Annual Review of Applied Linguistics (2010), **30**, 111–127.
© Cambridge University Press, 2010, 0267-1905/10 $16.00
doi:10.1017/S0267190510000036

Concurrent Verbal Reports in Second Language Acquisition Research

Melissa A. Bowles

This article provides an overview of the ways in which concurrent verbal reports, sometimes referred to as think-alouds, have been used in cognitivist second language acquisition (SLA) research. It addresses two issues related to the validity of verbal reports—reactivity and veridicality—and reviews studies that have examined the validity of verbal reports in SLA. On the basis of the results of a meta-analysis of studies comparing the performance of silent and think-aloud groups (Bowles, 2010), this article concludes with suggestions for further research into the issue of validity and recommendations for the careful use of think-alouds in research.

INTRODUCTION

Language acquisition research consists (at least in part) of measuring and describing learners' knowledge about a language. But it is well known that the evidence stemming from learners' language production is incomplete; some other method is needed to elicit a more complete data set. Verbal reports have been used in first (L1) and second (L2) language research to provide insight on a variety of issues that production data alone cannot address, such as language learners' thought processes and strategies. Simply put, verbal reports are verbalizations learners make either while they complete a task or sometime thereafter. Verbal reports completed during a task are referred to as concurrent reports (or think-alouds), and those completed after the task are referred to as retrospective reports (or stimulated recalls).

This article provides an overview of how think-alouds have been used in language research and seeks to provide some answers to questions about the validity and use of think-alouds. Readers interested in the use of stimulated recalls are referred to Gass and Mackey (2000), which is indispensable on the topic.

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

USE OF AND ASSUMPTIONS ABOUT VERBAL REPORTS

Decades before they were first applied in language research, verbal reports were used extensively in cognitive psychology, where they were traditionally used to investigate problem-solving strategies on nonverbal tasks and puzzles (J. H. Davis, Carey, Foxman, & Tarr, 1968; Gagné & Smith, 1962). Since the 1980s, "both concurrent and retrospective verbal reports [have become] generally recognized as major sources of data on subjects' cognitive processes in specific tasks" (Ericsson & Simon, 1993, p. xi). The collection of verbal reports has become standard in many fields, such as accounting (Anderson, 1985); care planning (Fowler, 1997); counseling (Bozarth, 1970); drug and alcohol addiction treatment (Midanik & Hines, 1991); ergonomics (Brinkman, 1993); marketing (Biehal & Chakravarti, 1989); nursing (Greenwood & King, 1995); software engineering (Hughes & Parkes, 2003); medicine, where verbal reports are routinely used in the treatment of autism and developmental disorders (Berg, 2002; Friedman & Mulhern, 1976); speech pathology (Karsenty, 2001); neurology (Chan, Hoosain, & Lee, 2002); and cardiology (Bernardi et al., 2000). In each of these fields, verbal reports provide insight into the cognitive processes of clients and patients.

Think-alouds have also been used to investigate many aspects of first and second language acquisition and use. In the L1 literature, there have been numerous studies on reading, which have tended to focus on reading strategies used by older and younger, more and less successful readers (Cohen, 1986, 1987; Earthman, 1992; Folger, 2001; Gordon, 1990). Think-alouds have also been used to study the L1 writing process (Breetvelt, 1994; Zellermayer & Cohen, 1996). In some cases, think-alouds have been used as one component of programs designed to improve students' reading (Baumann, Jones, & Seifert Kessell, 1993; Robertson, 1995; Walczyk, Marsiglia, Bryan, & Naquin, 2001; Wilhelm, 2001) and writing ability (Box, 2002; Cushman, 2002; Fresch, Wheaton, & Zutell, 1998; Scardamalia, 1984).

In second language acquisition (SLA) research, think-alouds have also been used to study reading (Carrell, 1989; Cohen, 1986, 1987; Cohen & Cavalcanti, 1987; Hosenfeld, 1976, 1977, 1979, 1984; Pressley & Afflerbach, 1995; Pritchard, 1990) and writing (Qi & Lapkin, 2001; Sachs & Polio, 2007). However, they have also been used to investigate a wide array of other phenomena, including L1 and L2 strategy use (Chamot & El Dinary, 1999; J. Davis & Bistodeau, 1993; Yamashita, 2002), lexical organization (Herwig, 2003), the use of translation in L2 learning (Kern, 1994), interlanguage pragmatics (Woodfield, 2008), the role of different levels of awareness in L2 learning (Leow, 1997, 1998a, 1998b, 1999, 2000, 2001a, 2001b; Rosa & Leow, 2004a, 2004b; Rosa & O'Neill, 1999), and the relationship between explicit and implicit L2 knowledge (Ellis, 2004; Hu, 2002).

As this brief review shows, verbal reports have been used to answer a wide array of research questions in SLA. Verbal reports have been used in cognitivist approaches to SLA, where they are viewed as a window into the minds of learners, that is, as a means of gathering information about their internal thought processes. Crucially, an underlying assumption is that verbalizations can

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

accurately reflect thought processes without altering them and can therefore be a data collection tool.

CONTROVERSY OVER THE USE OF VERBAL REPORTS

Indeed, despite the frequency with which verbal reports are gathered, their use is not uncontroversial. More than three decades ago cognitive psychologists raised questions about whether verbal reports actually capture thought processes accurately, or whether they alter those processes (Nisbett & Wilson, 1977; Payne, Braunstein, & Carroll, 1978). Similar concern has since been voiced in SLA. 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).

Specifically, the two main validity issues surrounding the use of verbal reports are reactivity and veridicality. Verbalization is said to be *reactive* if it alters task performance, and it is said to be *veridical* if it accurately reflects the participant's cognitive processes. Ericsson and Simon's (1993) classic model of verbalization speaks to the validity of think-alouds. The model does not treat all verbal reports alike; rather it distinguishes them along two dimensions: time of reporting (concurrent vs. retrospective) and level of detail. It predicts that concurrent reports will be 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 some time after completing a task. The model furthermore makes different predictions based on how detailed the verbalizations are required to be. Verbalizations that are generated as a normal part of the solution process (what Ericsson & Simon refer to as Type 1 verbalizations, or nonmetacognitive verbalizations) will be largely nonreactive; that is, they will reflect the nature of cognitive processes fairly accurately, while slowing processing slightly. Conversely, their model predicts that verbalizations that include additional reasoning or justifications that would not normally figure into the normal solution process (what Ericsson & Simon refer to as Type 2 and 3 verbalizations, or metacognitive verbalizations) are more likely to be reactive, or to affect task performance.

Experimental studies examining the effects of think-alouds on problem-solving and decision-making tasks have been conducted in the field of cognitive psychology since the 1950s. To assess the extent to which verbalization affected participants' task performance, one group would complete a task silently, while another would complete the same task while verbalizing. Overall, a synthesis of the studies reviewed in Ericsson and Simon (1993) reveals a clear pattern of results. When participants were asked to think aloud nonmetacognitively, their task performance (usually measured as accuracy) was generally not significantly different from that of the participants in the silent groups. On the other hand, when participants were asked to think aloud metacognitively, their task performance was, in most studies, significantly different from that of participants in silent groups. In some cases, participants in verbalization groups outperformed silent participants, and in others, the verbal ones

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

underperformed the silent, suggesting that sometimes verbalization is facilitative and sometimes it is detrimental. Furthermore, participants who verbalized (either nonmetacognitively or metacognitively) tended to take significantly more time to complete the task than silent participants, because the additional time needed for verbalization increases the overall solution time. These findings support the predictions of Ericsson and Simon's model. However, most of the studies used nonverbal and/or problem-solving tasks, limiting their relevance to SLA.

REVIEW OF SLA STUDIES ON REACTIVITY

Leow and Morgan-Short (2004) was the first SLA study to empirically investigate the reactivity of think-alouds. Their study investigated the effects of nonmetacognitive think-alouds on 77 beginning Spanish learners' text comprehension and intake and written production of formal imperative morphology in Spanish. Learners in the control condition read and completed the tasks silently, while learners in the experimental condition read and completed the tasks while thinking aloud. Results showed nonreactivity, since the two groups did not differ significantly on either text comprehension or posttask assessments of the targeted inflectional morphology. Time on task was not a dependent variable measured in the study, so it was not possible to determine whether thinking aloud affected solution time.

Expanding on Leow and Morgan-Short (2004), Bowles and Leow (2005) sought to investigate the reactivity of both metacognitive and nonmetacognitive think-alouds on text comprehension and item and system learning of the pluperfect subjunctive. Participants were 45 fifth-semester Spanish learners who were randomly assigned to either a control or to one of two verbalization groups (nonmetacognitive or metacognitive). All participants read a text that included tokens of the targeted structure and then completed the same comprehension and written production tasks. The only difference between the groups were verbalization instructions: Participants in the control group were silent; those in the nonmetacognitive group were instructed to "say whatever passed through [their] minds"; and those in the metacognitive group were instructed to comment specifically about their processes. Similar to Leow and Morgan-Short (2004), the results indicated that, compared with a control group, nonmetacognitive verbalization did not significantly affect either comprehension or written production of the targeted form. However, metacognitive verbalization caused a significant decrement in text comprehension but no significant difference for production when compared with either the control or to the nonmetacognitive group. Results also indicated that both verbalization groups took significantly more time to read the text and complete the postassessment tasks than the control group, but that there was no significant difference between the two think-aloud groups in terms of time on task.

Sachs and Polio (2007) examined the reactivity of think-alouds on an L2 writing task. The study was carried out in two experiments. In Experiment 1, 15 adult English as a second language (ESL) learners participated in a three-stage, composition-comparison-revision task. Each learner participated in the

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

three-stage process three times, once receiving written corrections on their compositions, once receiving reformulations of their errors, and once receiving reformulations of their errors and verbalizing while they compared their original composition to the reformulated version. Results of a Wilcoxon signed rank test indicated that when learners were silent during the comparison stage (reformulation), they subsequently revised more errors in their composition than when they verbalized during the comparison stage (reformulation + think aloud), although the effect size ($\eta^2 = 0.28$) was weak, suggesting that the difference in performance between the two groups, though statistically significant, was small. Experiment 2 was then conducted as a nonrepeated-measures study, this time with 54 ESL learners, randomly assigned to one of four conditions: (1) error correction, (2) reformulation, (3) reformulation + think aloud, or (4) control, in which learners received no feedback on their writing. A Mann-Whitney test comparing the reformulation and reformulation + think-aloud conditions revealed no significant differences ($p = .77$), indicating that in Experiment 2, learners who verbalized during the comparison stage and those who did not corrected a similar number of errors during the revision stage. In summary, verbalization was found to be reactive in Experiment 1, since it appeared to hinder learners' ability to make revisions to their compositions. In Experiment 2, however, verbalization was found to be nonreactive. In both Experiment 1 and Experiment 2, learners were instructed to verbalize in English (their L2), unlike in the previous SLA studies (Bowles & Leow, 2005; Leow & Morgan-Short, 2004), which gave learners the option to speak in either their L1 or L2. This difference between Sachs and Polio (2007) and previous studies, combined with the repeated-measures nature of Experiment 1, is cause for the results to be interpreted with caution.

Sachs and Suh (2007) investigated the issue of reactivity of think-alouds in the context of synchronous computer-mediated communication (CMC). Thirty ESL learners participated in the study, which targeted sequence of tense in English. All learners interacted individually with the researcher via CMC, first taking a written text completion and interactive story-retelling pretest, followed by a story-retelling task either with or without textually enhanced recasts, and with or without thinking aloud. Each learner then completed an interactive story-retelling posttest and a text completion posttest. Learners were allowed to speak in either their L1 (Korean) or in their L2 (English) while thinking aloud, as they preferred. Results indicated that the + and – think-aloud groups did not differ significantly in the amount of time spent on task ($t(28) = -0.113$, $p = .91$), indicating that verbalization was nonreactive for time on task, in contrast to the findings of Bowles and Leow (2005), the one previous SLA study that examined time as a dependent variable. Repeated-measures ANOVAs were then run with ± think-aloud as the between-subjects factor, time as the within-subjects factor, and text completion and production (story retelling) as the dependent variables. There was a significant between-group main effect on text completion ($F(1,26) = 6.478$, $p = .02$), which indicates that the group that verbalized performed differently on the text completion test than the group that did not verbalize, regardless of time. No significant interaction effect between time and group was found for either the story retelling or the text completion test, leading the authors to

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

conclude that strong generalizations regarding reactivity should not be drawn from the data.

In a study investigating the role of lexical temporal indicators, such as temporal adverbs, in the incidental acquisition of the Spanish future tense, Rossomondo (2007) compared groups of first-semester Spanish students who read a passage silently and those who read the same passage while thinking aloud nonmetacognitively. Following recommendations from Ericsson and Simon (1993) and from previous SLA studies, Rossomondo gave participants a sample think-aloud as well as a short warm-up text during which they were instructed to think aloud. After reading the experimental passage, all participants completed a 13-item multiple-choice comprehension test in English, followed by either a 13-item form-recognition task, or a 13-item form-production task. The form-recognition task consisted of 13 target sentences taken from the passage (presented without lexical temporal markers) with a blank where the inflected verb should have been. Participants had to choose the correct inflected form (the future) from among four options. The form-production task consisted of the same 13 target sentences taken from the text, but in this case each blank was followed by the infinitive form of a verb in parenthesis, and participants were instructed to conjugate the verb in the same form as in the passage. No reaction time or time on task data were collected. Results indicated that participants who thought aloud while reading scored statistically similarly on comprehension measures to participants who read the passage silently ($F(1,159) = 0.078$, $p = .781$). For recognition and production of the target form, results were quite different, with participants who thought aloud scoring significantly higher on both the form-recognition ($F(1,77) = 12.194$, $p < .001$) and form-production ($F(1,80) = 7.352$, $p < .008$) tests.

Bowles (2008) investigated the effects of completing nonmetacognitive and metacognitive think-alouds while performing an L2 problem-solving task on subsequent written production of previously encountered and new exemplars of a target form. A total of 194 first-semester learners of Spanish were randomly assigned to one of six experimental conditions, which differed in terms of the type of verbalization (metacognitive, nonmetacognitive, or silent) and the type of feedback (implicit or explicit). Learners were instructed to speak in either L1 (English) or L2 (Spanish), as they felt most comfortable. Results showed that metacognitive verbalization significantly increased time on task ($F(2,193) = 15.763$, $p < .0001$), with the metacognitive think-aloud group taking significantly more time to complete the task than either the nonmetacognitive or silent group. Metacognitive verbalization also hindered participants' ability to produce exemplars of the target structure that they had seen during the experimental task ($F(2,193) = 3.778$, $p < .05$). However, neither type of verbalization significantly affected participants' ability to produce novel exemplars of the target structure ($F(2,193) = 1.713$, $p = .183$), and there was no interaction between verbalization and feedback. These results as a whole indicate that nonmetacognitive verbalization was nonreactive for both item and system learning, in accordance with Ericsson and Simon's model. However, the results indicate that metacognitive verbalization was reactive on item learning but nonreactive on system learning.

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

Yoshida (2008) investigated the role that verbalization plays in an L2 reading task, thereby building on the work of Leow and Morgan-Short (2004) and Bowles and Leow (2005). In her study, 64 Japanese ESL students were randomly assigned to a verbalization condition (think-aloud or non-think-aloud) and to one of three reading conditions (a control passage, the same passage with questions embedded throughout, or the passage with guidance to help students produce an outline). Students were allowed to think aloud in their L1 (Japanese) or L2 (English), so that their verbalizations would not be constrained by their language ability. The instructions required students to produce metacognitive think-alouds, since they were asked to explain how they made each decision provided in their responses to written while-reading tasks. After reading the passage and completing a corresponding while-reading task, each learner completed a written recall of the propositions in the passage. The recall scores were taken as an indication of reading comprehension, and results of a two-way ANOVA found no significant main effect for reading condition ($F(1,58) = 0.795$, $p = .376$) or for task type ($F(2,58) = 0.392$, $p = .677$), and no significant interaction was found between reading condition and task type ($F(2,58) = 0.182$, $p = .834$). These results indicate that learners who verbalized performed similarly to learners who did not verbalize, regardless of the task they engaged in while reading. A two-way ANOVA revealed that the think-aloud group took considerably more time on task than the non-think-aloud group ($F(1,58) = 12.76$, $p = .001$).

Sanz, Lin, Lado, Bowden, and Stafford (2009) reported on two studies investigating the reactivity of concurrent verbal reports in an L2 instructional lesson. In Experiment 1, 24 L1 English-speaking students, 11 in the think-aloud group and 13 in a silent group, completed a computerized lesson on the Latin case system. After the lesson, participants took an aural interpretation test, a written grammaticality judgment test, and a sentence production test that required them to drag and drop appropriate morphemes to create a sentence describing an on-screen photograph. On all tests, there was a main effect only for time, indicating that students in both the silent and think-aloud groups learned as a result of the lesson, and that verbalizing during the lesson had neither a facilitative nor a detrimental effect on performance. Furthermore, Sanz et al. (2009) used a more precise measurement than that used in previous studies to measure time. They measured mean reaction times on correct responses from the three pre- and posttests and calculated a grand mean on which to base their overall time measure for each participant. The think-aloud group had longer reaction times than the silent group on just one of the posttests, the grammaticality judgment, indicating mixed results for latency in this experiment.

In Experiment 2, 24 college-age students, 11 in the think-aloud group and 13 in the silent group, completed a less explicit version of the treatment used in Experiment 1. The only difference between the two instructional treatments was that the treatment in Experiment 1 included an explicit grammar lesson on the case system, whereas the second treatment did not. Instead, students had to rely on knowledge they gained from task-essential practice and explicit feedback. All pre- and posttest measures were the same as in Experiment 1. As in the first experiment, there were main effects for time for all tests, indicating that the treatment caused students to improve in their ability to interpret and produce

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

sentences in Latin. But, in contrast to Experiment 1, there were also reactivity effects because students in the think-aloud group performed significantly better than those in the silent group on both the grammaticality judgment and production tests. In this case, verbalization had a facilitative effect on subsequent performance. As for latency, there were no significant differences in reaction times for the two groups across the three tests, indicating that in this experiment, although verbalization enhanced performance, it did not cause any change in the speed with which participants responded to stimuli in the L2.

Finally, in a study similar to Leow and Morgan-Short (2004), Polio and Wang (2005) examined the reactivity of think-alouds with advanced-proficiency learners completing a reading task. The authors hypothesized that the reading strategies of advanced learners might be different enough from those of the beginning learners in the original study to cause a change in reactivity. In Polio and Wang (2005), 30 Chinese L1 learners of English were randomly assigned to a think-aloud or to a silent reading group and were instructed to read a passage in English, seeded with both frequent and infrequent phrasal verbs. Upon completing the reading task, participants took a comprehension test (in the L2, English) as well as written production and recognition tests targeting the phrasal verbs found in the passage. Since there were 15 participants per group, less powerful nonparametric statistical tests (Mann-Whitney U tests) were used to compare the performance of the think-aloud and silent groups on the three measures. The only significant difference occurred in comprehension, where participants in the think-aloud group comprehended the text significantly worse than participants in the silent group ($p = .01$). Polio and Wang then examined the contents of the think-alouds to investigate the nature of the reading process for this group of learners. They determined that whereas in Leow and Morgan-Short, translation was the predominant reading strategy, for these advanced learners who used English for academic purposes, translation was relatively uncommon. The researchers hypothesized that this difference in strategy use may have been largely responsible for the differing findings between the original and the replication. Since learners in Leow and Morgan-Short relied heavily on translation, they were easily able to verbalize the contents of their short-term memory without any effect on comprehension. On the other hand, the learners in the replication were engaging in other, more complex strategies, so it may have been more cognitively demanding for them to verbalize, contributing to the finding of reactivity in comprehension.

META-ANALYSIS OF STUDIES ON REACTIVITY

As the review in the preceding section shows, there is a mounting body of primary research investigating the reactivity of think-alouds on verbal tasks. However, no one study on reactivity, regardless of its size, complexity, or number of participants, can provide reliable answers, since individual study findings are susceptible to chance variability and any number of idiosyncrasies in design and sampling. This section therefore presents the results of a quantitative research synthesis, or meta-analysis, which was carried out on reactivity studies

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

that used verbal tasks. A complete description of the meta-analysis is provided in Bowles (2010).

Besides providing a summary of existing research in different fields, a meta-analysis compares the outcomes of a range of studies with an array of independent variables, in an attempt to identify patterns. In SLA, meta-analyses have increased in frequency in the past decade, as researchers in the field have started to adopt the method to provide more comprehensive answers to long-standing questions, such as the role of instruction in SLA (Norris & Ortega, 2000), the efficacy of different types of corrective feedback (Russell & Spada, 2006), the effects of conversational interaction (Keck, Iberri-Shea, Tracy-Ventura, & Wa-Mbaleka, 2006; Mackey & Goo, 2007), the impact of attitude and motivation (Masgoret & Gardner, 2003), and visual/textual input enhancement (Lee & Huang, 2008) on language learning.

Identification of Studies

Since reactivity studies involving verbal tasks have been conducted both in psychology and in linguistics, Bowles searched the *PsycInfo* database, in addition to *Linguistics and Language Behavior Abstracts* (LLBA) and the *Education Resources Information Center* (ERIC) database, using the following subject, abstract, and keywords, and combinations thereof: *concurrent verbal report(s), concurrent verbal protocol(s), reactivity, think aloud(s), verbal task,* and *language task.* Then, the search was replicated in major SLA journals to ensure that published studies had not been missed. Specifically, separate searches were conducted in each of the following journals: *Applied Linguistics, Applied Psycholinguistics, Canadian Modern Language Review, Computer Assisted Language Learning, Foreign Language Annals, International Journal of Educational Research, Language Learning, Language Learning & Technology, Language Teaching Research, The Modern Language Journal, ReCALL, Second Language Research, Studies in Second Language Acquisition, System,* and *TESOL Quarterly.*

Following recommendations in the SLA literature (Norris & Ortega, 2006), unpublished papers, sometimes referred to as the *fugitive literature* (Rosenthal, 1994), were included in the meta-analysis to the extent possible.

Selection Criteria

The studies identified through the subject, abstract, and keyword searches were narrowed down by using the following selection criteria.

Inclusion Criteria

(1) The study was published, in press, or accepted for publication prior to November 2009.
(2) The study compared one or more verbalization groups to a silent (control) group, or it compared one or more verbalization groups to each other, in the absence of a silent (control) group.

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

(3) At least one of the verbalization groups required participants to verbalize concurrently, while performing some type of verbal task (involving either language comprehension or production).
(4) Participants in the studies could be either adults or children. It is important to note that children in all of the studies were verbalizing (and completing verbal tasks) in their L1. To disentangle the variety of variables, language of verbalization and language of the task were coded separately for each study.
(5) The study included sufficient descriptive statistics to enable the calculation of effect sizes. (Three studies—Rhenius & Deffner, 1990; Russo, Johnson, & Stephens, 1989; Stratman & Hamp-Lyons, 1994—were excluded from the meta-analysis on this basis.)
(6) The study included sufficient information to be coded according to the coding scheme adopted.
(7) Studies written in English, French, and German all appeared as results in the database searches and were evaluated for inclusion. That is, no study was excluded simply on the basis of language of publication.

Exclusion Criteria

Studies were excluded from the analysis if concurrent verbalization, although a part of the study design, was not an independent variable (Piolat & Olive, 2000; Witte & Cherry, 1994).

Based on the foregoing criteria, a total of 14 unique sample studies, described in 12 research reports, were identified for inclusion in the meta-analysis. Those studies are identified in the references with a single asterisk.

Combining Effect Sizes

One advantage of a meta-analysis is that it enables the combination of many studies to produce one effect size. Effect sizes of studies cannot be combined at random, however. Even when the meta-analyst has good reason to believe that studies are similar in terms of their substantive and/or methodological characteristics, it is necessary to quantitatively analyze the studies using a homogeneity test to determine whether the variability across effect sizes is greater than or less than what would be expected from sampling error alone. If the group of studies is found to be homogeneous, "the mean effect size is clearly a representative and meaningful summary of the distribution of effect size values" (Lipsey & Wilson, 2001, p. 162).

Sometimes even when studies have been grouped by substantive and methodological variables and are considered to be similar enough to warrant averaging their effect sizes, the homogeneity test determines that they are heterogenous (i.e., that the variability observed is larger than what would be due to different samples across the different studies). In these cases, it is not advisable to combine effect sizes but rather to determine whether some other variable explains a significant portion of the between-study variance. "When studies that presumably examine the same thing disagree, it is neither wise nor especially

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

meaningful to resolve their differences simply by averaging them all together. The average over contrary results is not likely to converge on the truth, just muddle it" (Lipsey & Wilson, 2001, p. 162).

Main Findings

It was not possible to average the effect sizes of all studies to produce grand weighted mean effect sizes for accuracy or for latency because those effect size distributions violated the assumption of homogeneity. This result indicates that the answer to the question of reactivity and think-alouds is not unidimensional but rather is dependent on a host of variables.

Nonetheless, the fact that almost all of the effect size means overlapped the zero value indicates that think-aloud groups do not consistently perform significantly differently than silent control groups. Furthermore, the small effect sizes of $d \leq 0.5$ for 18 of the 22 (81%) calculations done as part of the meta-analysis indicate that any posttest differences between silent and think-aloud groups tend to be small. In other words, compared to participants completing the same tasks silently, participants who think aloud tend to perform only slightly better or slightly worse on posttests.

The results for time on task were more decisive, indicating across the board that thinking aloud increases time on task. Nevertheless, effect sizes for latency ranged from small ($d = 0.16$) to very large ($d = 1.16$), with the largest effects demonstrated when participants were required to think aloud while performing reading tasks.

FUTURE RESEARCH DIRECTIONS

The systematic coding and analysis of reactivity studies that was conducted as part of the meta-analysis (Bowles, 2010) revealed several areas that, despite their theoretical importance, have been underinvestigated. At the same time, the meta-analysis highlighted the role that several previously untested variables unique to the context of language research may have on whether thinking aloud is or is not reactive.

Needed are studies such as Polio and Chiu (2007) that examine the effect of using just one language (either L1 or L2) in thinking aloud. This is because most of the studies conducted so far have examined the reactive effects of think-alouds when participants were allowed to select either the L1 or the L2, or to use a combination of the two languages, in their reports. Certainly, this is a gap in the research that should be filled.

Similarly, some task types have been investigated more heavily than others with regard to reactivity. Topping the list are studies that have examined the reactivity of think-alouds on reading ($n = 6$) and grammar learning tasks ($n = 7$). But the effect of think-alouds on the writing process has only been investigated in two unique sample studies, making it untenable to draw firm conclusions about reactivity with this task type.

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

More research is clearly needed overall in examining reactivity in conjunction with verbal tasks, given the paucity of reactivity research with language tasks as compared to nonverbal and problem-solving tasks. Empirical studies are needed to test the sources of between-study variance identified in this meta-analysis to determine to what extent they affect reactivity. Analyses to identify the sources of between-study variance revealed that a portion of the variance could be attributed to expected sources, such as type of report (metacognitive vs. nonmetacognitive), which has been discussed at length in the psychology literature (Ericsson & Simon, 1993). The other sources of between-study variance that were identified, L2 proficiency and explicitness of instruction, are unique to the language research context. Future research is clearly needed to determine how verbalizing affects learners (a) at different proficiency levels and (b) in conjunction with more and less explicit instruction.

Although reactivity has been the focus of the majority of research on think-alouds, both in psychology and in SLA, the issue of veridicality should not be forgotten. In research studies that use think-alouds as a data elicitation tool, once the think-aloud group has been shown to perform statistically similarly to a silent control group, it is also important to examine the issue of veridicality. The small number of studies in psychology that have investigated veridicality have demonstrated that concurrent verbal reports tend to be more accurate (Robinson, 2001) and to include more information (Kuusela & Paul, 2000) than retrospective reports. In SLA, only Polio and Chiu (2007) have addressed veridicality directly, using a within-subjects design to examine concurrent and retrospective reports of the L2 writing process. In the study, advanced ESL learners completed concurrent verbal reports while writing and were then asked to provide a retrospective report (stimulated recall). Results showed that learners were indeed able to provide additional information in the retrospective reports, indicating that although concurrent verbalization seemed to accurately reflect thought processes, they were somewhat incomplete. So in cases where think-alouds are shown not to have been reactive, researchers should next examine the question of veridicality to determine how accurate and complete the reporting was.

CONCLUSIONS

On the basis of the body of research on reactivity in SLA, it is possible to draw some conclusions. The findings of the meta-analysis detailed in Bowles (2010) have important implications for researchers from a variety of different theoretical perspectives. Crucially, a variety of factors seem to play a role in determining whether or not verbalization is reactive on a given task.

Based on the evidence available to date, it appears that in some cases verbal reports can indeed be a source of learning, as sociocultural theory would suggest. For instance, Bowles (2010) reported that participants who thought aloud nonmetacognitively scored moderately higher ($d = 0.67$) on receptive tests of form learning than did participants who completed the same task silently. This result suggests that verbalizing facilitated form learning, at least as measured by

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

receptive tests. On tests of productive form learning, however, thinking aloud had a small, detrimental effect ($d = -0.12$).

By the same token, the small effect sizes of $d \leq 0.5$ in 81% of the effect size calculations indicate that any posttest differences between silent and think-aloud groups tend to be small. In 86% of the effect size calculations, the 95% confidence interval overlaps zero, indicating that the *d* value is not significantly different from zero. This finding suggests that verbal reports can be used as a data collection tool. The results of the meta-analysis in Bowles (2010) stress the importance of including a (silent) control group in the design of any study using verbal reports as a data collection tool.

Since the first reactivity study in SLA (Leow & Morgan-Short, 2004) was conducted less than a decade ago, there are undoubtedly more questions than answers about verbal reports and reactivity. This safeguard allows the researcher to determine whether verbalizing caused significantly different task performance compared to silent participants. The findings also highlight the need for more fine-grained research to determine under what circumstances verbalizations are likely to accurately reflect processing and in what circumstances they are not.

REFERENCES

Anderson, M. (1985). Some evidence on the effect of verbalization on process: A methodological note. *Journal of Accounting Research, 23*, 843–852.

Baumann, J. F., Jones, L. A., & Seifert Kessell, N. (1993). Using think alouds to enhance children's comprehension monitoring abilities. *Reading Teacher, 47*, 184–193.

Berg, H. P. (2002). *Remediating cognitive perspective-taking in children with autism.* Unpublished doctoral dissertation, Columbia University Teachers College, New York.

Bernardi, L., Wdowczyk-Szulc, J., Valenti, C., Castoldi, S., Passino, C., Spadacini, G., et al. (2000). Effects of controlled breathing, mental activity, and mental stress with or without verbalization on heart rate variability. *Journal of the American College of Cardiology, 35*, 1462–1469.

Biehal, G., & Chakravarti, D. (1989). The effects of concurrent verbalization on choice processing. *Journal of Marketing Research, 26*, 84–96.

*Bowles, M. (2008). Task type and reactivity of verbal reports in SLA: A first look at a task other than reading. *Studies in Second Language Acquisition, 30*, 359–387.

Bowles, M. (2010). *The think-aloud controversy in language acquisition research*. New York: Routledge.

*Bowles, M., & Leow, R. P. (2005). Reactivity and type of verbal report in SLA research methodology: Expanding the scope of investigation. *Studies in Second Language Acquisition, 27*, 415–440.

Box, J. A. (2002). Guided writing in the early childhood classroom. *Reading Improvement, 39*, 111–113.

Bozarth, J. (1970). Verbal protocol patterns of college dormitory counselors. *Counselor Education and Supervision, 10*, 23–29.

Breetvelt, I. (1994). Relations between writing processes and text quality: When and how? *Cognition and Instruction, 12*, 103–123.

Brinkman, J. A. (1993). Verbal protocol accuracy in fault diagnosis. *Ergonomics, 36*, 1381–1397.

Carrell, P. L. (1989). Metacognitive awareness and second language reading. *Modern Language Journal, 73*, 121–134.

Chamot, A. U., & El Dinary, P. B. (1999). Children's learning strategies in language immersion classrooms. *Modern Language Journal, 83*, 319–338.

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

Chan, R. C., Hoosain, R., & Lee, T. M. (2002). Talking while performing a task: A better attentional performance in patients with closed head injury? *Journal of Clinical Experimental Neuropsychology, 24*, 695–704.
Cohen, A. D. (1986). Mentalistic measures in reading strategy research: Some recent findings. *English for Specific Purposes, 5*, 131–145.
Cohen, A. D. (1987). Recent uses of mentalistic data in reading strategy research. *Revista de Documentação de Estudos em Lingüística Teorica e Aplicada, 3*, 57–84.
Cohen, A. D., & Cavalcanti, M. C. (1987). Viewing feedback on compositions from the teacher's and the student's perspective. *ESPecialist, 16*, 13–28.
Cushman, D. (2002). From scribbles to stories. *Instructor, 111*, 32–33.
Davis, J., & Bistodeau, L. (1993). How do L1 and L2 reading differ? Evidence from think aloud protocols. *Modern Language Journal, 77*, 459–472.
Davis, J. H., Carey, M. H., Foxman, P. N., & Tarr, D. B. (1968). Verbalization, experimenter presence, and problem solving. *Journal of Personality and Social Psychology, 8*, 299–302.
Earthman, E. A. (1992). Creating the virtual work: Readers' processes in understanding literary texts. *Research in the Teaching of English, 26*, 351–384.
Ellis, R. (2004). The definition and measurement of L2 explicit knowledge. *Language Learning, 54*, 227–275.
Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.
Folger, T. L. (2001). Readers' parallel text construction while talking and thinking about the reading process. *Dissertation Abstracts International, 62* (4), 1329A. (UMI No. 3012966)
Fowler, L. P. (1997). Clinical reasoning strategies used during care planning. *Clinical Nursing Research, 6*, 349–361.
Fresch, M. J., Wheaton, A., & Zutell, J. B. (1998). Thinking aloud during spelling word sorts. *National Reading Conference Yearbook, 47*, 285–294.
Friedman, P., & Mulhern, S.T. (1976). Relationship of clinician feedback to child-initiated verbalization during language training. *Journal of Communication Disorders, 9*, 289–299.
Gagné, R. H., & Smith, E. C. (1962). A study of the effects of verbalization on problem solving. *Journal of Experimental Psychology, 63*, 12–18.
Gass, S., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Erlbaum.
Gordon, C. J. (1990). Modeling an expository text structure strategy in think alouds. *Reading Horizons, 31*, 149–167.
Greenwood, J., & King, M. (1995). Some surprising similarities in the clinical reasoning of expert and novice orthopaedic nurses: Report of a study using verbal protocols and protocol analysis. *Journal of Advanced Nursing, 22*, 907–913.
Herwig, A. (2003). Plurilingual lexical organisation: Evidence from lexical processing in L1-L2-L3-L4 translation. In J. Cenoz, B. Hufeisen, & U. Jessner (Eds.), *Cross linguistic influence in third language acquisition: Psychological perspectives* (pp. 115–137). Clevedon, UK: Multilingual Matters.
Hosenfeld, C. (1976). Learning about learning: Discovering our students' strategies. *Foreign Language Annals, 9*, 117–129.
Hosenfeld, C. (1977). A preliminary investigation of the reading strategies of successful and nonsuccessful second language learners. *System, 5*, 110–123.
Hosenfeld, C. (1979). Cindy: A learner in today's foreign language classroom. In W. Borne (Ed.), *The foreign language learner in today's classroom environment* (pp. 53–75). Montpelier, VT: Northwest Conference on the Teaching of Foreign Languages.
Hosenfeld, C. (1984). Case studies of ninth grade readers. In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 231–249). London, UK: Longman.
Hu, G. (2002). Psychological constraints on the utility of metalinguistic knowledge in second language production. *Studies in Second Language Acquisition, 24*, 347–386.
Hughes, J., & Parkes, S. (2003). Trends in the use of verbal protocol analysis in software engineering research. *Behaviour & Information Technology, 22*, 127–141.

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

Jourdenais, R. (2001). Cognition, instruction, and protocol analysis. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 354–375). Cambridge, UK: Cambridge University Press.

Karsenty, L. (2001). Adapting verbal protocol methods to investigate speech systems use. *Applied Ergonomics, 32*, 15–22.

Keck, C. M., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). Investigating the empirical link between task-based interaction and acquisition: A quantitative meta-analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91–131). Amsterdam: John Benjamins.

Kern, R. G. (1994). The role of mental translation in second language reading. *Studies in Second Language Acquisition, 16*, 441–461.

Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *The American Journal of Psychology, 113*, 387–404.

*Lass, U., Klettke, W., Lüer, G., & Ruhlender, P. (1991). Does thinking aloud influence the structure of cognitive processes? In R. Schmid & D. Zambarbieri (Eds.), *Oculomotor control and cognitive processes* (pp. 385–396). New York: North-Holland.

Lee, S.-K., & Huang, H.-T. (2008). Visual input enhancement and grammar learning: A meta-analytic review. *Studies in Second Language Acquisition, 30*, 307–331.

Leow, R. P. (1997). Attention, awareness, and foreign language behavior. *Language Learning, 47*, 467–505.

Leow, R. P. (1998a). The effects of amount and type of exposure on adult learners' L2 development in SLA. *Modern Language Journal, 82*, 49–68.

Leow, R. P. (1998b). Toward operationalizing the process of attention in SLA: Evidence for Tomlin and Villa's (1994) fine-grained analysis of attention. *Applied Psycholinguistics, 19*, 133–159.

Leow, R. P. (1999). The role of attention in second/foreign language classroom research: Methodological issues. In F. M.-G. J. Gutiérrez-Rexach (Ed.), *Advances in Hispanic Linguistics: Papers from the 2nd Hispanic Linguistics Symposium* (pp. 60–71). Somerville, MA: Cascadilla Proceedings Project.

Leow, R. P. (2000). A study of the role of awareness in foreign language behavior: Aware versus unaware learners. *Studies in Second Language Acquisition, 22*, 557–584.

Leow, R. P. (2001a). Attention, awareness, and foreign language behavior. *Language Learning, 51*(Suppl. 1), 113–155.

Leow, R. P. (2001b). Do learners notice enhanced forms while interacting with the L2? An online and offline study of the role of written input enhancement in L2 reading. *Hispania, 84*, 496–509.

*Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition, 26*, 35–57.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 407–452). Oxford, UK: Oxford University Press.

Masgoret, A. M., & Gardner, R. C. (2003). Attitude, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning, 53*, 123–163.

*Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J. R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 1083–1100.

Midanik, L. T., & Hines, A. M. (1991). "Unstandard" ways of answering standard questions: Protocol analysis in alcohol survey research. *Drug and Alcohol Dependence, 27*, 245–252.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*, 417–528.

Norris, J. M., & Ortega, L. (Eds.). (2006). *Synthesizing research on language learning and teaching*. Amsterdam: John Benjamins.

Payne, J. W., Braunstein, M. L., & Carroll, J. S. (1978). Exploring predecisional behavior: An alternative approach to decision research. *Organizational Behavior and Human Performance, 22*, 17–44.

Piolat, A., & Olive, T. (2000). Comment étudier le coût e le déroulement de la rédaction de textes? La méthode de la triple tâche: Un bilan méthodologique [How can the process and cost of writing texts be studied? The triple task methodology]. *L'Annee Psychologique, 100*, 465–502.

Polio, C., & Chiu, S. C.-H. (2007, April). *Reactivity, veridicality, and language choice in L2 writing concurrent verbal protocols.* Paper presented at the American Association of Applied Linguistics Conference, Costa Mesa, CA.

*Polio, C., & Wang, J. (2005, October). *Another look at the reactivity of concurrent verbal protocols in second language reading research.* Paper presented at the Second Language Research Forum, New York.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Erlbaum.

Pritchard, R. (1990). The effects of cultural schemata on reading processing strategies. *Reading Research Quarterly, 25*, 273–295.

Qi, D., & Lapkin, S. (2001). Exploring the role of noticing in a three-stage second language writing task. *Journal of Second Language Writing, 10*, 277–303.

Rhenius, D., & Deffner, G. (1990). Evaluation of concurrent thinking aloud using eye-tracking data. In M. E. Wiklund (Ed.), *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 1265–1269). Santa Monica, CA: Human Factors Society.

Robertson, B. (1995). Why think along? Using "think alouds" in the classroom. *State of Reading, 2*, 19–22.

Robinson, K. M. (2001). The validity of verbal reports in children's subtraction. *Journal of Educational Psychology, 93*, 211–222.

Rosa, E., & Leow, R. P. (2004a). Awareness, different learning conditions, and L2 development. *Applied Psycholinguistics, 25*, 269–292.

Rosa, E., & Leow, R. P. (2004b). Computerized task-based exposure, explicitness, type of feedback, and Spanish L2 development. *Modern Language Journal, 88*, 192–216.

Rosa, E., & O'Neill, M. (1999). Explicitness, intake, and the issue of awareness. *Studies in Second Language Acquisition, 21*, 511–556.

Rosenthal, M. C. (1994). "The fugitive literature." In H. Cooper & L. Hedges (Eds.), *Handbook of research synthesis*. New York: Russell Sage Foundation.

*Rossomondo, A. E. (2007). The role of lexical temporal indicators and text interaction format in the incidental acquisition of the Spanish future tense. *Studies in Second Language Acquisition, 29*, 39–66.

Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar: A meta-analysis of the research. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133–164). Amsterdam: John Benjamins.

*Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory and Cognition, 17*, 759–769.

*Sachs, R., & Polio, C. (2007). Learners' uses of two types of written feedback on an L2 writing revision task. *Studies in Second Language Acquisition, 29*, 67–100.

*Sachs, R., & Suh, B. R. (2007). Textually enhanced recasts, learner awareness, and L2 outcomes in synchronous computer-mediated interaction. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 197–227). Oxford, UK: Oxford University Press.

*Sanz, C., Lin, H.-J., Lado, B., Bowden, H. W., & Stafford, C. A. (2009). Concurrent verbalizations, pedagogical conditions, and reactivity: Two CALL studies. *Language Learning, 59*, 33–71.

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press

Scardamalia, M. (1984). Teachability of reflective processes in written composition. *Cognitive Science: A multidisciplinary journal of artificial intelligence, 8*, 173–190.

Stratman, J. F., & Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols. In P. Smagorinsky (Ed.), *Speaking about writing: Reflections on research methodology* (pp. 89–112). London: Sage.

Walczyk, J., Marsiglia, C. S., Bryan, K. S., & Naquin, P. J. (2001). Overcoming inefficient reading skills. *Journal of Educational Psychology, 93*, 750–757.

Wilhelm, J. (2001). Getting kids into the reading game: You gotta know the rules. *Voices from the Middle, 8*, 25–36.

Witte, S. P., & Cherry, R. D. (1994). Think-aloud protocols, protocol analysis, and research design: An exploration of the influence of writing tasks on writing processes. In P. Smagorinsky (Ed.), *Speaking about writing: Reflections on research methodology* (pp. 20–54). London: Sage.

Woodfield, H. (2008). Problematising discourse completion tasks: Voices from verbal report. *Evaluation and Research in Education, 21*, 43–69.

Yamashita, J. (2002). Reading strategies in L1 and L2: Comparison of four groups of readers with different reading ability in L1 and L2. *ITL, Review of Applied Linguistics, 135–136*, 1–35.

*Yoshida, M. (2008). Think-aloud protocols and type of reading task: The issue of reactivity in L2 reading research. In M. Bowles, R. Foote, S. Perpiñán, & R. Bhatt (Eds.), *Selected proceedings of the 2007 Second Language Research Forum* (pp. 199–209). Somerville, MA: Cascadilla Proceedings Project.

Zellermayer, M., & Cohen, J. (1996). Varying paths for learning to revise. *Instructional Science, 24*, 177–195.

https://doi.org/10.1017/S0267190510000036 Published online by Cambridge University Press