At the same time, the addition of control and filler items to the stimuli set will inadvertently increase the duration of the procedure. Longer protocols are difficult to use with young children, who tend to lose concentration even in short tasks. A means of counteracting this problem is to break the task into parts and allow children small breaks in between the parts. If children are rested (or sufficiently distracted), they will be more motivated to complete the entire experimental paradigm. Another problem with question-and-answer tasks is that respondents might be unable or unwilling to produce the target response. Inability to provide
an answer might indicate insufficient knowledge of the target structure, or that the task’s linguistic and cognitive demands are unattainable by the participants. Rather, unwillingness to give an answer tends to be due to factors that are beyond the researchers’ control.

**Sentence completion**

Should researchers want to decrease the potential number of non-target responses, they can maximise their control over the participant’s production. To do so, they can use a task that is more prescriptive in terms of the responses that might be given. A sentence-completion task is one such example. This type of task involves participants being exposed to a stimulus, listening to part of a sentence that describes it, and then being asked to complete the sentence in their own words. This paradigm has been used in numerous L1 acquisition studies (e.g. Berko, 1958; Kirjavainen, Lieven, & Theakston, 2017; Marchman, 1997; Räsänen, Ambridge, & Pine, 2016).

Given that sentence-completion tasks (and other tasks like them) force the respondent to produce lexically restrained utterances, these tasks are more suitable for testing the acquisition of morphological or syntactic structures that appear in the middle or the end of the sentence (such as verbal inflections and clause-markers) rather than sentence-initial structures (like subjects) or whole-sentence properties (like word order patterns) – although, notably, a careful design could allow for the latter structures to be tested as well. Sentence-completion tasks are ideal for testing rare and complex structures since respondents are more likely to produce structures that they might otherwise avoid in other circumstances.

Despite their advantages, designing sentence-completion tasks (and similar prescriptive tasks) involves additional complications. First – and as mentioned earlier for question-and-answer tasks – sentence-completion tasks should involve control and filler items presented alongside the experimental items in a randomised and counterbalanced order. These manipulations are necessary to ensure that participants do not overperform due to familiarisation with the structure or the procedure as a whole. However, adding in filler items risks making the procedure longer and less engaging. Second, and related, sentence-completion tasks tend to lack communicative sense, as respondents typically have to look at the picture stimuli together with the experimenter, who is also responsible for uttering the first part of the sentence. Thus, the procedure might come across as a test, which is bound to affect participants’ performance in the task.

In summary, there are numerous language-elicitation tasks that have been profitably used in the literature which fall on a continuum regarding how prescriptive and constraining the task is in shaping the language produced. Advantages of these tasks (and all those many tasks that fall on this continuum that were not discussed here) include the control offered to the researcher in capturing the linguistic phenomenon of interest, and the greater likelihood that these tasks will be sensitive to changes in participants’ development. If it is a task that has been used in past research, it will be more likely that comparative analyses are possibly drawing direct connections across different research studies. Disadvantages include a comparative lack of reliability and validity statistics (in relation to standardised tasks). Ideally researchers would consider carefully how well the task maps on to the research questions and use tasks that have been used already in the literature to help mitigate against this difficulty.

**Concluding remarks**

There are many choices to be made in carrying out research, and research on language production is no different. Standardised tasks are useful because they offer us tried and tested assessments for which there is known reliability and validity statistics. They are problematic in that
they can be insensitive to small changes in linguistic growth, and when used on bilingual or L2 samples, may not be norm-referenced against a similar population. This makes their standardised scores somewhat irrelevant and uninformative (though their raw scores can still be used), and indeed, using such a test that was not developed for these different populations could be construed as inappropriate.

An alternative then is a researcher-developed task, of which there are many, and which range considerably in terms of types of linguistic feature they can measure and how they might be administered. These can be less controlling in terms of the language produced by the participant, or they can be highly constrained, forcing the participant to use very specific linguistic features. Less controlling language elicitation tasks (like spontaneous speech and narration) are natural for learners and fairly straightforward to administer. However, highly constrained, methodologically rigorous language elicitation tasks (like question-and-answer and sentence-completion) are sensitive to small changes in development and more likely to tap in to the key features of speech that are of interest to the researcher. Yet researcher-developed tasks tend to not have formal reliability and validity indices; though if they have been used appropriately in previous studies, we do know something about their use. Thus, ideally, researchers would not invent their own tasks but rather, choose a task that has already been used in the literature. When choosing a researcher-developed language elicitation task, researchers should consider the age and linguistic background of the participants, and the structure under investigation. Less restrained tasks are easier for younger (monolingual and bilingual) participants but might be unable to shed light on learners’ knowledge of rare and/or more complex structures. However, highly restrained tasks are more difficult to administer with young (bilingual) participants but can be highly informative regarding the use of specific, infrequent, and difficult features of speech.

Knowing what a learner can do in terms of language production is a fundamental interest in applied linguistics generally and can also have important implications and applications across contexts (such as educational settings). One of the hallmarks of methodological rigour is ensuring that the measures we use tap in to the construct under investigation, produce appropriate responses that yield to analyses, and move our understanding forward. Thus, researchers should ensure that their tasks have communicative sense, meet the learners’ linguistic and cognitive level, and provide sufficient contexts for the learners to produce the target linguistic structures. Thinking carefully about this choice of task in conjunction with careful piloting can go a long way towards achieving our desired outcome of measuring learners’ verbal behaviour accurately, thus leading to informed inferences about their underlying linguistic competence.

Notes
1 In many sentence-completion studies, as well as language acquisition studies in general, the experimenter does not utter the experimental stimuli live. Instead, the stimuli are pre-recorded and replayed in order during the experiment. Pre-recording the stimuli allows the researcher to be consistent with their intonation and secures the reproducibility of the procedure.
2 A good place to start is the IRIS database: www.iris-database.org/iris/app/home/index;jsessionid=C0C2F0004668OB6BFFD20193E6F4758 (Marsden, Mackey, & Plonsky, 2016). This database provides a repository of instruments, materials, stimuli, and coding and analysis tools used for research in L2 development.

References


Oral language elicitation tasks


