Verbal Reports in Instructed SLA Research
Opportunities, Challenges, and Limitations

Melissa A. Bowles

Introduction/Definitions
Verbal reports are verbalizations that learners make either while completing a task (concurrent reports, or think-alouds) or some time thereafter (retrospective reports). In instructed SLA research, both think-alouds and retrospective reports are used to provide information about how learners process the second language. Specifics of the research design determine which report is appropriate in a given study. In a standard retrospective report, which typically takes place during the debriefing session immediately following an experimental session, participants are asked to think back to the experiment they just completed and to answer a set of probe questions, ranging from general to more specific. For instance, in studies on implicit learning, the experimenter might begin by asking whether the participant noticed any patterns or irregularities in the input and then move on to more detailed questions related to the target structure (Rebuschat, 2013). On the one hand, in this type of retrospective report, the experimenter’s prompts are the only ties back to the task. On the other hand, in one particular type of retrospective report known as stimulated recall, which has been used in more than 125 published studies between 2000 and 2015 (Gass & Mackey, 2016), participants are provided with some sort of stimulus (most often an audio- or video-recording of themselves completing the task) to help them recall what they were thinking at the time of task completion.

Given that verbal reports are versatile and can be used to answer a multitude of research questions, they have been used widely by researchers of different theoretical orientations, including conversational analytic (CA) (Walters, 2007), sociocultural (Poehner, 2007; Smagorinsky, 2001; Swain, 2006), and cognitivist perspectives (Hama & Leow, 2010; Leung & Williams, 2011; Rebuschat, Hamrick, Riestenberg, Sachs, & Ziegler, 2015; Sachs & Polio, 2007). There have, however, been questions surrounding the validity of verbal reports. This chapter provides an overview of the history and use of both concurrent and retrospective verbal reports in instructed SLA research, synthesizing research addressing their validity and outlining the opportunities, challenges, and limitations associated with them. For further detail and step-by-step data collection and analysis tips that are beyond the scope of this chapter, readers are referred to Bowles (2010) for think-alouds and to Gass and Mackey (2016) for stimulated recall.
Historical Perspectives

Origins in Cognitive Psychology

Starting in the early twentieth century, verbal reports were used in cognitive psychology, most extensively to investigate problem-solving strategies on non-verbal tasks and puzzles. By the 1980s, their use was so widespread that “both concurrent and retrospective verbal reports [had become] generally recognized as major sources of data on subjects’ cognitive processes in specific tasks” (Ericsson & Simon, 1993, p. xi). Their use had expanded to a host of task types in a wide array of fields.

Stimulated recall, which is a special type of retrospective report, was first used by Bloom (1953), who sought to investigate college students’ thought processes and learning outcomes in lecture classes. He argued that the contents of stimulated recalls should be more reliable and accurate than standard retrospective reports, which lack a stimulus and would be more affected by memory decay, stating, “the subject may be enabled to relive an original situation with vividness and accuracy if he is presented with a large number of the cues or stimuli which occurred during the original situation” (p. 161). The technique was further refined by Siegel, Siegel, Capretta, Jones, and Berkowitz (1963), known for first implementing video-recorded stimuli, and stimulated recall has since become widely used not only in education, where it has been used to evaluate teacher effectiveness and decision-making processes, but also in first and second language acquisition and in a variety of other fields, including conflict resolution, management, athletic coaching, counseling, and medicine.

Use in SLA

In large part due to their versatility in capturing processes and processing that would otherwise remain hidden, verbal reports have been prevalent in language research since the 1980s. First language reading and writing research have relied heavily on think-alouds, focusing on strategies used by different groups of students, such as more and less successful readers and writers. In some cases, think-alouds have been used as a component of instructional programs designed to improve students’ reading and writing ability. (See Bowles, 2010, for references.)

In L2 research, both concurrent and retrospective verbal reports have also been used to study reading and writing. Additionally, they have been used to investigate a wide array of other phenomena, including L1 and L2 strategy use, lexical organization, the use of translation in L2 learning, interlanguage pragmatics, and vocabulary acquisition through incidental reading. (See Bowles, 2010, for references.) They have also been used extensively as part of the validation process for large-scale assessments of L2 learners’ reading, speaking, and writing abilities (e.g., Wei & Losa, 2015).

Concurrent verbal reports have proven particularly useful in the attention and awareness strands of L2 research, where they have shed light on participants’ allocation of attention to or noticing of targeted forms in the input (Alanen, 1995; Leow, 2001), as well as the roles of different levels of awareness (Rosa & Leow, 2004), unawareness (Hama & Leow, 2010; Leow, 2000), different levels or depths of processing (Adrada Rafael, 2017; Leow, Hsieh & Moreno, 2008; Morgan-Short, Heil, Botero-Moriarty, & Ebert, 2012; Qi & Lapkin, 2001), and different types of processing (Leow, 1998). They therefore allow researchers a means to “operationalize and measure the roles of cognitive processes postulated to play a role in the learning process” and have the additional methodological advantage of being able to demonstrate the representativeness of participants in experimental conditions (Leow, Grey, Marijuan & Moorman, 2014, p. 114). In so doing, verbal reports increase the internal validity of any such study. Verbal reports are
logistically simple to set up and use in both classroom and laboratory settings and do not require expensive software or equipment, unlike some other concurrent data elicitation procedures, such as eye-tracking. Verbal reports also have the added advantage that they make thought processes directly observable, whereas techniques such as eye-tracking or reaction times are indirect, relying on assumptions about what different fixations, eye movement patterns, and response times mean to make claims about cognitive processes.

Since think-alouds are collected during task completion, they are thought to provide richer data than retrospective reports during the input processing and intake stages of L2 learning. However, some types of SLA research, such as oral interaction studies or other sorts of research involving speaking tasks, are not well suited to thinking aloud and are better addressed by stimulated recall, since reporting occurs after task completion.

**Critical Issues and Topics**

Despite the widespread use of verbal reports, there has been controversy surrounding their validity, which hinges on (1) whether verbalizing alters the very thought processes under investigation (reactivity) and (2) whether verbalizations are a true reflection of thoughts (veridicality). Concerns about reactivity were first raised in cognitive psychology in the 1970s (Nisbett & Wilson, 1977; Payne, Braunstein, & Carroll, 1978) and more recently in SLA as well. For instance, Jourdenais (2001) specifically cautioned that “the think aloud data collection method itself acts as an additional task which must be considered carefully when examining learner performance” (p. 373). Concern has also been expressed about the accuracy of participants’ memory when they provide retrospective reports (Ericsson & Simon, 1993, in cognitive psychology) or stimulated recalls (Gass & Mackey, 2016; Smagorinsky, 1994, in SLA). Because of differences between think-alouds and retrospective reports/stimulated recall, the validity concerns are manifested differently depending on the report methodology in question, and each is discussed in turn in the following sections.

**Reactivity With Think-Alouds**

In their classic model, Ericsson and Simon (1993) distinguished verbal reports according to time of reporting (concurrent vs. retrospective) and level of detail, indicating that concurrent reports are more complete and accurate (veridical) than retrospective reports, since participants who think aloud during a task are not subject to memory decay, unlike those who are asked to report their thoughts after completing a task. Therefore, Ericsson and Simon’s model established that the chief validity concern for think-alouds is reactivity rather than veridicality, and most subsequent research on think-aloud validity has been concerned with reactivity.

Regarding level of detail of reporting, the model predicts that verbalizations generated as a normal part of the solution process (what Ericsson and Simon refer to as Type 1 verbalizations, or non-metalinguistic or non-metacognitive verbalizations in the SLA literature) will be generally non-reactive; that is, they should reflect the nature of cognitive processes fairly accurately, although slowing processing slightly. Conversely, the model predicts that verbalizations that include additional reasoning or justifications that would not figure into the normal solution process (what Ericsson and Simon refer to as Type 2 and 3 verbalizations, or metalinguistic verbalizations) are more likely to be reactive (to affect task performance).

Systematic investigation into the validity of verbal reports has been undertaken since the 1950s in psychology, typically by comparing the task performance of a group that thinks aloud to a group that completes the same task silently. Statistically significant differences in task performance, when all other variables are held constant, is taken as an indication that the act of verbalizing
had an effect on participants’ thought processes, or that it was reactive. In a meta-analysis of 94 studies with a total of approximately 3,500 participants, Fox, Ericsson, and Best (2011) confirmed the predictions of Ericsson and Simon’s (1993) model, finding that (1) think-aloud participants took longer to complete tasks compared to their silent counterparts and that (2) non-metalinguistic verbalizations did not significantly affect task performance compared to silent controls: “the think-aloud effect size [was] indistinguishable from zero (r = -.03)” (p. 1). The meta-analysis also confirmed that providing metalinguistic verbalizations did have a significant effect on task performance.

Fox et al. (2011) did not specifically set out to investigate the reactivity of think-alouds used in conjunction with verbal tasks, however. Most of the studies included in their meta-analysis came from psychology and involved problem-solving tasks, making the applicability of their findings to language research uncertain.

Bowles (2010) sought to fill this gap, setting out to isolate, among other variables, the effects of task type, investigating the issue of reactivity when think-alouds were used with verbal tasks. In her meta-analysis, which included 14 unique sample studies, results were not as decisive as those in Fox et al. (2011). However, she was able to conclude that thinking aloud took significantly longer than silent task completion with all verbal tasks, although effect sizes varied substantially, ranging from small ($d = 0.16$) to very large ($d = 1.16$), with the most pronounced increase for reading tasks. Regarding the impact of thinking aloud on task performance, “in 86 per cent of the effect size calculations, the 95 per cent confidence interval overlapped [zero] indicating that the $d$ value [was] not significantly different from zero” (Bowles, 2010, p. 138). It was not possible to produce grand weighted mean effect sizes because the distributions violated the assumption of homogeneity, but this finding, taken in conjunction with that of Fox et al. (2011), suggests that with both verbal and non-verbal tasks, thinking aloud does not generally impact task performance or, by extension, alter thought processes.

**Veridicality With Think-Alouds**

Because the main validity concern for think-alouds is reactivity, their veridicality has been sparsely investigated, with the few psychology studies (Kuusela & Paul, 2000; Robinson, 2001) confirming the predictions of Ericsson and Simon’s model, finding that think-alouds tend to be more accurate and to include more information than retrospective reports. Winikoff (1967) examined the veridicality of think-alouds by triangulating participants’ eye movements and verbalizations, finding that the two methodologies corroborated each other. In the SLA literature, just two studies to date have investigated veridicality, with mixed results. Similar to Winikoff, Godfroid and Spino (2015) triangulated think-alouds with an additional measure, asking participants to think aloud and simultaneously use their index finger to track words as they read a text. Although there was overlap in what the two methods revealed, finger tracking captured some processes at a low level of awareness that were not revealed by the think-alouds. The second SLA study, Barkaoui (2011), asked essay raters to think aloud while rating, finding that more than one-third of raters reported in a subsequent interview that they had been unable to verbalize all of the thoughts that went through their minds as they rated. Because rating essays relies heavily on judgment and evaluation, it is not clear to what extent this finding is task-dependent.

Rather than taking these studies’ findings as a challenge to the validity of think-alouds on the grounds that they do not accurately represent thought processes, I would argue that the results in fact indicate that think-alouds are accurate, although they are almost certainly not exhaustive, including every fleeting thought or sensation on every level of awareness that a participant experiences. Along these lines, Leow et al.’s (2014) review demonstrated that eye
tracking and reaction times appear to be more sensitive at capturing cognition at lower levels of awareness, whereas think-alouds appear to be better suited for capturing cognition at higher levels of awareness.

**Reactivity and Veridicality With Retrospective Reports/Stimulated Recall**

Turning now to retrospective reports, their main validity concern is non-veridicality (that participants may not accurately recall and report what they were thinking at the time of task completion). The effects of memory decay are mitigated by the amount of time that lapses between task completion and reporting. In the case of retrospective reports, such as those described in Rebuschat (2013), since reporting occurs immediately after the testing session, during debriefing, the time delay is very small, which should minimize the risk of a participant forgetting what s/he was thinking during task completion.

With stimulated recall, memory decay is mitigated by the fact that there is a stimulus to aid in the recall. If the time lapse between task completion and stimulated recall is short, and if the stimulus provided is strong, such as a video-recording of the participant completing the task, memory decay should be minimal. Gass and Mackey (2016) recommend that the delay between task completion and reporting should be less than 48 hours, but depending on the amount and type of information participants have to process during the task, such a delay could arguably allow memory decay. Taking a more conservative approach, it is advisable for retrospective reporting to occur as soon after task completion as possible. As a final safeguard against non-veridicality, the verbalization instructions should be carefully written and uniformly presented to ensure that participants state what they were thinking at the time they completed the task rather than what they think at the time of reporting. Careful adherence to these guidelines and to the advice on designing a study using stimulated recall in Chapter 3 of Gass and Mackey (2016) should serve to minimize the risk of non-veridicality.

Reactivity is not typically a major validity concern for retrospective reports, since verbalization happens after the task has already been completed. However, in the case of stimulated recall, when posttests take place after a recall session, verbalization does have the potential to be reactive, especially since it provides participants with “double exposure to the targeted item under investigation in the interaction” (Leow, 2015: 149). Two SLA studies have specifically investigated this question, drawing mixed conclusions. In Egi (2007), L2 learners of Japanese received recasts while engaging in task-based interaction and verbalized either during the task (immediate recall group) or immediately thereafter (stimulated recall group). There were no significant differences in performance between the groups on the immediate posttest, but stimulated recall participants significantly outperformed immediate recall participants on the delayed posttest. Thus, Egi found stimulated recall to be reactive, having a facilitative impact on L2 learners’ performance. In a follow-up study, Egi (2008) attempted to identify the source of the reactivity, but her results showed no significant differences between groups, suggesting non-reactivity for stimulated recall in this case.

Taken as a whole, research suggests that if appropriate safeguards are taken, verbal reports can provide rich data on cognitive processes and processing that would otherwise be inaccessible. Validity concerns must always be paramount in researchers’ minds, both as they design and plan a study using verbal reports and as they analyze and interpret the resulting data.

**Current Contributions and Research**

Verbal reports are inherently flexible in the sense that they can be used to provide insight into learners’ cognitive processes and strategies on virtually any topic. In part because of this flexibility,
they are routinely used not only by researchers who take a cognitive approach to SLA, with which they are perhaps most associated (Leow et al., 2014), but also by researchers from sociocultural and conversation analytic (CA) perspectives as well. However, the theoretical frame affects the starting assumptions about verbal reports, as well as the analysis and interpretation of the data. Whereas cognitivist SLA researchers believe that verbal reports are a window into learners’ minds and a reflection of their individual processing and processes, sociocultural and CA researchers tend to view verbal reports less as a reflection of individual thought processes and more as a means of mediating the internalization of new knowledge (Vygotsky, 1987, p. 86) or, in the case of collaborative dialogue or languaging, a way of socially co-constructing knowledge (e.g., Toth, Wagner, & Moranski, 2013).

Perhaps unsurprisingly, given the diverse perspectives of researchers using verbal reports as a data elicitation tool, it is common to find both purely quantitative and purely qualitative studies using think-alouds and/or stimulated recall, as well as studies that combine both quantitative and qualitative methods. In fact, nearly 20 years ago, Leow (2000) first argued that using a combination of both online and offline measures, relying on both quantitative and qualitative analyses, is beneficial not only in providing evidence of attentional processes and processing but also “to explicate more clearly the findings” of any such study (p. 573). Gass and Mackey (2016) similarly espouse such an approach, writing that

combining analytical techniques is desirable wherever possible. The shortcomings of one type of analysis may be addressed through the strengths of another. . . . For example, . . . qualitative data can be used to shed light on the findings of any quantitative analysis.

(pp. 153–154)

Along the same lines, verbal reports can be used with a wide variety of different research designs, ranging from experimental designs in laboratory settings and quasi-experimental designs involving research on intact classes of learners, to teacher-initiated action research in classrooms and single or multiple case study designs. Depending on the specific goals and research questions being investigated, sometimes verbal reports are the sole source of data, as in Mackey, Gass, and McDonough (2000), and sometimes they complement other data sources, as in Godfroid and Schmidtke (2013), which used a combination of think-alouds, eye tracking, and posttests to investigate incidental vocabulary learning. Most studies that include verbal reports as a data elicitation measure use either think-alouds or retrospective reports, not both, and given that it is easier to minimize validity threats if only one type of verbal report is used, this is generally advisable. Nevertheless, some SLA studies have profitably incorporated both think-alouds and retrospective reports into their designs (e.g., Nassaji, 2003; Rebuschat et al., 2015; Upton & Lee Thompson, 2001).

**Recommendations for Practice**

**Recommendations for Research Designs**

As the previous sections have shown, verbal reports are a versatile data elicitation tool that can be used either by themselves or in combination with other data sources to answer a seemingly infinite number of research questions in quantitative, qualitative, or mixed-method designs. Along with these benefits come challenges as well, which are the focus of this section. Perhaps the most significant challenge to using verbal reports to elicit introspective data is that it is a labor-intensive methodology from start to finish, from the design phase through the analysis phase, and at all points in between. Although more complete guidance is presented
in Bowles (2010) for think-alouds and in Gass and Mackey (2016) for stimulated recall, this section will briefly outline key aspects of the data collection and analysis process that require careful planning and attention.

Once the researcher has determined which type of verbal report is best suited for the research questions and design of the study, participant instructions for reporting must be carefully written and piloted to ensure that they are clear and elicit the desired type of information. For instance, if a researcher wants to elicit non-metalinguistic think-alouds, the instructions need to make it clear to participants that they should verbalize whatever thoughts naturally go through their mind, without providing additional explanations or justifications for their thoughts or actions while they complete the task. Instructions should also specify what language participants are expected to verbalize in (L1, L2, or a combination). Verbal reports are more likely to be incomplete if participants cannot fully express their thoughts because of language proficiency limitations, so veridicality is a concern particularly when verbal reports are not completed in the L1 or in a language in which the participant is highly proficient (Sachs & Polio, 2007). Therefore, the language of verbalization is not a decision to be taken lightly, but rather must be considered in light of participant characteristics and logistical constraints (e.g., whether the researcher has the resources to translate verbal reports in participants’ L1).

For stimulated recalls, the researcher must identify an appropriate stimulus to use and plan the logistics of recording the stimulus. Video- and audio-recordings of task performance are by far the most common stimuli, but depending on the task type, there may be other suitable possibilities (e.g., transcripts of computer-mediated chat (CMC) sessions in Smith, 2012). With the stimulus chosen, the researcher must decide how to prompt participants to talk about what they were thinking at the time of the task rather than at the time of report, how to select specific excerpts of the stimulus to inquire about during the recall session, and how to inquire about participants’ thoughts without asking leading questions. In all cases, standardized scripts should be used during data collection. As with any experimental study, all experimental tasks and procedures should be piloted to ensure, among other things, that instructions are clear, sufficient time is planned for each reporting session, and all recording equipment is working as expected.

Although verbal reports are logistically simpler to set up than measures like eye tracking, which require specialized equipment, software, and programming, it is still necessary to have a protocol in place to ensure consistent administration. If possible, it is recommended that researchers sit in on a verbal report session before conducting one on their own (Gass & Mackey, 2016). The training protocol should focus not only on logistical issues (e.g., when to turn recording devices on and off) but also on how to avoid common pitfalls. For think-alouds, this includes how to prompt participants to speak if they fall silent for a pre-determined period of time. Normally, this jars participants into verbalizing if they have unintentionally fallen silent, but in some cases participants may report not being able to verbalize any thoughts at a given time (or not “thinking anything”). If this occurs only sporadically, such brief lapses are unlikely to affect the quality of the data. However, if a participant is largely silent, the representativeness of the verbal report could be called into question and would likely need to be excluded from analysis. In stimulated recall, the researcher periodically refers to particular portions of the stimulus and asks the participant to state what s/he was thinking at that time. This pushes the participant to respond, and researchers must be trained to accept responses that indicate a lack of recall (e.g., “I don’t remember thinking anything in particular”) and have a script including some language that they can use to provide back-channeling without pushing the participant to say what s/he thinks the researcher wants to hear.

Once necessary adjustments have been made based on piloting, data collection can begin. Collecting verbal report data is time-consuming in and of itself, since it is most commonly done one-on-one. Once verbal reports have been recorded, the labor-intensive data analysis phase
begins. Participants’ entire verbal report, or in some cases, selected portions of the report that have been identified in advance, must be transcribed. Although the level of detail of the transcription varies according to the research question and framework being adopted, producing even a simple orthographic transcript is time-consuming, with 15 minutes of speech taking about 1 hour to transcribe. Transcriptions should then be checked for accuracy, and all researchers involved in coding should be trained to use the established coding scheme. In many cases, the coding scheme may need to be modified as a result of rater socialization, particularly for high-inference coding categories. Then all of the data are coded (typically by one rater), and a randomly selected subset is chosen to be coded by a second rater and used to calculate inter-rater reliability, which should be included in any published reports. With coding complete, quantitative and/or qualitative data analysis can take place and inferences can be drawn.

One final challenge that researchers should bear in mind when they use verbal reports is individual differences, which have been suggested to have more of an effect on think-alouds than on stimulated recall, since the former require participants to perform the dual task of verbalizing while completing another cognitive task, whereas the latter do not. Anecdotally, some participants have a greater propensity to verbalize than others, for instance, although it should be noted that some tasks seem to lend themselves more than others to thinking aloud (c.f., problem-solving vs. reading tasks in Leow et al., 2014). In the vast majority of studies on verbal reports (in both psychology and SLA), no mention is made of individual difference variables. An implicit assumption seems to be that, at least in group-level analyses (as opposed to case studies), any individual differences that affect verbal reporting will “come out in the wash” and not significantly impact the study’s findings. Although neither cognitive psychology studies nor SLA studies employing verbal reports have consistently identified individual difference variables as having a significant effect on verbalizations, the results of one SLA study, Goo (2010), showed that thinking aloud was reactive for L2 learners with high working memory but not for those with low working memory. Such findings need to be replicated in order to be generalized, but they suggest that researchers should consider collecting individual difference data in order to determine what impacts such variables have on reactivity.

**Recommendations for Classrooms**

In the SLA literature, verbal reports have been discussed almost exclusively as a research tool, rather than as a pedagogical technique. Yet, given the insights they provide, it stands to reason that they could be of great benefit in the classroom as teachers seek to understand their students’ cognitive processes. As has been shown in the preceding sections, verbal reports are non-reactive more often than not, and in cases where reactivity has been found, its effect is overwhelmingly positive, improving participants’ performance rather than hindering it. Indeed, of the 22 studies from non-SLA fields reviewed in Bowles (2010) in which verbalizations were found to be reactive, 17 (77%) showed either positive reactivity, with participants who thought aloud having improved performance compared to those who completed the same task silently or had mixed results, with facilitative results in one experiment but not in another (see Bowles, 2010, Chapter 2 for detailed reviews of all included studies). The remaining five studies (23%) found that verbalization led to negative reactivity, or decreased performance compared to silent controls. Turning to SLA studies, a very similar pattern emerges. Only four studies to date have found reactivity, and of these three were positive (Rossomondo, 2007; Sanz et al., 2009 with L2 learners; Yanguas & Lado, 2012 with heritage learners). Just one study (Sachs & Polio, 2007) found negative reactivity, albeit in just one of their two experiments, but since learners in this study were asked to verbalize exclusively in their L2, language of reporting could have contributed to this anomalous result.
These findings seem to bode well for the use of think-alouds as a pedagogical technique, and in fact English language arts classes have been using think-alouds since the 1980s to model reading and writing. Research has found that skilled readers share particular online processing strategies as they read a text, and think-alouds have been used to teach effective reading strategies to native English speakers (Crain-Thoreson, Lippman, & McClendon-Magnuson, 1997) and, to a lesser extent, to L2 English speakers (Pritchard & O’Hara, 2006; Wolsey & Lipp, 2018). Professional development books targeting K–12 teachers (Wilhelm, 2001) and even professional organizations tout the use of think-alouds in not just English language arts classrooms but in mathematics and inquiry-based science as well (Martin-Hansen & Caton Johnson, 2006). In fact, although think-aloud is not commonly covered in methodology courses for second/foreign language teachers, there is good reason to believe that it would be effective in such contexts as well. For instance, language learners often struggle when they come across unknown words or achieve local comprehension at the expense of global text comprehension. Appropriate modeling of reading strategies could lead students through a few sentences or a short paragraph of text, with the teacher providing a think-aloud demonstrating exemplary processing, predicting what will happen next, encouraging guessing word meanings from context, based on morphological patterns, and so forth. Then students could be instructed to read the remaining text on their own, either individually or in a group while thinking aloud using the strategies they have just been shown (see Brown & Lytle, 1988, for group think-aloud). Similarly, teachers could model good listening comprehension strategies and lead students through short excerpts. The technique is versatile in the classroom and could be adapted to different levels of proficiency and used in upper-level content courses on literature, culture, and linguistics, as students seek to build increasingly accurate and precise language competency.

Future Directions

As the preceding sections show, verbal reports are a flexible data elicitation tool that can be used to gain insight into cognitive processes and help to answer a wide range of research questions. As with any methodology, there are limitations with verbal reports that must be acknowledged. Cohen (1996, 1998) provides a thorough treatment of the expectations and limitations of verbal reports in the context of SLA research. The most central limitation has to do with the potential reactivity and non-veridicality of verbal report data, but if proper protocols are followed, as outlined in this chapter, these validity threats can be minimized. As an additional precaution, it is recommended that all studies with a think-aloud group have a silent control group that is otherwise matched for all relevant variables. This way, the silent and think-aloud groups’ performances can be compared to determine whether there was reactivity for that group on the specific task(s) in question. Along the same lines, although it is not mentioned in Gass and Mackey (2016), in research where stimulated recall precedes a posttest, it is advisable to have a control group that participates in all phases except the stimulated recall session, to ensure that double exposure is not impacting the results. The second main limitation is that some aspects of cognition are likely to be inaccessible by means of verbal reports; processing at lower levels of awareness may not be captured by verbal report but may require a more sensitive or fine-grained method (Leow et al., 2014). As such, at the outset of a project, researchers should carefully consider whether verbal reports will be suitable to answer their specific research questions or whether they would be better served by another data elicitation instrument.

Looking forward, verbal reports have a bright future in applied linguistics research. Increasingly, researchers have used them to triangulate data from other measures, such as eye-tracking or reaction times (Godfroid & Schmidtke, 2013; see also Godfroid [Chapter 4]; Gor & Chrabaszcz [Chapter 21]; Indrarathne [Chapter 23]; Issa [Chapter 32]; Lee & Choi [Chapter 20]; Lee &
Doherty [Chapter 25, this volume). Introspective data can only help researchers to gain a more nuanced view of language learning and use as the field becomes increasingly sophisticated and integrates new, more sensitive measures into its methodological repertoire.

Without a doubt, as verbal reports expand in their popularity as a data elicitation tool, further research is still needed regarding their reactivity and validity in language research. Since Leow and Morgan-Short (2004) first explored the topic of reactivity in SLA, more than a dozen published studies have examined the reactivity of various types of reports with a number of task types in L2 research. Yet among the topics that have been under-researched to date are (1) how participants’ individual difference variables (e.g., working memory and L2 proficiency) affect verbalization and (2) the extent to which thinking aloud is reactive on L2 tasks other than reading, which has been the focus of the lion’s share of current reactivity studies. Given the positive effects that think-aloud instruction has had in other fields, research on the effects of think-alouds as an instructional strategy in L2 classrooms is also warranted.

References


Further Reading

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York, NY: Routledge. (In addition to presenting a quantitative meta-analysis of findings from studies involving verbal tasks and think-alouds, this book also provides guidance for researchers wishing to use think-alouds in their own applied linguistics research.)

Gass, S. M., & Mackey, A. (2016). *Stimulated recall methodology in applied linguistics and L2 research* (2nd ed.) New York, NY: Routledge. (Now in its second edition, this how-to guide demonstrates the use of stimulated recall in a range of different applied linguistics studies to address a variety of research questions.)


labs, can be used in quantitative or qualitative research designs to generate evidence for claims about higher-level thinking and learning, with a focus on K–12 and postsecondary educational settings.)

Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design*. Oxford: Oxford University Press. (This book demonstrates how think-alouds can be applied to questionnaire design to develop survey questions that are valid, reliable, and well-understood, thereby producing low levels of response error.)