University of Alberta

Verbal Report Validity and Children's Subtraction

By

Katherine MacLeod Robinson

A thesis submitted to the Faculty of Graduate Studies and Research

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Psychology

Edmonton, Alberta

Fall 1999

0-612-46911-5

University of Alberta

Library Release Form

Name of Author: Katherine MacLeod Robinson

Title of Thesis: Verbal Report Validity and Children's Subtraction

Degree: Doctor of Philosophy

Year this Degree Granted: 1999

15 Selkirk Road West
Lethbridge, Alberta
T1K 4N4

July 16, 1999

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled "Verbal Report Validity and Children's Subtraction" in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dr. Jeffrey Bisanz

Dr. C. Donald Heth

Dr. Connie Varnhagen

Dr. Patrick Lemaire

Dr. Stephen Norris

Dr. Edward Cornell

Date: 12 July 1999

Abstract

Verbal reports have frequently been used as measures of cognitive development in recent years. However, the appropriateness of verbal reports has often been a controversial topic. Study of verbal report validity has been infrequent and results have been mixed. In this study, the validity of verbal reports in children's subtraction was investigated. Arithmetic problems lend themselves well to verbal reporting and in the process, the development of children's subtraction skills could also be investigated. Children in Grades 1, 3, and 5 were asked to solve a set of subtraction problems and were placed in the No Report, Retrospective Report, or Concurrent Report Condition. Students in all grades and in both report conditions were able to provide veridical reports of their solution strategies and the instruction to verbal report had little effect on their performance. However, in general, retrospective reports were recommended as the ideal type of verbal report for students in all grades although students in Grade 1 also performed well with concurrent reports. Based on the valid verbal reports, the development of children's subtraction solution strategies was also examined. While students in Grade 1 relied primarily on counting and Grade 5 students most often used retrieval to solve the subtraction problems, Grade 3 students most frequently used a third type of strategy, special tricks. Grade 3 students also used more diverse strategies than the other students suggesting that Grade 3 students are in a state of transition.

Acknowledgements

Jeff Bisanz has guided my way for several years now and has not only provided support to me in my work but also in my personal life. His care and concern have been especially invaluable during these last stages of the degree. My only regret is that he has been unable to break me of my habit of having research and studies do things that only people can do.

Connie Varnhagen has been incredibly helpful to me throughout my graduate degrees. She has always been easily accessible, a great listener, and has provided the happy medium between mentor and practical onlooker.

Don Heth has been a supportive committee member ever since the beginning of my days at the University of Alberta. He has always provided astute insights into my work and somehow always seems to be able to find a simple solution to especially knotty problems.

My time at the University of Alberta has been enriched by several valuable friends. Connie Korpan has been a great friend for many years and has lightened many moments with her good humour and caustic wit. I can think of no one I would rather travel to conferences with. Juliette Sandwell Klein and I have shared many common interests both within and outside of our graduate lives. Thanks to Juliette I have become passably inept at ballet and I have found someone who will listen to me rave interminably about the wonders of my animals. Finally, Nat Hardy has not only been a wonderful friend to my husband, but he has also been incredibly generous in giving me of his time and energy.

Barb McLeod of Red Deer College helped guide me through my first academic position and in the process shared her delightfully dry sense of humour with me. Kathy MacKinnon, whom I first met at age 13, has been a dynamic, enthusiastic, and speed-walking friend who has provided constant encouragement and support in all my endeavours.

A great debt is also due to my incredibly supportive and loyal parents, Ruby and Stuart Robinson. For all of these years they have been a weekly long-distance source of comfort, love, and encouragement. They have had constant, unconditional faith in me and my abilities to get this degree done, even when I myself had lost faith.

My husband, Christian Riegel, and I have journeyed together through this sometimes excruciating process of completing a doctoral degree. During all the highs and lows, Chris has been a steady source of comfort and love. He has unwaveringly supported my goals and dreams, both academic and non-academic, for over 11 years. During these years, Chris has taken time out of his studies to help me collect the data for this degree, has edited and critiqued my work, has patiently listened to me complain, has tolerated me in all my moods, has made me laugh, and has provided me with so much joy and love that I consider myself incredibly lucky to have found such a wonderful life companion.

Finally, the warm, furry support of my other life companions, both equine and canine, has been immeasurably important to me. Nothing can quite compare to a good dose of fuzz therapy and unconditional love.

This degree was supported by a Social Sciences and Humanities Research Council of Canada Doctoral Fellowship.

Table of Contents

List of Tables

List of Figures

Verbal Report Validity and Children's Subtraction

The study of children's mathematical problem solving is a rapidly growing area of research within cognitive development. The solution strategies that children use on simple arithmetic problems such as addition and subtraction provide a window to how cognitive skills and strategies evolve and change across development. Indeed, Siegler (1996; Siegler & Jenkins, 1989) has formulated a compelling theory of cognitive development with foundations that are at least partially based on how children solve arithmetic problems. More specifically, several models of children's solution strategies in arithmetic have been proposed (e.g., Groen & Parkman, 1972; Siegler & Shipley, 1995; Siegler & Shrager, 1984). Traditionally, accuracy and latency data have been used by researchers to infer what strategies participants use to solve different types of mathematical problems (e.g., Svenson, 1975; Svenson & Hedenborg, 1979; Woods, Resnick, & Groen, 1975), and models of children's arithmetic are based on these inferences.

In recent years there has been a growing trend to use verbal reports to assess solution processes on arithmetic problems (e.g., Boulton-Lewis, 1993; Carr & Jessup, 1997; Kerkman & Siegler, 1997; LeFevre, Sadesky, & Bisanz, 1996, Robinson, 1993; Siegler, 1987a, 1989b; Svenson, Hedenborg, & Lingman, 1976). Siegler (1987a, 1989b) has provided compelling evidence for the use of verbal reports in the investigation of strategy use in both simple addition and simple subtraction. In separate studies of addition and subtraction, Siegler (1987a, 1989b) demonstrated that if only students' accuracy and latency data were used to determine what solution strategy the children were using, the pattern of inferred strategies was markedly different than the pattern established from their verbal reports of solution strategies. Siegler found converging evidence for the validity of the strategies reported by the students and thus concluded that inferences based on verbal reports were more likely to be accurate than inferences based on chronometric data. For example, from chronometric data, previous researchers had hypothesized that children were predominantly using one type of strategy to solve addition or subtraction problems (e.g., Groen & Parkman, 1972; Svenson, 1975; Svenson & Hedenborg, 1979; Woods, Resnick, & Groen, 1975), but the verbal reports the children were providing indicated a range of strategies.

The findings by Siegler (1987a, 1989b) have provided the basis for an increased use of verbal reports, especially in mathematical cognition (e.g., Boulton-Lewis, 1993; Carr & Jessup, 1997; Leiken & Zaslavsky, 1997) and other areas of cognitive development (e.g., Steffler, Varnhagen, Friesen, & Treiman, 1998). However, there are several validity issues in the use of verbal reports as data that still remain to be addressed (Ericsson & Simon, 1980, 1984, 1993; Cooney & Ladd, 1992; Fidler, 1983; Nisbett & Wilson, 1977; Norris, 1991; Payne, 1994; Russo, Johnson, & Stephens, 1989; Wilson, 1994). Ericsson and Simon have proposed a well-grounded theoretical framework for the use of verbal reports, but actual testing of the theory has been minimal and unsatisfactory (Austin & Delaney, 1998; Critchfield & Epting, 1998; Russo, Johnson, & Stephens, 1989). However, the advantages of verbal reports are many. As Crutcher (1994) stated: "within the domain of psychology, the use of verbal reports promises to be one of the most important methodological developments of the past 25 years" (p.244).

Therefore, the first goal of this study, further delineated in the next section of the introduction, was to investigate the issue of the validity of verbal reports as a measure of cognitive development. Pressley and Afflerbach most succinctly and eloquently summed up the issue of verbal reports: "What is needed is systematic study of these issues, not continued conjecturing about them" (1995, p. 129). The second goal of this study, as described in the second section of this introduction, was to investigate the cognitive development of arithmetic skills in subtraction problems. Siegler (1987a, 1989b) has compellingly demonstrated the importance of collecting verbal reports to provide crucial information on how children solve arithmetic problems. Furthermore, subtraction problems provide a useful focus for investigating cognitive development. However, the utility of verbal reports as a valid source of data must first be established before they can be considered an appropriate measure of children's subtraction problem-solving strategies.

Verbal Reports in Research

The trend of using verbal reports as data can be found in many diverse areas of research such as decision making (Biggs, Rosman, & Sergenian, 1993), frequency estimation (Brown, 1995), marketing (Biehal & Chakravarti, 1989), accounting (Anderson, 1985), clinical research on psychopathology (Brewin, Andrew, & Gotlib, 1993), product design (Denning, Hoiem, Simpson, & Sullivan, 1990), academic skills (Klahr, Fay, Dunbar, 1993; Lawson & Rice, 1987), behaviour analysis (Critchfield, 1993; 1996; Critchfield & Epting, 1998), assessment (Ericsson, 1987), eyewitness testimony (Christianson, 1992), reading (Pressley & Afflerbach, 1995), text comprehension (Laszlo, Meutsch, & Viehoff, 1988), memory (Anders, 1971), and problem solving (Robertson, 1990; Stinessen, 1985). This increasing trend to use verbal reports as data in many different research areas may be linked to the work of Ericsson and Simon (1980, 1984, 1993) who posited the merits of asking subjects to report their thoughts either while, or just after, performing a task. Their work has been extremely influential and has been touted as the bible of cognitive science methodology (Lesgold, 1993).

Validity Issues for Verbal Reports

Ericsson and Simon (1980, 1984, 1993) set out guidelines under which verbal reports should be valid forms of data. Nisbett and Wilson (1977) had compellingly demonstrated that participants were often unable to accurately report on the reasons for their decisions. Ericsson and Simon theorized that if verbal reports are requested for the motives or reasons for participants' thinking, then the reports are likely to be faulty because reasons and motives are rarely part of short-term memory, and thus participants are being asked to reconstruct motives and reasons rather than to remember them. Ericsson and Simon clearly specified that verbal reports will only be valid if they are reports of the contents of short-term memory (although other research perspectives are not quite as strict about this viewpoint as the information processing perspective, e.g., Hayes, White, & Bissett, 1998). Thus, according to Ericsson and Simon, participants should only be asked to provide verbal reports of the contents of their short-term memory, which might include thoughts, actions, or feelings rather than the reasons for their thoughts, actions, or feelings.

Ericsson and Simon (1980, 1984, 1993) also proposed that concurrent reports, or thinking aloud or talking aloud, be used as much as possible so that processing and verbalization occur at the same time. In concurrent reporting, no thought, feeling, or action should be omitted because the participant has had time to forget that thought, feeling, or action. However, not all tasks are suitable to concurrent verbal reports. For example, extremely rapid tasks occur so quickly (e.g., slapping at a mosquito that is stinging you, or answering the question "2 - 1 = ?") that concurrent reporting during the process will interfere with the process itself. Another task that might not be suitable for concurrent reporting is one that has such a very high cognitive load that the added task of concurrently reporting interferes with the primary task itself.

If subjects are asked for a verbal report after completing a task, that is, for retrospective reports, then subjects presumably will only be able to report the contents of their short-term memory after the task is completed. This type of reporting has two implications. First, if the task is too difficult or too lengthy, information will be lost. Second, if the task is automatized, such as the rapid retrieval of the answer to 2 x 3, then subjects should not be able to report how they solved the problem, as the solution process was too rapid and never entered short-term memory. However, people should be able to report that the answer was directly retrieved from long-term memory.

Overall, Ericsson and Simon's (1980, 1984, 1993) prognosis for the use of verbal reports as data was extremely positive, and they concluded that "it is now time for verbal reports to reassume their position as a rich source of data, combinable with other data, that can be of greatest value in providing an integrated and full account of cognitive processes and structures" (p. 373). Though Ericsson and Simon's recommendations are based on an extensive review of research using verbal reports as data in many areas, their work remains a largely untested theory that is not based on well-researched findings.

Ericsson and Simon cited supporting evidence but the evidence is not necessarily as conclusive as its presentation would suggest. Indeed, Ericsson and Simon have in some instances simultaneously cited supporting evidence for the validity of verbal reports as data while ignoring contrary evidence from within the same study (e.g., Norris, 1990). As well, some of the studies cited by Ericsson and Simon have methodological flaws that should discount their value in supporting the theory of verbal reports (e.g., Biehal & Chakravarti, 1989). Thus, although verbal reports have been used with apparent success in a wide variety of areas, including children's subtraction, their validity is not as clear as Ericsson and Simon posit. Careful investigations of their use must be conducted.

Validity of Verbal Reports: Cause for Concern?

Why should researchers be concerned about whether or not verbal reports are a valid source of data? Theoretical implications drawn from research are always constrained by the methods that are utilized in the research. If inappropriate methods are being used, then resulting theories may yield grossly inaccurate reflections of the questions being investigated (e.g., Siegler, 1987a, 1989b). Therefore, before using verbal reports as data, researchers need to know whether, indeed, verbal reports are an appropriate measure.

Ericsson and Simon's (1980, 1984, 1993) work appears to be well accepted and researchers using verbal reports often do not acknowledge any issues regarding validity (e.g., Boulton-Lewis, 1993; Hamann & Ashcraft, 1985; Harel, Behr, Post, & Lesh, 1992; Kliman, 1987; Siegler, 1988; Toward & Kerr, 1994; Williams & Santos-Williams, 1980). Researchers who do refer to the validity issue for verbal reports allude to the issue briefly—usually by citing Ericsson and Simon's work as strong evidence that verbal reports are a valid form of data (e.g., Cooney & Swanson, 1990; Cooney, Swanson, & Ladd, 1988; McGilly & Siegler, 1989; Short, Cuddy, Friebert, & Schatschneider, 1990; Short, Schatschneider, Cuddy, Evans, Dellick, & Basili, 1991).

Not enough is known about when and, indeed, whether verbal reports are valid sources of data, despite the important groundwork by Ericsson and Simon (1980, 1984, 1993). Generally, even detractors of verbal reports (e.g., Wilson, 1994) who previously suspected the use of verbal reports as data (Nisbett & Wilson, 1977) now concede that research should focus on when and not on whether verbal reports are valid (Smith & Miller, 1978).

There are two main issues concerning verbal report validity: reactivity and veridicality (Russo, Johnson, & Stephens 1989). Performance on a task is said to be reactive if the act of giving a verbal report affects or changes performance on that task in any way. For example, being asked to verbally report while solving a mathematical word problem may slow down the solution process for most people. Ericsson and Simon (1984, 1993) have documented many examples of the slowing down of task completion times when concurrently reporting. Therefore concurrent reports are, according to Ericsson and Simon, expected to have altered solution latencies because of the time needed to verbalize. Thus, when participants are concurrently reporting, changes in measures other than solution times, such as reported solution strategies or accurate performance, would be more appropriate for assessing task reactivity. However, alterations in solution times for participants who are retrospectively reporting would be indicative of a reactive effect of verbal reporting on task performance.

Verbal reports are veridical to the extent that they accurately reflect the contents of short-term memory. If verbal reports are nonveridical, then the implication is that the verbal reports are not measuring what they are supposed to measure. Verbal reports can be nonveridical in two different ways. First, errors of commission can occur when subjects report thoughts, actions, or feelings that were not actually experienced. For example, in an arithmetic task, participants may report using retrieval when they actually used a counting procedure, perhaps because the subjects felt they *should* be using retrieval and so reported that they had. In this case, participants are reporting fictional events, and any conclusions or theories based on these reports will be erroneous. Second, errors of ommission can occur when subjects do not give a complete report of the thoughts, actions, or feelings they experienced. For example, participants may report using a counting procedure on an arithmetic problem but fail to report that they also used a special counting trick (e.g., counting by threes) within that counting procedure. In this instance, verbal reports would not be erroneous per se but would be incomplete, and thus conclusions or theories based on the data would not be full accounts of the phenomena under investigation. However, with other types of data, such as reaction time and accuracy, the accounts would not be complete either, so the possibility of incompleteness is not unique to verbal reporting. Indeed, as Critchfield and Epting (1998) posited, "all measurement necessarily distorts, in that data records are imperfect reproductions of target events, so that information is both lost and altered during measurement" (p.67).

As well, even if the verbal reports are accurate representations of participants' thought processes (that is, veridical), the performance of the task under investigation may be altered by the act of verbal reporting such that the study is no longer an investigation of what the researcher intended to study (that is, reactive). Russo, Johnson, and Stephens therefore state that "concerns about reactivity naturally take precedence over lack of veridicality because there is little point to testing whether or not a report is veridical if verbalization has already changed the primary process being reported" (1989, p. 760). Interestingly, several studies that have taken into careful consideration the validity of verbal reports have concentrated solely on the veridicality issue and have ignored the reactivity issue (Siegler, 1987a, 1989b; Steffler, Varnhagen, Friesen, & Treiman, 1998).

Advantages of Verbal Reports

There are several important advantages and reasons for the use of verbal reports, even if there are reasons to doubt their validity. First, the amount of information yielded by this method is often much different than that of traditional forms of data, such as latency and accuracy. For example, Siegler (1987a, 1989b) and Robinson (1993) found that verbal reports yielded information about an unexpected arithmetic strategy that would not have been noted had they relied solely on accuracy and chronometric data. Robinson found a new shortcut strategy, one that was not under investigation, that was often reported by subjects as a strategy of choice for solving certain arithmetic problems. This new shortcut would never have been identified using only accuracy or latency data.

Second, sometimes the information from verbal report data demonstrates that the use of traditional methodologies can lead to incorrect theories. An example of this advantage is the work by Siegler (1987a, 1989b) who found that chronometric data alone obscured the various strategies that children use on simple addition and subtraction problems. When verbal reports and chronometric data were used in conjunction, a completely different theory of strategy use emerged. Therefore, the use of verbal reports can be valuable in the development of theories (Crutcher, 1994) and for the demonstration of the erroneous theories that can arise from more traditional types of data. Green (1995) stated: "protocol analysis has much to contribute to the development and evolution of theories across a wide ranges of tasks" (p. 129). Overall, verbal reports necessitate more stringent criteria for the development of theories (Bellezza, 1986) because investigators must account for both verbal report data and other types of overt data in their theories.

Third, as more research is being focused on individual differences, verbal report data are an appropriate tool for assessing these differences. Siegler (1987a, 1988, 1989a, 1989b; Siegler & Crowley, 1991) demonstrated the utility of verbal reports for assessing individual subjects' strategies for solving different types of arithmetic problems. For example, he found that he was able to differentiate three types of problem solvers based primarily on verbal reports of their solution processes: good students, not-so-good students, and perfectionists.

Fourth, there is a growing realization that not only are there important individual differences in strategy use but also intra-individual differences in strategy use (e.g., Siegler, 1989a). Verbal reports are often the easiest method for collecting information involving these types of differences. Subjects often use different solution strategies on the same type of problems and indeed, sometimes use different strategies on the same problem when it is presented more than once.

Fifth, verbal reports are an important tool for measuring or examining process.

That is, verbal reports complement both the information processing viewpoint and the area of assessment with their greater emphasis on process over product, whereas more traditional types of data are most appropriate for assessing product (Crutcher, 1994; Ericsson, 1987; Green, 1995; Siegler, 1989b). That is, verbal reports can be utilized to assess problems that a student may have with the strategy that they are using to try and solve a problem (e.g., Lawson & Rice, 1987; Short, Cuddy, Friebert, & Schatschneider, 1990) rather than simply getting information about whether the student got the correct answer.

Toward and Kerr (1994) stated that "the collection and analysis of verbal reports both concurrent and retrospective of task performance provide the investigator a rich source of information" (p.514). As verbal reports become an increasingly common and ideal source of data in many areas of research, including the area of mathematical development, it is essential to establish whether confidence about research and theory based on verbal reports is warranted. Thus, one of the goals of this dissertation was to examine specific conditions under which verbal reports might be appropriate. The task used to examine verbal report validity is presented next followed by the conditions under which verbal reports might be more or less appropriate as data.

Investigating the Validity of Verbal Reports in the Development of Subtraction Skills

In this study, arithmetic problems, specifically subtraction problems, were chosen as the primary task, in part because they lend themselves well to tests of verbal report validity. Researchers need to know under what conditions verbal reports are an appropriate tool. Clear advantages to using verbal reports and clear substantive validity issues exist and are associated with the use of verbal reports. Verbal reports have been used to investigate many diverse tasks in many diverse research areas. However, assessing the validity of verbal reports is most easily done on certain types of tasks, such as tasks that have a finite set of correct sequences or a finite number of solution paths for getting the correct response (Austin & Delaney, 1998). Critchfield and Epting (1998) stated that "well-defined tasks provide obvious benchmarks against which verbal reports can be compared, and thus offer a means of estimating validity" (p. 70). Both Austin and Delaney and Critchfield and Epting cited arithmetic problems as obvious examples of well-defined tasks. Indeed, researchers have suggested the use of and/or have used arithmetic problems for getting participants to practice verbal reporting before introducing the primary task (Ericsson & Simon, 1984, 1993; Hayes, White, & Bissett, 1998; Wulfert, Dougher, & Greenway, 1991).

The investigation of children's problem solving strategies in arithmetic comprises a well-researched area with a history of successful use of verbal reports. Other advantages to using simple arithmetic problems for investigating the validity of verbal reports, include the following:

> Children in the process of learning arithmetic use a wide variety of strategies to solve what for them are real problems. The strategies are crisp and easily described, children can accurately report which strategy they used on each trial, and the distribution of strategies that they use changes substantially over a period of several years (Siegler, 1996, p. 61).

As well, because the area has been well documented, an additional advantage to using an arithmetic task is that specific patterns of accuracy and latency are known to be associated with specific solution strategies (e.g., Siegler, 1987a, 1989b). Thus, convergent types of data are available to assess the validity of subjects' verbal reports—a rare advantage that exists in this research area. Veridicality can then be assessed by matching accuracy and latency data with the verbal report data.

Verbal Report Validity and Type of Verbal Report

Arithmetic problems also lend themselves to the investigation of other conditions under which verbal reports may or may not be valid forms of data. The two main types of verbal reports, concurrent and retrospective, have been successfully used in the study of how arithmetic skills develop (Boulton-Lewis, 1993; Flaherty, 1974; Ilg & Ames, 1951; Lawson & Rice, 1987; Secada, Fuson, & Hall, 1983; Svenson, 1975).

Concurrent reports are most highly recommended by Ericsson and Simon (1980, 1984, 1993) because if participants report their thought processes while working on a task, they are less likely to omit or fabricate. That is, participants are simply reporting the contents of short-term memory as they work through a task (Green, 1995). However, generally concurrent reporting also slows the task because verbalizing the contents of short-term memory takes additional time (Ericsson & Simon, 1980, 1984, 1993).

Retrospective reports are also considered appropriate, especially if "the problems are all similar, solved quickly, and presented in great numbers" (Svenson, 1989, p.67). Crutcher (1994) also recommended the use of verbal reports on tasks similar to those Svenson posited but suggested that the retrospective reports be gathered as quickly as possible after the task is completed (see also Ericsson & Simon, 1980, 1984, 1993). This last recommendation stems from the basic assumption of Ericsson and Simon's theory that verbal reports are simply reports of the contents of short-term memory. If delays occur, then there is an increasing chance that the verbal reports will either have missing or erroneous information. However, because the verbal reports are being collected after the task is completed, they should be less likely to interfere with the problem solving process.

Little actual investigation has been done to assess which type of report is better, if either, although there are clear advantages and disadvantages to both types of reports. Some researchers have found both types of reports to be non-reactive (Norris, 1990; 1991), to be veridical (Steffler, Varnhagen, Friesen, & Treiman, 1998), to be both veridical and non-reactive (Biggs, Rosman, & Sergenian, 1993; Bowers & Snyder, 1990), or to be both non-veridical and reactive (Russo, Johnson, & Stephens, 1989).[1] Therefore, the question remains whether it makes a difference if the verbal reports are concurrent or retrospective and, if so, how much of a difference there is (see also Pressley & Afflerbach, 1995). Because both concurrent and retrospective verbal reports have a successful history of use in the area of mathematical development, asking students to solve subtraction problems provides an ideal task to test which, if either, report type is best.

Verbal Report Validity and Children

The characteristics of the participants under investigation was another condition that may affect the validity of verbal reports. Ericsson and Simon (1980, 1984, 1993) cited

---

[1] Although Cooney and Ladd (1992) investigated the validity of both types of verbal reports on an arithmetic task with children, their study had a number of serious flaws that made any conclusions about verbal report validity impossible.

Siegler's (1987a, 1989b) work as evidence for the validity of verbal reports but they did not specifically refer to the issue of whether there are differences in veridicality and reactivity as a function of age or cognitive maturity or the development of verbalization skills. This omission is of concern as verbal reports are becoming increasingly popular in all areas of cognitive developmental research (e.g., Pressley & Afflerbach, 1995; Siegler 1987a; 1989b; Steffler, Varnhagen, Friesen, & Treiman, 1998). The area of eyewitness testimony is an excellent example of why this is such an important issue. In this area there is a belief that children are much more prone to giving incomplete or erroneous testimony, although this assumption is not necessarily valid (Poole & White, 1991, 1993).

Within the area of mathematical development, though, the issue of whether or not verbal reports should be used with children has yielded contradictory evidence. Ginsburg, Kossan, Schwartz, and Swanson (1983), in their investigation of children's concurrent verbal reports on mathematics problems, concluded that verbal reports were not sufficient for gathering information about children's cognitive processes and should be reserved for exclusive use with adults. Whereas adults were sophisticated enough for concurrent reporting, children lacked the verbal skills, had poor comprehension, and sometimes were minimally cooperative such that the use of concurrent reports was not deemed appropriate. Hamann and Ashcraft (1985) asked children in Grades 1, 4, 7, and 10 to solve simple and complex addition problems. Subjects in Grades 1 and 4 gave interesting and informative delayed retrospective reports of their strategies, whereas the older students gave verbal reports that were relatively uninformative (note, though, that Hamann and Ashcraft were not arguing that the verbal reports were not veridical). This was probably due to the lack of problem difficulty for the older students and the high frequency of direct retrieval. There seems to be evidence that verbal reports may be inappropriate with younger, or cognitively more immature, subjects, but successful use of verbal reports with children has been demonstrated in other studies of arithmetic skills (Robinson, 1993; Siegler, 1987a, 1989b). As questions still remain about the use of verbal reports for children, this study examined the validity of verbal reports given by children of different ages.

The two main goals of this study were to (1) investigate the validity of verbal reports and (2) to investigate the development of arithmetic skills. Subtraction problems provide an ideal task for fulfilling both of the goals of this study. To achieve the first goal, the validity of verbal reports was investigated by asking students to solve a series of simple subtraction problems. The two most common types of verbal reports, concurrent and retrospective, were gathered and analyzed to determine which, if either, yielded veridical and non-reactive data. Finally, to determine whether verbal reports are an appropriate tool for non-adult participants, children in Grades 1, 3, and 5 were asked to provide reports of how they solved subtraction problems, either concurrently or retrospectively. To achieve the second goal of investigating the development of arithmetic skills, subtraction problems were used. In the next section research on the development of subtraction skills is presented and how this study provided the opportunity to contribute to the existing knowledge of how subtraction strategies and types of knowledge develop is delineated.

The Development of Subtraction Skills

Investigating the development of arithmetic skills has provided much information on children's cognitive development (Bisanz & LeFevre, 1990; Siegler, 1996, 1998). Using a task such as arithmetic has given researchers insight into how skills and strategies change dramatically across development (Fuson, 1984; Geary, 1994). Studies of mathematical cognition have also demonstrated children's flexibility in choosing and applying a multitude of different strategies, depending on problem difficulty and characteristics (Hamann & Ashcraft, 1985; Houlihan & Ginsburg, 1981; Lemaire & Siegler, 1995; Siegler, 1987a; 1989b; Siegler & Shipley, 1995; Siegler & Shrager, 1984). In this study, one type of simple arithmetic problem was chosen, subtraction problems.

Because they are introduced in Grade 1, subtraction problems are an appropriate task for students in a wide range of grades. Even young children use a variety of strategies to solve these problems (e.g., Boulton-Lewis, 1993; Siegler, 1987a, 1989b). Typically, as students get older, there is a heavier reliance on simply retrieving the answer from memory, especially on familiar and relatively easy problems (Geary, 1994; Siegler, 1987a, 1989b). Subtraction problems are typically considered more difficult than addition problems and therefore may lead to more variety in strategy use, even for older children (cf. Hamann & Ashcraft, 1985). Thus, subtraction problems lend themselves well to this study.

Subtraction Solution Strategies

Multiple subtraction solution strategies were expected to be reported by the students in this study as students tend to use several strategies to solve even fairly simple subtraction problems (e.g., problems such as 7 - 2 and slightly more difficult problems such as 16 - 13). According to Geary (1994), children tend to rely on counting fingers when they are first learning subtraction. For example, on a problem such as 6 - 1, children might raise one finger at a time until they have six fingers raised then they will lower one finger and finish by counting how many fingers remain raised. Children then tend to start using verbal counting strategies (Geary, 1994). For example, for the problem 14 - 4, children might solve it by counting down four times (14, 13, 12, 11), and so the answer would be the next number (10). Conversely, children could use a different counting procedure, counting up. Using the same example, children would count up from 4 to 14: 5, 6, 7, 8, 9, 10, 11, 12, 13, 14. Note that both of these counting procedures can be done verbally or by using fingers.

Geary (1994) postulated that the next type of strategy that students use predominantly involve decomposition strategies. Decomposition strategies are numerous but all involve breaking a problem into smaller and/or easier parts. For example, on a problem such as 14 - 6, a student may break the problem into 14 - 4 = 10 and then subtract 2 more from the 10 to get the answer, 8. Another example would be on a problem such as 17 - 13 on which a student might simply "delete the 10s" and be left with 7 - 3 and then the answer to this easier subtraction problem might be retrieved from memory or counted. Decomposition strategies are more sophisticated than counting strategies and rely on the student having at least some understanding of the concepts involved in subtraction.

Retrieval involves retrieving the answer to a problem from memory. Students who use retrieval will often report that the answer to a problem such as 8 - 3 just pops into their heads. As such it is a quick and automatic process that is usually accurate (Siegler & Shipley, 1995; Siegler & Shrager, 1984). Conversely, counting tends to be the slowest and most error-prone family of solution strategies. These differential latency and accuracy data characteristics of counting and retrieval are one of the reasons that the veridicality of verbal reports of subtraction solution procedures can be verified.

The strategy chosen to solve a subtraction problem depends partly on problem difficulty. Basic strategies such as counting are used on problems that are more difficult for children whereas retrieval is used most often on problems that children consider to be easy. Some decomposition strategies are used solely on problems with special characteristics (e.g., the delete 10s strategy can only be used on problems where at least one of the numbers is two digits). Retrieval is often used on the simplest subtraction problems as they are the ones that are known the best. Students in all grades were also expected to use more than one strategy, regardless of what grade they were in. For example, Siegler (1989b) found that 99% of children used at least two strategies, 91% used at least three, 72% at least four, and 41% at least five to solve the same set of subtraction problems used in this study. Students were expected to use several strategies, regardless of grade but, based on Geary (1994), some hypotheses could be made about which strategy would dominate across the different grades used in this study. More counting was expected to be reported in Grade 1. Grade 3 students were expected to report less counting than Grade 1 students and less retrieval use than Grade 5 students, and possibly more decomposition use, than Grade 1 students. Retrieval was expected to be reported most frequently in Grade 5.

This increased use of retrieval for older students was expected and yet this does not mean that the younger students do not try to retrieve whenever possible. According to Siegler's (1989b) model of subtraction strategy development, children automatically attempt to retrieve and only when the attempt fails do the children then fall back on "backup strategies" such as counting and decomposition. Siegler is referring here to what other researchers have labeled factual and procedural knowledge (Bisanz & LeFevre, 1990). In the area of mathematical development, there are thought to be three forms of knowledge that children use to help them solve arithmetic problems.

Types of Mathematical Knowledge

The first type of mathematical knowledge is factual knowledge. This refers to knowledge of arithmetic facts, that is, simply knowing that the answer to 6 - 3 is 3 (Bisanz & LeFevre, 1990). Retrieval is the accessing of factual knowledge. The second type of mathematical knowledge is procedural knowledge. Procedural knowledge includes knowledge of appropriate procedures or strategies to use to solve problems. For example, a child in Grade 1 may not be able to use factual knowledge to solve the problem 17 - 14. Thus, the student must use what Siegler (1987a, 1989b; Siegler & Shipley, 1995; Siegler & Shrager, 1984) calls a "backup strategy" or procedural knowledge such as counting. Thus, factual and procedural knowledge are intertwined forms of knowledge that operate and develop in conjunction.

Young children tend to have relatively few procedures at their disposal but as procedural knowledge develops, children will have more procedures available to them and these procedures will become increasingly efficient. For example, a child might switch from counting on their fingers to verbal counting, or a child could switch from a verbal counting strategy to using a decomposition strategy. However, it must be emphasized that even young children have a number of procedures at their disposal and they are able to use them adaptively (Lemaire & Siegler, 1995; Siegler & Shipley, 1995; Siegler & Shrager, 1984). Thus, Siegler's (1989b) study was a study of mainly procedural knowledge, but also of factual knowledge.

In this study, not only factual and procedural knowledge was examined, but also conceptual knowledge, the third type of mathematical knowledge. Conceptual knowledge has traditionally been the form of arithmetical knowledge most difficult to define and assess (Bisanz & LeFevre, 1990). Bisanz and LeFevre defined conceptual knowledge as "the principles that reflect the underlying structure of mathematics and that can be inferred from the selective use of effective procedures where those principles apply" (pp. 216-217). As with factual knowledge, conceptual knowledge is closely linked with procedural knowledge. Conceptual knowledge is hypothesized to aid in the appropriate selection of one procedure from a set of many and the underlying understanding of the principles of arithmetic involved in a particular problem will affect the procedure selected.

Procedural knowledge was assessed in this study by collecting verbal reports as well as accuracy and latency data on the same set of subtraction problems utilized by Siegler (1989b). Factual knowledge was separately assessed by asking students to solve a series of similar subtraction problems as quickly as possible in a specified amount of time. Students were thus forced to work quickly, meaning that students who were able to solve the most problems with the fewest errors would probably be the ones with the best factual knowledge of subtraction. Conceptual knowledge was also assessed by presenting students with demonstrations of several correct and incorrect solution strategies for subtraction problems and asking them whether or not the solution strategies were appropriate or not and to justify their response. Thus, even if students had not reported using the demonstrated strategies in the procedural knowledge task, their conceptual knowledge of subtraction strategies was still assessed.

Overall, this study provided the opportunity to investigate several issues in the development of arithmetic skills. First, it was an attempt to partially replicate Siegler's (1989b) findings. Second, because three separate grades were studied, the development of subtraction strategies within and across grades could be studied more closely. And finally, this study assessed all three forms of arithmetical knowledge and thus how the forms of knowledge develop both independently and together could be investigated.

Overview

The first goal of this study, to determine under what circumstances verbal reports yield veridical and non-reactive data, was assessed by having students in Grades 1, 3, and 5 provide concurrent or retrospective reports of their solution strategies on a series of simple subtraction problems. To determine whether either type of report changed performance on the subtraction problems, a control group of students solved the same set of subtraction problems without providing verbal reports. Few tests of reactivity have been performed on this type of task with children, and therefore few results could be predicted beforehand. Accuracy and latency data as well as strategy reports were used to investigate whether both or either of the verbal report types affected task performance. However, there is fairly strong evidence in the literature that providing concurrent reports does slow the solution process. Thus slower latencies were expected for the students who reported concurrently.

Few tests of veridicality have been performed on this type of task with students but what little evidence exists has been favourable (e.g., Siegler, 1987a, 1989b concluded that retrospective reports yielded valid data). Patterns of accuracy and latency data were examined for the different types of solution strategies reported by the students to determine verbal report veridicality. Differences in reactivity and veridicality across development were also examined.

The second goal of this study, to investigate the development of arithmetic skills through the use of subtraction problems, was fulfilled by first asking students in all grades to solve the aforementioned series of subtraction problems. Accuracy, latency, and strategy verbal reports were all used to determine how performance changes and develops. This first subtraction task was a measure of procedural knowledge, one of the three types of mathematical knowledge that aid students in understanding and solving arithmetic problems. The other types of knowledge, factual and conceptual, were assessed by asking students to complete two additional tasks. Thus, how performance on each task was related to performance on the other tasks could provide a window on when the different forms of knowledge are mastered across development. Conceptual knowledge is usually the last type of arithmetical knowledge to develop so performance on this task was expected to be best for older students. However, the younger students were expected to have more difficulties on all tasks. Additionally, the procedure recognition and fluency tasks provided further tests of reactivity. That is, these tasks were used to determine whether or not the instruction to provide verbal reports could alter subsequent performance on tasks related to the primary task.

METHOD

Participants

A total of 178 students participated, including 30 boys and 30 girls in Grade 1, 31 boys and 27 girls in Grade 3, and 27 boys and 33 girls in Grade 5. Median ages for students in Grades 1, 3, and 5 were (in years:months) 6:10 (range 6:5 to 7:9), 9:0 (8:4 to 9:6), and 10:10 (10:2 to 11:6), respectively. The study was conducted during the last few months of the school year.

Materials and Procedures

Students participated in two sessions. In the first session, they completed a subtraction problem solving task and a procedure recognition task. This session, which lasted approximately 30 minutes for the younger children and 20 minutes for the older children, was conducted with one student at a time and were videotaped. In the second session, students completed a fluency task in approximately 5 minutes. This session was not videotaped and was conducted in small groups.

Subtraction Problem Solving Task

The problem set was composed of 36 problems taken from a study by Siegler (1989), including 9 problems for each of four types. Type 1 problems had small subtrahends and small minuends (e.g., 6 - 2) with a medium difference, Type 2 problems had large subtrahends and small minuends (e.g., 15 - 1) with a large difference, Type 3 problems had large subtrahends and medium minuends (e.g., 14 - 8) with a medium difference, and Type 4 problems had large subtrahends and large minuends (e.g., 16 - 14) with a small difference. The problems were designed by Siegler such that a counting-up strategy would be much easier, more difficult, or equal to a counting-down strategy depending on problem type. Problems were presented in vertical format in a large font on letter-size paper with six problems per page. Ten orders of 36 problems (five orders and their reverses) were constructed with the following constraints for problem order: problems were ordered in random blocks of four with one problem of each type within each block; the position of problem type within each block was counterbalanced as closely as possible; and problems with the same subtrahend, minuend, or answer were never presented consecutively. See Appendix A for a list of all 36 problems.

Students were assigned randomly to one of three conditions, balancing sex as closely as possible. Students in all report conditions solved the same 36 problems. No feedback on responses was given. Accuracy, latency, and verbal report data were collected. Latencies were timed with a stopwatch by the experimenters and accuracy was ensured by checking the videotapes later. When times were rounded to the nearest tenth of a second, few discrepancies were found (11 out of 648) and all discrepancies were less than 0.8 s. For the verbal report data, during the sessions the experimenters recorded the students' verbal reports on the data sheets. All sessions were viewed later on videotape to ensure that the records were complete.

No Report Condition

The student was given the following instructions:

> We are going to do some subtraction problems today. I'll show a problem to you, and when you have an answer, tell me what it is. You can do anything you want to get the right answer. You can count or use your fingers or do whatever you want to do. It doesn't matter how you get the right answer, as long as you do the best you can.

The student then practiced on as many as six practice problems until he had familiarized himself with the task.

Retrospective Report Condition

The student was given the same instructions as the No-Report condition students, but after solving the first practice problem she was told:

> I'm really interested in how kids your age figure out the answers to these problems. How did you figure out the answer to that problem? (If necessary, experimenter prompted with the following: Did you count in your head? Count on your fingers? Use a different way?).

The student then proceeded with the same practice problems as the No-Report condition but was asked to report how she had solved each problem immediately after she had given an answer. The student practiced on as many as six problems until she appeared to understand the instructions for retrospective reporting.

Concurrent Report Condition

The student was given the same instructions as in the No Report condition but, before solving the first practice problem, he was told:

> I'm really interested in how kids your age figure out the answers to these problems. I'll show a problem to you, and I want you to tell me how you are trying to figure out the answer while you are working on the problem. Let's practice telling me how you are trying to figure something out while you work on it. Here's a shoe with the laces undone. Do you think you could tie the laces up and, at the same time, tell me exactly what you are doing?

The student then proceeded with the same practice problems as the No-Report condition but was reminded to tell the experimenter how he solved each problem as he worked on it. The student practiced on as many as six problems until he appeared to understand the instructions for concurrent reporting.

Procedure Recognition Task

Immediately following the subtraction problem solving task, students were asked to do the procedure recognition task. The task consisted of six different solution procedures that were demonstrated to the students. Four procedures were legitimate ways of solving subtraction problems (counting down, counting up, deleting tens, and addition fact) and two procedures were illegitimate ways of solving subtraction problems correctly (adding the subtrahend and minuend and combining digits). Illegitimate procedures were included to ensure that students were not simply answering the same way for all of the procedures. Eight orders of the procedures (four orders and their reversals) were used. Two legitimate procedures always were followed by an illegitimate procedure and this pattern was then repeated. Students' responses were recorded on a data sheet and videotaped. See Appendix B for the complete task and instructions.

The student was told the following:

> I know this boy/girl who's learning to add and subtract. I gave him/her some addition and subtraction problems to solve. He/she told me about different ways to solve the problems. I would like you to tell me if the ways the boy/girl tried to solve each problem would work. Some of the ways may be "good" ways of solving the problems and some of the ways may be "silly" way of solving the problems.

The student was first given two practice addition procedures, one of which demonstrated a legitimate way (counting on) and a illegitimate way ("4 + 3 is 7 because I'm 7 and 5 + 6 is 11 because my sister is 11"). For each addition procedure, the student was asked for her opinion about whether the procedure was good or silly and then asked to justify her answer. After this practice, the student was randomly assigned to one of the eight orders (balancing for sex as closely as possible) and was asked for her opinions on the six subtraction procedures.

Fluency Task

The fluency task consisted of two pages of subtraction problems presented in vertical format. On the first page, all 45 subtraction problems with subtrahends and minuends between 1 and 9 were presented. On the second page, all 90 subtraction problems with subtrahends between 11 and 19 and minuends between 1 and 9 were included. Problems were presented in the same random order for all students with the constraint that the same minuend or the subtrahend could not appear consecutively more than twice. See Appendix C for the complete set of problems.

Students completed the fluency task in small groups. Each student was asked to complete as many problems on the page as possible within the time limit while trying not to make any mistakes and without skipping any problems. Three students in Grade 1 and two students in Grade 3 did not follow task instructions so no fluency data for these students was used for purposes of analysis. Students in Grades 1 and 3 were given 90 seconds to work on the first page and Grade 5 students were given 60 seconds. Grade 1 students were not asked to work on the second page. Students in Grade 3 and 5 were given 90 seconds to work on the second page. Students were then thanked for their participation and given the opportunity to ask questions about the study.

RESULTS AND DISCUSSION

The results are divided into two major sections: the validity of verbal reports in children's subtraction and the development of subtraction skills. Within the first section, findings from the subtraction problem solving task, the procedure recognition task, and the fluency task are presented. Each task is divided into subsections of findings involving grade, gender, and sometimes problem type, that help clarify the findings involving verbal report reactivity and veridicality. Though no specific gender differences were expected, gender was included in the analyses on the basis of known gender differences in later mathematics performance (Leder, 1992). Any gender differences found in this study might then be related to differences which occur later on in development. In the second section, results will be presented on how children's accuracy, solution times, and solution strategies for subtraction problems change across development, and findings regarding the different forms of mathematical knowledge, as measured with the three tasks, also are delineated.

Verbal Report Validity in Children's Subtraction

Verbal report validity can take two forms: reactivity and veridicality. Reactivity was considered first. The instruction to provide verbal reports, whether retrospectively or concurrently, could potentially affect performance on all three of the tasks in this study. To provide tests of the potential reactive effects of verbal reporting, all three measures (accuracy, latency, and strategy reports) from the subtraction problem solving task were analyzed separately, as were the data from the procedure recognition and fluency tasks. The following analyses differ from most tests of reactivity found in the recent literature in that more tasks, and more types of data, are analyzed (cf. Cooney & Ladd, 1992; Russo, Johnson, & Stephens, 1989). However, tests of reactivity are not independent of variables such as grade, gender, and problem type and thus results involving these three variables are presented first for each measure of each task.

The second validity issue for verbal reports is veridicality. The main question to be addressed in this section is whether or not students' reports of their solution strategies are accurate representations of their solution processes. To provide tests of the veridicality of retrospective and concurrent reporting, verbal report data from the subtraction problem solving task will be examined. Accuracy and latency data from the subtraction problem solving task will also be examined to determine whether there is converging evidence for the veridicality of verbal reports. In this study different students participated in the two report conditions, in contrast to other attempts to verify the validity of verbal reports (cf.

Cooney & Ladd, 1992; Fidler, 1983; and Russo et al., 1989). Thus, in this study, there is no possibility of practice effects influencing performance in subsequent report conditions. However, differences in students' retrospective and concurrent reports cannot be directly compared to each other.

Subtraction

Measures used in the subtraction problem-solving task included accuracy, latency, and verbal reports of strategy use. For each measure, analyses of the possible effects of grade, gender, and problem type are presented first, followed by analyses of the potential reactive effects of verbal reporting. An alpha level of .05 was used for all statistical tests and all reported differences were significant ($\underline{p}s \leq .05$), unless noted otherwise. Finally, for strategy use, tests of verbal report veridicality are also presented.

Accuracy

Analyses involving accuracy were based on the proportion of correct responses either for all 36 problems (overall accuracy) or for each of the 4 problem types (Types 1-4 accuracy). Incorrect responses included both wrong answers and cut-offs. A response was coded as a cut-off when students were unable to provide an answer within approximately 40 seconds or else told the experimenter that they were unable to solve the problem. Cut-offs occurred on 9.9% of the problems for Grade 1 students versus 0.1% and 0% for Grade 3 and 5 students, respectively. Proportions of correct responses were subjected to a 3 (Grade: 1, 3, and 5) x 3 (Report Condition: No Report (NR), Retrospective Report (RR), and Concurrent Report (CR)) x 2 (Gender: male and female) x 4 (Problem Type: 1 through 4) analysis of variance with repeated measures on the last variable.

Grade, gender, and type. Not surprisingly, there was an effect of grade, $F(2, 160) = 96.86$, $p < .001$, with Grade 1 students having significantly smaller proportions of accurate responses than both Grade 3 and 5 students, who did not differ from each other (.60 vs. .90 and .92, respectively). Grade 3 and 5 students were close to ceiling which could account for the lack of grade effect for these two groups. No effects involving gender were found. A main effect of type was found, $F(3, 480) = 73.27$, $p < .001$. Examples of problems of each type are: 5 - 1 (Type 1), 13 - 2 (Type 2), 14 - 9 (Type 3), and 15 - 13 (Type 4). Tests of simple effects showed that Type 1 problems had a higher proportion of accurate responses than the Type 2 through 4 problems (.94 vs. .86, .72, and .73, respectively). Type 2 problems were easier than both Type 3 and 4 problems and there was no difference between the last two problem types.

Although main effects for grade and type were found, both effects were qualified by an interaction between the two variables, $F(6, 480) = 30.86$, $p < .001$. Within each problem type, grade followed the same pattern as the main effect of grade. For each of the problem types, students in Grade 1 performed significantly worse than either Grade 3 or 5 students who did not differ from each other (see Table 1). Within both Grades 1 and 3, the type interaction variable followed the same pattern as the main effect of type. In Grade 5 however, performance on Type 1 and 2 problems differed from that on Type 3 and 4 problems. In Grade 5 only, Type 2 problems were not more difficult than Type 1 problems.

Table 1

Proportion of Correct Responses on all Problem Types as a Function of Grade

	Problem Type				
Grade	Type 1	Type 2	Type 3	Type 4	Mean
1	.896	.715	.389	.413	.603
3	.969	.923	.866	.856	.904
5	.959	.928	.900	.909	.924
Mean	.942	.856	.719	.727	.811

Overall, the younger students were less accurate on all problem types than the older students, but within each problem type, Types 1 and 2 tended to be more accurate than Types 3 and 4. For older students, accuracy was very close to ceiling, especially on Type 1 and 2 problems, but proportions of accurate responses on Type 3 and 4 problems were still quite high, thereby requiring cautious interpretations. Though students in all grades made more errors on Type 3 and 4 problems, the drop in proportion of accurate responses was larger for the younger students. Thus, on problems where students are required to subtract a medium to large amount (between 8 and 14), younger students start having dramatic difficulties providing accurate solutions, and even the older students had a somewhat harder time with these problems.

Reactivity. If verbal reporting itself affects task performance, then students in either or both of the retrospective report (RR) and concurrent report (CR) conditions might be expected to have significantly more or less accurate responses than the students in the no report (NR) condition. Overall, no main effects or interactions involving the Report Condition were found (see Appendix D). However, note that task performance was close to ceiling for the older students on all of the problems and close to ceiling for all students on Type 1 problems.

Accuracy overview. No surprising findings were revealed regarding grade, gender, or problem type effects. As expected, younger students made more errors or were unable to solve more problems than the older students. The Grade 3 and 5 students had very high proportions of accurate responses. Performance was not related to gender. Problem type was a significant factor on accurate performance: In general, regardless of grade, problems with smaller subtrahends were easier than problems with medium to large subtrahends. Reactive effects of verbal reporting were not found on accuracy measures, thus lending support to the conclusion that the instruction to verbally report, whether retrospectively or concurrently, does not affect task performance.

Latency

Analyses of solution time latencies were based on correct responses only. Median latencies were calculated for a maximum of all 36 problems (overall latency) and for the maximum of 9 problems for each problem type (Types 1- 4). The large variability found in the Grade 1 latencies could obscure the smaller but possibly still significant differences in the older grades, and therefore latency analyses were conducted separately for each grade. Because Grade 1 students had a high rate of both inaccurate and cut-off responses, especially on the more difficult problems (e.g., Types 3 and 4), latency data from these students are assumed to be underestimates of true values and should be interpreted cautiously. Also, students in the Concurrent Report condition were expected to have longer solution latencies as the process of verbalizing during problem solving usually slows the solution process (Ericsson & Simon, 1980, 1984, 1993). Median latencies of correct responses were subjected to a 3 (Report Condition: NR, RR, and CR) x 2 (Gender: male and female) x 4 (Problem Type: 1 through 4) analysis of variance.

Grade, gender, and type. Though each grade was analyzed separately, an examination of the median latencies indicates that latencies differed across grade (10.5, 3.9, and 2.1s for Grades 1, 3, and 5, respectively). No effects involving gender were found in any grade, although a significant interaction between problem type and gender was found in Grade 3, which is presented below. In all grades, as with the accuracy data, an effect of type was found, $F(3, 105) = 26.32$, $p < .001$, $F(3, 153) = 19.41$, $p < .001$, and $F(3, 162) = 13.07$, $p < .001$, for Grades 1, 3, and 5, respectively (see Table 2).

Table 2

Median Latencies (in seconds) of Correct Responses on all Problem Types as a Function of Grade

	Problem Type				
Grade	Type 1	Type 2	Type 3	Type 4	Mean
1	6.6	7.4	16.9	11.6	10.5
3	2.9	3.7	4.8	4.2	3.9
5	1.7	2.0	2.8	2.1	2.1
Mean	3.7	4.4	6.1	6.0	5.5

In Grade 1, all types differed significantly from each other, with Type 1 problems having the fastest solution latencies, followed by Type 2 problems then Type 4 problems, and finally Type 3 problems had the longest latencies. In Grades 3 and 5, the pattern of differences was exactly the same. Type 1 problems had the shortest solution latencies and were different from all other problem types; Type 2 and 4 problems did not differ from each other but were slower than Type 1 problems and faster than Type 3 problems. In Grade 3, however, the main effect of type was qualified by an interaction between type and gender, $F(3, 153) = 2.90$, $p = .037$. For males, the type effect followed the same pattern as the main effect: Type 1 problems had faster solution latencies than all other problem types, no differences were found between problems of Types 2 and 4, but both were faster than Type 3 problems (see Figure 1). For females, Type 1 problems had smaller latencies than all other types, Type 2 problems had smaller latencies than both

Type 3 and 4 problems, but these last two problem types did not differ from each other. There were no gender differences on problems of Types 1, 2, or 3, but on Type 4 problems, males had faster solution latencies than females.

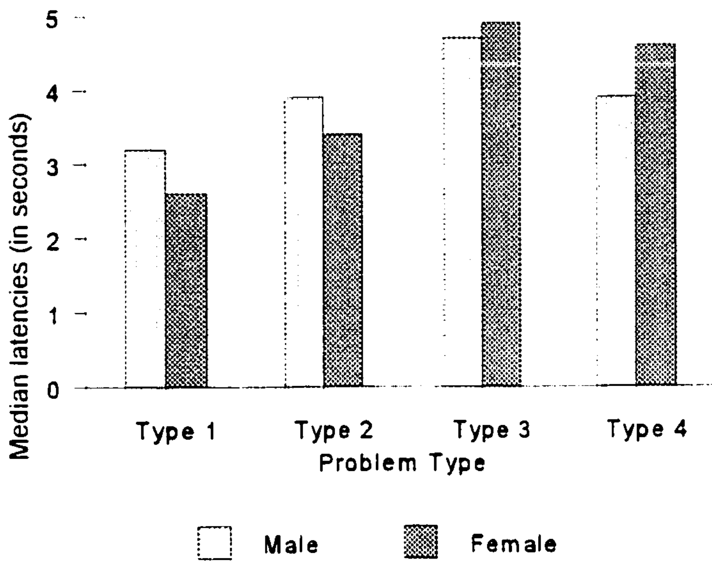

Figure 1. Interaction between Gender and Problem Type in Grade 3.

Reactivity. If verbal reporting itself affects task performance, then students in either or both of the RR and CR conditions might have significantly faster or slower solution latencies than the students in the NR condition. However, solution latencies were expected to be longer for the CR students because of the possibility for interference between verbalizing and problem solving (Ericsson & Simon, 1980, 1984, 1993). Therefore, the finding that there were no main effects or interactions involving Report Condition in the analyses of the Grade 1 and Grade 5 median latencies (see Table 3) was unexpected. Note, however, that the means were in the expected direction (see Appendix E for a more detailed presentation of the means). In Grade 3, though, there was a main effect involving Report Condition, $F(2, 51) = 7.86$, $p = .001$. Students in the NR and RR groups had latencies that did not differ from each other but both groups were faster than the CR group. Thus, for all grades, students in the NR and RR groups did not have significantly different latencies. However, though the means were in the expected direction for Grades 1 and 5, only in Grade 3 were the latencies significantly slower in the CR group.

Table 3

Median Latencies (in seconds) of Correct Responses for each Report Condition as a Function of Grade

	Grade		
Report Condition	1	3	5
NR	8.1	2.8	1.8
RR	11.3	3.6	2.0
CR	12.4	5.2	2.6
Mean	10.6	3.9	2.2

Report condition difficulties. One possible explanation for the unexpected lack of significantly slower latencies for the CR students in Grades 1 and 5 is that a large number of students in the CR group, regardless of grade, reported retrospectively at least once in the concurrent report condition. Though students in the RR condition usually found the instruction to provide a report of their problem-solving process straightforward, students in the CR condition often had difficulty solving the subtraction problems and reporting their solution process concurrently. Instead, they often retrospectively reported their solution processes, despite instructions to the contrary, and thus these students might be expected to be more similar to the students in the RR condition. Consequently, students were re-categorized as belonging to a new report condition group (RR in CR) if they retrospectively reported at least once in the concurrent report condition. Although there were no grade differences in the number of students having difficulty concurrently reporting (14 out of 20, 13 out of 19, and 14 out of 20 for Grades 1, 3, and 5, respectively), the mean percentage of problems (out of 36) for which they failed to provide a concurrent report was somewhat higher for the older students (19.6, 31.7, and 25.7%, for Grades 1, 3, and 5, respectively). Reasons for this difficulty in following concurrent report instructions are discussed in a later section.

The same analyses were redone separating the students into four report groups: NR, RR, CR, and RR in CR. Note that because students who even reported retrospectively once in the CR condition were re-categorized as now belonging in the RR in CR group, any new results involving report condition would be a conservative estimate of the effects of retrospectively reporting in the CR condition. Once again, no effect of Report Condition was found in Grade 1, suggesting that Grade 1 students were slow, regardless of which report group they belonged to (see Table 4). In Grade 3, Report Condition again had an effect, although the group differences were more complex in this new analysis, $F(3, 49) = 11.17$, $p < .001$. Students in the pure CR group had slower latencies than the other three report groups. The NR group did not differ from the RR group but was faster than the RR in CR group, and finally the RR group did not differ from the RR in CR group. In Grade 5, an effect of Report Condition was found in the new analyses, $F(3, 52) = 4.61$, $p = .006$. Following the pattern found in the Grade 3 data, students in the CR group had slower latencies than the other three report groups but these three groups did not differ from each other (see Appendix F for more detailed means). Analyses of the accuracy data were also redone for the four report groups but once again no differences involving Report Condition were found.

Table 4

Median Latencies (in seconds) of Correct Responses for each Report Condition (including RR in CR) as a Function of Grade

	Grade		
Report Condition	1	3	5
NR	8.1	2.8	1.8
RR	11.3	3.6	2.0
RR in CR	12.4	4.3	2.1
CR	12.3	7.3	3.9
Mean	10.6	3.9	2.2

Note. NR: ns = 20 for each grade. RR: ns = 20, 19, and 20 for Grades 1, 3, and 5, respectively. RR in CR: ns = 14, 13, and 14 for Grades 1, 3, and 5, respectively. CR: ns = 6 for each grade.

Latency overview. As with the accuracy data, few findings regarding grade, gender, or problem type were of note. Though analyses were done separately for each grade, latencies were in the expected direction with Grade 1 students having slower latencies than the older students. Following similar patterns to the accuracy data, latencies were slowest on problems with medium to large subtrahends and fastest on problems with small subtrahends. Only in Grade 3 was gender a significant factor as it interacted with problem type. This interaction involved one problem type that was solved faster for males than females. There is no obvious explanation for this finding.

The task of concurrent verbal reporting itself was expected to slow solution times as postulated by Ericsson and Simon (1980, 1984, 1993). In initial analyses of reactivity, only Grade 3 students were significantly slower in the CR condition than in the NR and RR conditions. Means were in the expected direction for Grades 1 and 5, however. The lack of reactivity in these two grades may be due to the difficulty that many of the students had with providing concurrent reports. The majority of students in all grades retrospectively reported at least once in the CR condition.

Analyses were redone with a new report condition, RR in CR, to determine whether "pure" concurrent reporters had slower latencies than the other three report groups. Slower latencies for the concurrent reporters were found in Grades 3 and 5 but once again in Grade 1 there was no effect on solution latencies based on the report condition students were in. This anomalous finding may be due to the tendency of many of the younger children to spontaneously concurrently report or "think aloud" while solving problems, regardless of report condition instructions. Indeed, in the Grade 1 NR group, students spontaneously reported a solution strategy on 37.4% of the problems. Overall, no unexpected evidence of reaction was found (although unexpected non-reactivity was found). Thus, the potential reactive effects of verbal reporting do not appear to be of significant concern based on both the latency and accuracy data. Ericsson and Simon (1980, 1984, 1993) cited a number of studies in which performance, although slower, did not otherwise change with the instruction to concurrently report. However, because solution times were indeed affected in Grades 3 and 5, this finding, though expected, may indicate that researchers concerned about any changes in performance, expected or unexpected, may prefer to use retrospective reports rather than concurrent reports.

Strategy Reports

Strategy reports were collected in two report conditions: Retrospective Report and Concurrent Report. Students in the No Report condition were not asked to provide a verbal report. However, Grade 1 students in the NR group sometimes provided spontaneous verbal reports (on 37.4% of the problems). A total of 65 strategies were identified as being used at least once (see Appendix G for a description of all 65 strategies). These strategies fell into nine categories: counting, retrieval, suspected retrieval, derived fact, special trick, addition fact, unknown, guess, and other. A description of each strategy category is provided in Appendix H.

Reliabilities for each of the nine categories were calculated based on the verbal reports of 18 of the 178 students. Six students were selected from each grade, half from the RR condition and half from the CR condition. Gender was balanced within each grade. Strategy coding was based on viewing videotapes of the selected students solving the complete problem set. Discrepancies were resolved by the primary coder (the experimenter). Reliabilities were calculated by dividing the number of agreements by the number of agreements plus the number of cases in which the secondary coder disagreed with the primary coder. Overall reliability for the categories was 94.2%, with reliabilities ranging from 50.0 to 100.0% (see bottom of Table 5). Reliabilities were lowest on the categories used most infrequently (unknown, other, and guessing) for which even a small number of disagreements could affect reliability dramatically. Both coders used each of these three categories on less than 4% of the trials.

Because students in each grade tended to report using different strategies with differing frequencies, analyses involving each strategy category were conducted separately for each grade. As well, analyses were performed separately for each strategy category only if it was reported frequently enough for meaningful interpretation. For example, no analyses were performed on the guessing, derived fact, or other categories in any grade due to the low frequency of use. Also, strategy categories such as addition facts were analyzed in Grades 3 and 5 but not in Grade 1 (see Table 5). Proportions of reported frequencies for each strategy category were calculated for each student by dividing the number of reported instances of a specific strategy category by the number of possible opportunities to use that strategy (e.g., nine times on Type 1 problems). Proportions of reported frequencies were then subjected to a 2 (Report Condition: RR and CR) x 2 (Gender: male and female) x 4 (Problem Type: 1 through 4) analysis of variance.

Table 5

Proportions of the Frequency of the Strategy Categories Reported by the Retrospective (RR) and Concurrent report (CR) Students as a Function of Grade

		Strategy Category								
Grade	Report Condition	Counting	Retrieval	Suspected Retrieval	Derived Fact	Special Trick	Addition Fact	Unknown	Guess	Other
1										
	RR	.43	.13	.07	.01	.11	.01	.15	.07	.03
	CR	.49	.13	.08	.00	.12	.01	.14	.02	.01
	Mean	.46	.13	.07	.01	.12	.01	.14	.04	.02
3										
	RR	.24	.16	.04	.02	.42	.06	.04	.00	.02
	CR	.15	.17	.11	.03	.22	.19	.11	.00	.01
	Mean	.20	.17	.08	.02	.32	.13	.07	.00	.02
5										
	RR	.23	.24	.09	.04	.25	.13	.01	.00	.01
	CR	.12	.27	.35	.03	.13	.09	.02	.00	.00
	Mean	.17	.26	.22	.03	.19	.11	.01	.00	.01
Reliability		.96	1.00	.94	.91	1.00	1.00	.50	.79	.60

Grade, gender, and type. An examination of the proportions of frequencies for each strategy category revealed a number of differences across grade. For example, as expected, counting was much more frequent in Grade 1 than in Grades 3 and 5 (.50 vs. .16 and .15) and retrieval more common in the older students (.21 vs. .29 and .34). Special tricks were also reported more frequently by the older students (.07 vs. .14 and .10) as were addition facts (.01 vs. .19 and .10). In general, the distribution of frequencies was more evenly dispersed across the categories for the older students, suggesting that they are more flexible in their strategy use. This hypothesis is discussed in more detail in a later section. Significant effects involving problem type on counting, retrieval, and special trick trials are presented here but are examined in more detail in the section on the veridicality of verbal reports.

No differences involving gender were found in Grade 1. An effect of type was found in all analyses of strategy categories that were used frequently: counting, retrieval, suspected retrieval, and special tricks (see Figure 2), as well as the unknown category. Counting was reported more frequently on both problem types 1 (e.g., 5 - 1) and 3 (e.g., 13 - 8) than on problem types 2 (e.g., 13 - 1) and 4 (e.g., 13 -12), $F(3, 108) = 2.85$, $p = .041$. Retrieval was reported most frequently on Type 1 problems than any of the other problem types, followed by Type 2 problems, which had higher reported frequencies of retrieval than on Types 3 and 4 , $F(3, 108) = 11.99$, $p < .001$. Results for suspected retrieval were expected to parallel those for the retrieval category, $F(3, 108) = 7.41$, $p < .001$. Though the means were in the same direction as for retrieval, there was no significant difference between problem Types 1 and 2, which had higher reported suspected retrieval use than Type 4 problems, which in turn had higher frequencies of reported use than Type 3 problems.

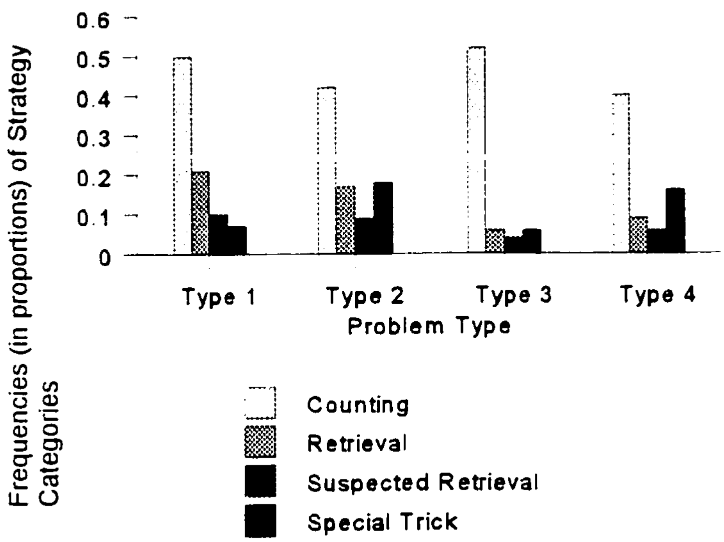

Figure 2. Main Effects of Problem Type on Reported Strategy Frequencies in Grade 1.

Use of special tricks varied across types, $F(3, 108) = 6.49$, $p < .001$, and the pattern of reported use was opposite to that of counting, with problem types 1 and 3 having smaller frequencies of reported category use than problem types 2 and 4. The final strategy category, not shown in Figure 2, was the unknown category. In Grade 1, many of the responses in the unknown category stemmed from cut-off protocols, when students were unable to solve the problem. The unknown category was used more frequently on problem types 3 and 4 than on problem types 1 and 2 (.21 and .18 vs. .09 and .10). It is no surprise that the problem types that produced the most errors and slower solution latencies also produced the largest proportion of unknown strategies, $F(3, 108) = 4.82$, $p = .003$. Generally, in Grade 1, counting was the predominant strategy. Retrieval and suspected retrieval were used on the problem types associated with the fewest errors and the fastest solutions times (Types 1 and 2), and special tricks tended to be used on problem types 2 and 4.

In Grade 3, there were no main effects involving gender but in the suspected retrieval category gender interacted with the type and report condition variables (these interactions are presented below). An effect of type was found in all but one of the strategy categories that were used frequently enough to be analyzed: counting, retrieval, suspected retrieval, and special trick (see Figure 3), as well as addition fact. No effect of type was found in the unknown category. For counting, $F(3, 102) = 3.29$, $p = .024$, Type 2 problems did not differ from any of the other problem types. However, Type 3 problems had higher reported frequency of counting than on problem types 1 and 4. For retrieval, $F(3, 102) = 13.00$, $p < .001$, Type 1 problems had the highest frequency of reported use and there were no differences between the other three problem types.

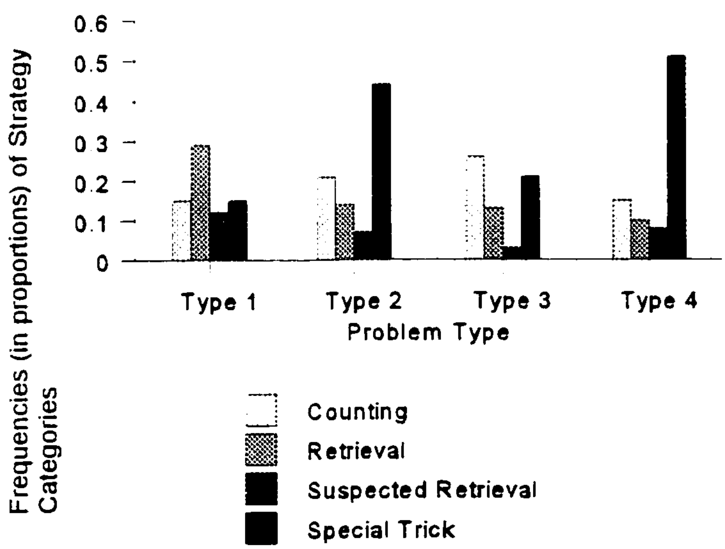

Figure 3. Main Effects of Problem Type on Reported Strategy Frequencies in Grade 3.

Reported use of suspected retrieval varied across problem type, $F(3, 102) = 6.23$, $p = .001$, but did not follow the same pattern as the retrieval category. Though Type 1 problems had higher suspected retrieval use than problem Types 2 and 3, Type 1 problems did not differ from Type 4 problems. Additionally, suspected retrieval was used on Type 2 and 4 problems more frequently than on Type 3 problems. However, this main effect of type is qualified not only by an interaction between type and gender, $F(3, 102) = 3.72$, $p = .014$, but also by an interaction between type, gender, and report condition, $F(3, 102) = 5.96$, $p = .001$ (see Figure 4). Additionally, there was an interaction between report and gender, $F(3, 34) = 7.35$, $p = .01$. From Figure 4, it can be seen that the CR female group had the highest proportion of suspected retrieval use overall and that with the exception of Type 1 problems, the CR male group had very low use of suspected retrieval. In the RR condition, the males tended to have higher proportions of suspected retrieval use, especially on Type 1 and 2 problems. Differences involving gender in the RR condition were negligible on problem Types 3 and 4. For problem type, Type 1 problems tended to have the highest frequencies of use (except for RR females). Generally, frequencies were lower on problem types 2 and 3 and then increased again on problem type 4. Differences involving report condition will be examined in closer detail in the reactivity and veridicality sections.

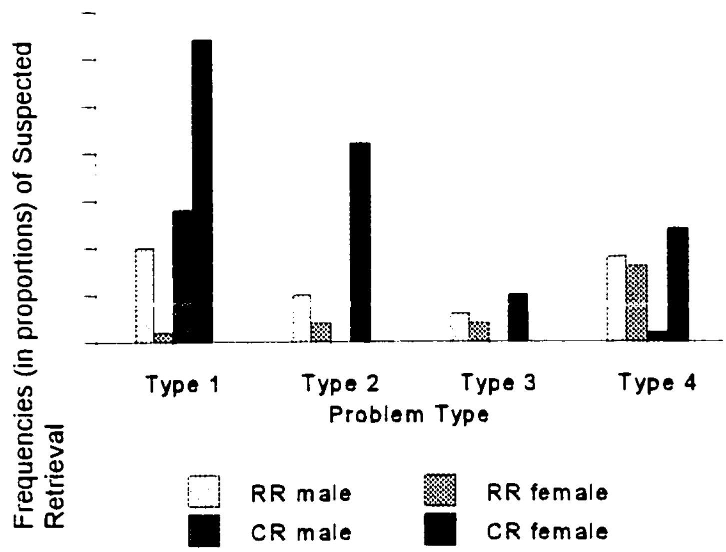

Figure 4. Interaction Between Report Condition, Gender, and Problem Type for Suspected Retrieval Use in Grade 3.

Special tricks, $F(3, 102) = 30.60$, $p < .001$, had higher frequencies on problem types 2 and 4 than on problem types 1 and 3. Addition facts, however, had the opposite pattern, with higher frequencies of reported strategy use on problem types 1 and 3 than on problem types 2 and 4 (.20 and .18 vs. .07 and .06), $F(3, 102) = 11.26$, $p < .001$. Thus, in Grade 3, counting no longer dominated the strategies that students reported using and instead, strategy use was more evenly divided although the highest reported frequencies overall were for special tricks.

In Grade 5, analyses were done on five of the nine strategy categories: counting, retrieval, suspected retrieval, special tricks, and addition fact. No main effects or interactions involving gender were found except in the retrieval category, with an interaction between report condition and gender, $F(1, 36) = 6.87$, $p = .013$. In the RR condition, females reported using retrieval more frequently than males, while in the CR condition there was no gender difference (see Figure 5). For males, there was more reported retrieval use in the CR than in the RR condition, while for females there was no difference in reported use between the report conditions. There is no clear reason for this interaction and it is unique to this grade and strategy category

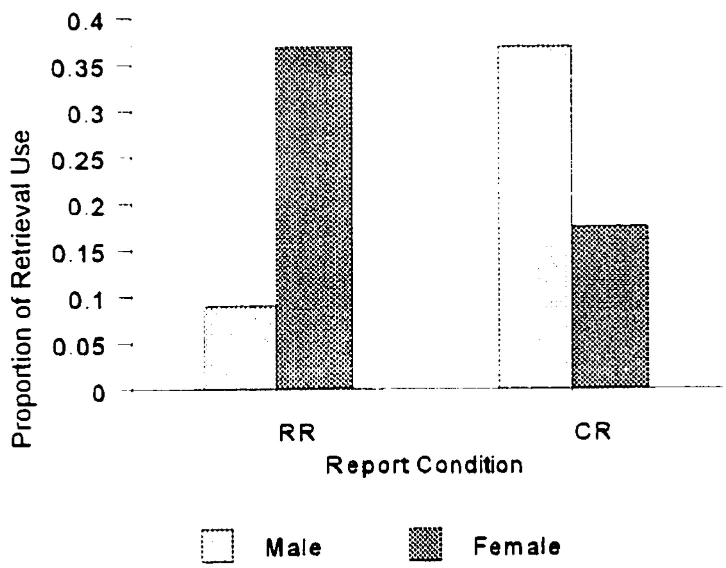

Figure 5. Interaction between Report Condition and Gender for Retrieval Use in Grade 5

For all five of the strategy categories analyzed, an effect of type was found. For counting, frequencies of reported use were highest on Type 3 problems with no differences between the other three problem types, $F(3, 108) = 4.04$, $p = .009$ (see Figure 6). Reported retrieval use was highest on Type 1 problems. Type 2 and 4 problems had the next highest retrieval use and did not differ from each other but retrieval was reported more often than on Type 3 problems, $F(3, 108) = 10.18$, $p < .001$. On suspected retrieval trials, the pattern of effects was exactly the same as for retrieval, as expected, $F(3, 108) =$

9.11, p < .001. The same pattern for special tricks was found in Grade 5 as had been found in both Grades 1 and 3. Frequency of reported special tricks use was higher on Type 2 and 4 problems than on Types 1 and 3, $F(3, 108) = 8.48$, $p < .001$. Finally, for addition facts, reported use was highest on Type 3 problems and did not differ among the other three problem types (.15 vs. .10, .09, and .09), $F(3, 108) = 4.77$, $p = .004$. The higher frequency of counting on Type 3 problems matches the finding of the slowest solution latencies on this type of problem. The fastest solution latencies were found on Type 1 problems which also coincides with the highest retrieval and suspected retrieval use on this problem type.

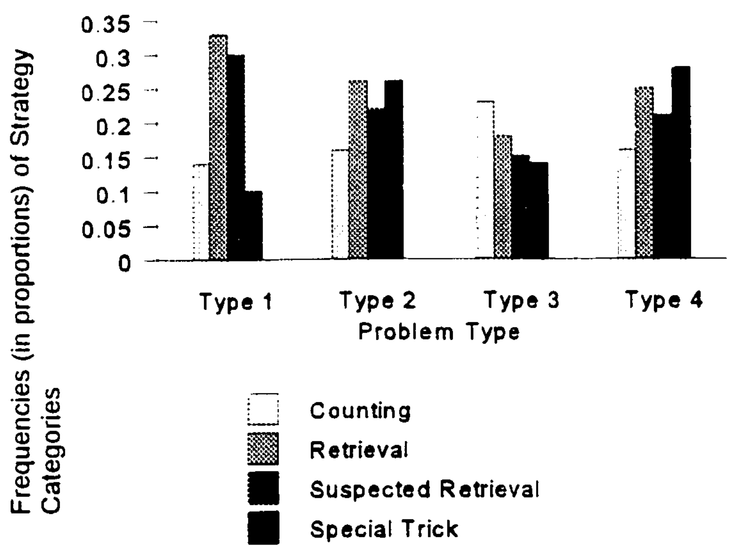

Figure 6. Main Effects of Problem Type on Reported Strategy Frequencies in Grade 5

Overall, in all grades a significant main effect of problem type was found on many strategies. In Grade 1, counting was reported more frequently on Type 1 and 3 problems, retrieval and suspected retrieval were reported more frequently on Type 1 problems followed by Type 2 problems, and special tricks were reported most frequently on Type 2 and 4 problems. In Grades 3 and 5, counting was reported most frequently on Type 3 problems, retrieval and suspected retrieval were reported more frequently on Type 1 problems, and special tricks were reported most frequently on Type 2 and 4 problems.

The tendency to use counting most frequently on Type 3 problems in all grades parallels the finding that there were more errors and longer solution times on this problem type. Thus, Type 3 problems could be considered the most difficult problem type, and students needed to use counting as a "backup strategy" to solve these problems. Type 1 problems had the fewest errors and the quickest solution latencies and thus the use of retrieval and suspected retrieval corresponds well to the relative easiness of this problem type. Grade 1 students also reported counting frequently on Type 1 problems. This finding may be due to the large number of cut-off protocols implemented on Type 3 and 4 problems. Also, counting was relatively easy to use on Type 1 problems. On problems with medium error rates and medium solution latencies, Type 2 and 4 problems, special tricks were reported most frequently by students in all grades. That is, on problems of medium difficulty students had difficulty retrieving the answers and yet were able to use strategies that were quicker and more efficient than counting to attain the solution. Counting may also have been too difficult on these problems, especially in Grade 1, thereby forcing students to come up with more novel solution strategies, such as special tricks. Overall, specific strategies tended to be used on specific problem types, regardless of grade.

Reactivity. Unlike the accuracy and latency data, reactivity with strategy reports cannot be done by comparing the results in the RR and CR conditions to the NR condition results. One alternative is to compare the frequencies of reported strategy use in the RR and CR conditions. If the patterns of frequencies are similar in both groups, this would provide evidence that the demands of these two types of verbal reporting are similar. If, however, the patterns of frequencies are not similar in both groups, conclusions are harder to make. The source of the pattern differences could then be due to either reactivity and/or non-veridicality. That is, there might be some aspect of being asked to provide one or both types of reports that changes the types of strategies that students use to solve these problems. Alternatively, there may be some aspect about providing verbal reports that leads to non-veridical strategy reports in one or both report conditions. Even a cursory glance at the frequencies of reported strategy categories in Table 5 demonstrates that there are at least a few noticeable group differences for a number of the strategy categories in Grades 3 and 5 (and see Appendix I for a complete table of means).

In Grade 1, there were no significant effects or interactions involving Report Condition in any of the analyses of the strategy categories that had high frequencies of reported use. Students in both conditions reported using all strategy categories equally frequently.

In Grade 3, three strategy categories had significant effects or interactions involving Report Condition. Main effects for Report Condition were found for the special trick category, $F(1, 34) = 10.51$, $p = .003$, and the addition fact category, $F(1, 34) = 5.57$, $p = .024$. In the special trick category, students in the RR condition had higher frequencies of use than in the CR condition (.42 vs. .22), and in the addition fact category, students in the RR condition had lower frequencies of use than in the CR condition (.06 vs. .19).

For suspected retrieval, there was a main effect of Report Condition, $F(1, 34) = 4.47$, $p = .042$, which was qualified by interactions between Report Condition and Gender, $F(1, 34) = 7.35$, $p = .01$ (see Figure 7), and, as previously reported, between Report Condition, Gender, and Problem Type (see Figure 4). Overall, suspected retrieval was more frequent in the CR condition than in the RR condition. This finding is consistent with the notion that students had more difficulty concurrently reporting their solution strategy than retrospectively reporting it. That is, though retrieval was suspected, students in the CR condition failed to report their solution strategy themselves and retrieval was inferred by the coder. The proportions of suspected retrieval use were not different between report conditions for males, but for females, there was more suspected retrieval use in the CR condition. However, within each report condition, opposing gender differences were found: within the RR condition, males had higher frequencies of suspected retrieval use than females, and within the CR condition, exactly the opposite pattern was found. For the interaction involving type, no clear patterns emerged. Generally, Type 1 problems had a higher proportion of suspected retrieval (Type 1 problems were the easiest problems in the problem set consisting of problems such as 6 - 2) and most of the differences between males and females occurred in the CR condition in which females usually had higher proportions of suspected retrieval use.

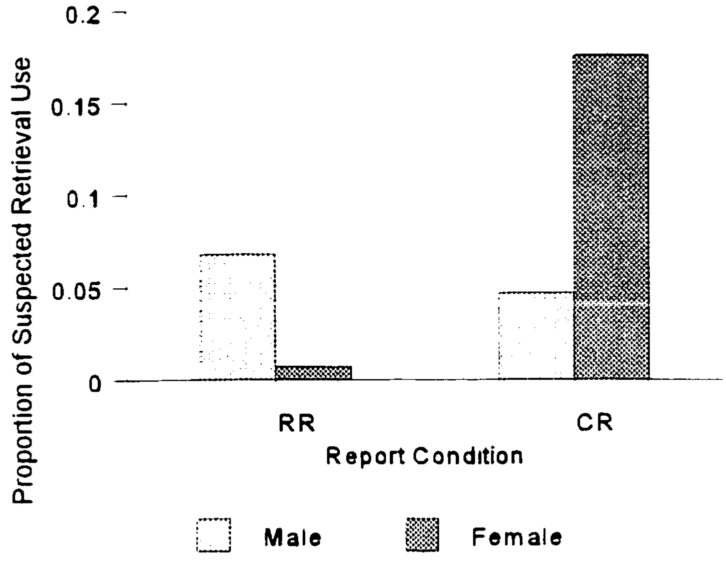

Figure 7. Interaction between Report Condition and Gender for Suspected Retrieval Use in Grade 3.

In Grade 5, significant differences involving Report Condition were found in two of the strategy categories. In the retrieval category, as previously reported, there was an interaction between report condition and gender (see Figure 5). The proportions of frequency of the retrieval category were lower in the RR than in the CR condition for males, but for females, there was more retrieval use in the RR condition. Within the RR report condition, females had higher frequencies of use than males, and within the CR condition, there was no difference between males and females. Finally, in the suspected retrieval category, similar to the Grade 3 students, Grade 5 students in the CR condition had higher proportions of suspected retrieval than in the RR condition, (.35 vs .09), $F(1, 36) = 6.93$, $p = .012$.

Thus, in Grade 5 two strategy categories had differing frequencies of use depending on report condition. Thus, there is moderate evidence that either or both of the two main components of verbal report validity, reactivity and veridicality, have been jeopardized. In Grade 1 no differences involving report condition were found but in Grade 3, three categories had differences involving report condition. Therefore, there is some support, at least for the older students, for the hypothesis that providing a verbal report can either have a reactive effect on task performance or lead to non-veridical strategy reports. Whether these effects involve either or both of the report conditions, however, cannot be determined.

As with the accuracy and latency data, the above results may be misleading due to the number of students in each grade who reported retrospectively rather than concurrently at least once in the CR condition. Thus, analyses were redone using 3 Report Condition groups: RR, CR, and RR in CR (see Table 6) (see Appendix J for a more detailed table of means). Analyses should be interpreted cautiously because the number of students remaining in the CR group was only 6 for all grades. Proportions of reported frequencies were subjected to a 3 (Report Condition: RR, CR, and RR in CR) x 2 (Gender: male and female) x 4 (Problem Types: 1 through 4) analysis of variance.

Table 6

Proportions of the Frequency of the Strategy Categories Reported by the Retrospective (RR), Concurrent (CR), and Retrospective in Concurrent (RR in CR) Report Condition Students as a Function of Grade and Problem Type

Grade	Report Condition	Strategy Category: Counting	Retrieval	Suspected Retrieval	Derived Fact	Special Trick	Addition Fact	Unknown	Guess	Other
1[a]										
	RR	.43	.13	.07	.01	.11	.01	.15	.07	.03
	RR in CR	.45	.11	.09	.00	.15	.01	.14	.02	.02
	CR	.57	.19	.05	.00	.05	.00	.14	.00	.00
	Mean	.46	.13	.07	.01	.12	.01	.14	.04	.02
3[b]										
	RR	.24	.16	.04	.02	.42	.06	.04	.00	.02
	RR in CR	.17	.24	.11	.03	.24	.13	.09	.00	.00
	CR	.13	.03	.11	.02	.18	.33	.16	.00	.04
	Mean	.20	.17	.08	.02	.32	.13	.07	.00	.02
5[a]										
	RR	.23	.24	.09	.04	.25	.13	.01	.00	.01
	RR in CR	.07	.26	.50	.04	.08	.05	.00	.00	.01
	CR	.23	.30	.00	.00	.25	.17	.04	.00	.00
	Mean	.17	.26	.22	.03	.19	.11	.01	.00	.01

[a]ns = 20, 6, and 14 in the RR, CR and RR in CR groups. [b]ns = 19, 6 and 13 in the RR, CR, and RR in CR groups.

With the inclusion of the new report group, RR in CR, the number of main effects and interactions involving report condition increased. In Grade 1, there was an interaction between Report Condition and Gender, $F(2, 34) = 3.90$, $p = .030$, involving the proportion of retrieval use (see Figure 8). There was also a main effect of Gender, $F(1,34) = 9.11$, $p = .005$, in which males reported using retrieval more often than females (.27 vs. .07). However, within both the RR and RR in CR groups, males and females were not different in their reported retrieval use, although in the CR group, males reported retrieval more often than females. This difference in the CR group should be interpreted cautiously as it involves only 2 males and 4 females and means are near floor for females. Males had higher frequencies of retrieval in the CR group than in the other two groups (which were not different from each other). Females reported retrieval more often in the RR group than in the CR group but there were no other group differences.

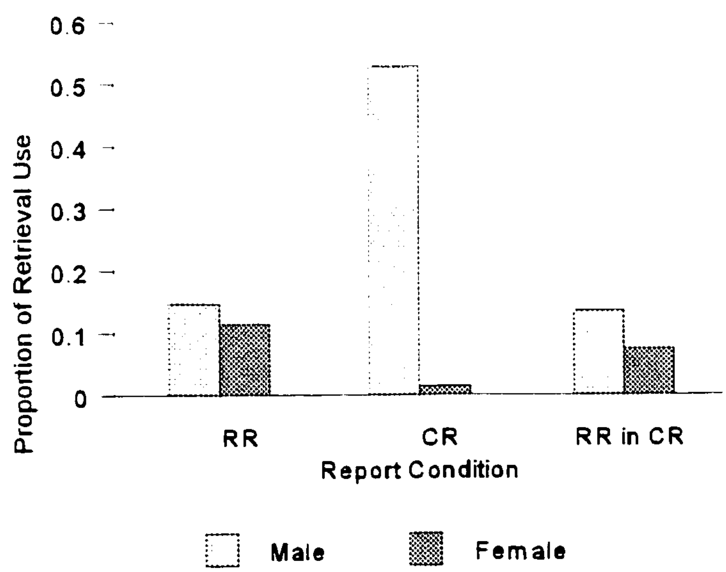

Figure 8. Interaction between Report Condition and Gender for Retrieval Use in Grade 1.

In Grade 3, main effects or interactions involving Report Condition were found for the same three strategy categories as before. In the special trick and addition fact categories, a main effect was found again, $F(2, 32) = 5.21$, $p = .011$ and $F(2, 32) = 7.33$, $p = .002$. In the special trick category, students in the RR condition had higher frequencies of reported use than in either the CR or RR in CR conditions (.43 vs. .17 and .24), while in the addition fact category, students in the CR condition had higher frequencies of reported use than in the RR and RR in CR conditions (.33 vs. .07 and .13).

In the suspected retrieval analysis there was an interaction between Report Condition and Gender, $F(2, 32) = 4.27$, $p = .023$. This interaction, however, was qualified by a three-way interaction between Report Condition, Gender, and Problem Type, $F(6\ 96) = 3.54$, $p = .003$. No main effect involving Report Condition was found, as in the previous analysis. In the 2-way interaction between Report Condition and Gender, RR males had higher frequencies of suspected retrieval use than CR males (whose means were at floor) and RR females had lower suspected retrieval frequencies than in either the CR or RR in CR groups (see Figure 9). There was no gender difference for the RR condition but in the RR in CR and CR conditions, females had higher proportions of suspected retrieval.

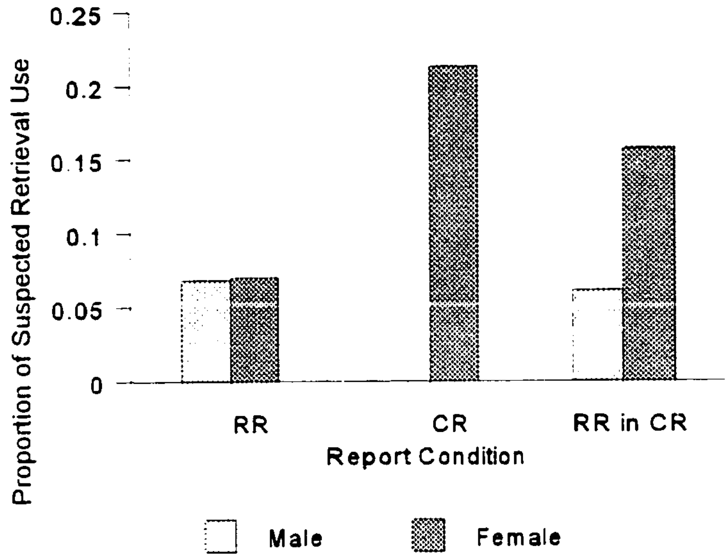

Figure 9. Interaction between Report Condition and Gender for Suspected Retrieval Use in Grade 3.

In the 3-way interaction involving type, no striking patterns involving report condition were found (see Table 7). Generally, the CR condition had the highest frequencies of suspected retrieval, followed by the RR in CR condition and then the RR condition. Only within the RR condition did the suspected retrieval frequencies for males exceed those of the females who usually had higher suspected retrieval use in the CR and RR in CR conditions. As with the previous analyses involving only the two report conditions, Type 1 problems tended to have a higher proportion of suspected retrieval than the other problem types, regardless of report condition.

Table 7

Interaction between Report Condition, Gender, and Problem Type for Grade 3 Students on the Proportion of Suspected Retrieval Strategy Use.

	Report Condition			
Gender	RR	CR	RR in CR	Mean
		Problem Type 1		
Male	.101	0	.064	.055
Female	.014	.321	.259	.198
Mean	.058	.161	.162	.120
		Problem Type 2		
Male	.051	0	.032	.028
Female	.022	.222	.204	.149
Mean	.037	.111	.118	.089
		Problem Type 3		
Male	.030	0	.064	.031
Female	.044	.037	.056	.046
Mean	.037	.019	.060	.039
		Problem Type 4		
Male	.091	0	.111	.067
Female	.078	.143	.111	.111
Mean	.084	.072	.111	.089

Note. Males: 11, 3, & 7 in RR, CR, RR in CR. Females: 8, 3, & 6 in RR, CR, RR in CR.

In Grade 5, significant differences involving Report Condition were found in two strategy categories. Once again, in the retrieval category, Report Condition and Gender interacted, $F(2, 34) = 5.62$, $p = .008$ (see Figure 10). For males, the proportions of frequency in the retrieval category were higher in the CR condition than in either the RR or RR in CR conditions, and the RR in CR condition reported retrieval more often than in the RR condition. For females, however, there was more retrieval use reported in the RR condition than in the other two conditions and frequencies were higher in the RR in CR condition than in the CR condition. Within the RR report condition, females had higher frequencies of use than males, within the CR condition males were higher in their use, and in the RR in CR condition there was no difference between males and females. A main effect of Report Condition was again found in the suspected retrieval category, $F(2,34) = 12.46$, $p < .001$, with no differences between the RR and CR conditions but both conditions had lower frequencies than the RR in CR condition (.09 and .03 vs. .50).

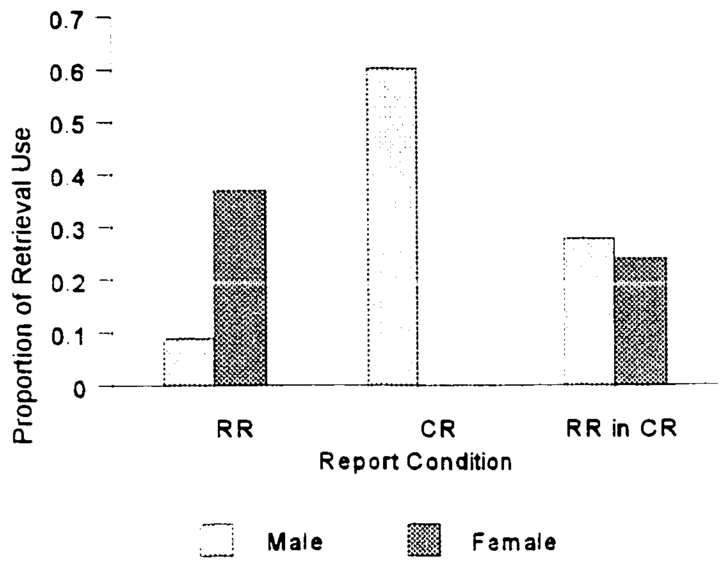

Figure 10. Interaction between Report Condition and Gender for Retrieval Use in Grade 5.

The results from the analyses involving the RR, CR, and RR in CR report conditions for the most part followed the same patterns as the analyses that were conducted only on the RR and CR report condition groups. Only one new significant effect was found in Grade 1. Therefore, when the students in the Concurrent Report condition were divided into two subgroups, there was only slightly stronger evidence for the inconsistency across verbal report conditions and there were only six subjects in one of the new report conditions.

Overall, the evidence for reactivity due to verbal reports is limited (see Table 8). Analyses of accuracy and latency data revealed only a few, and expected, reactive effects of verbal reporting, and analyses of strategy reports only yielded weak to moderate evidence that some type of invalidity may exist. Whether this evidence stemmed from task performance being affected by the instruction to verbally report or whether it was due to one or both report conditions yielding non-veridical reports, is indeterminable. Also, the issue of task performance being affected by the instruction to either concurrently or retrospectively report is not, according to the above analyses, a major cause for concern. However, there are enough small discrepancies to suggest that tests of reactivity should still be included when studying this type of task with verbal reports. In the next analyses, verbal reports are analyzed to determine whether they are veridical.

Table 8

Summary of Reactivity for Analyses Involving the Original Report Conditions (NR, RR, CR) and the Modified Report Conditions (NR, RR, CR, RR in CR).

		Grade		
Data	Conditions	1	3	5
Accuracy				
	Original	No	No	No
	Modified	No	No	No
Latency				
	Original	No	Yes (expected)	No
	Modified	No	Yes (expected)	Yes (expected)
Strategy Reports				
	Original	No	Maybe (3)	Maybe (2)
	Modified	Maybe (1)	Maybe (3)	Maybe (2)

Veridicality. A preliminary method for assessing veridicality of verbal reports is to determine whether reported strategies are associated with the expected accuracy and latency patterns across and within problem type. As well, certain strategies tend to be used more frequently on some types of problems, and less frequently on others. Obviously, not all strategy categories used in this study will be appropriate for this type of approach. For example, there is no theoretical basis for why the strategy category of "other" should have a certain latency or accuracy pattern. A second approach to assessing veridicality is to determine whether known predictors of subtraction performance (cf. Siegler, 1989; Woods, Resnick, & Groen, 1975) also predicted performance for students in either or both report conditions.

If verbal reports are veridical, then the reported strategies should correspond to known patterns of frequency, accuracy and latency. For example, retrieval is generally known to be a fast and accurate process while counting tends to be slower and more error prone. Because several problem types were included in this study, predictions also can be made about when a strategy will tend to be used. For example, retrieval would be expected to occur most frequently on "easy" problems (e.g., 6 - 1) and counting would be expected to occur most frequently on "harder" problems (e.g., 17 - 13). Thus, patterns of frequencies of strategy use, accuracy of strategy use, and median latency of correctly used strategies, were examined within and across problem type to determine whether they matched theoretical expectations.

Within each grade, report condition, and problem type, all reported instances of a strategy category for all subjects were aggregated together and frequency, accuracy, and the median latency on correct instances were calculated (see Appendix K). As outlined above, counting and retrieval have known patterns of frequency, accuracy, and latency. For suspected retrieval, of course, fast solution times and a high percentage of accurate responses would be expected. Because these two characteristics were the basis of being placed in this category, however, the comparison merely becomes circular with little value for establishing veridicality. For derived facts, special tricks, and addition facts, little is known about their expected accuracy and latency patterns, but frequency of use does tend to increase with age (Geary, 1994). As well, one might expect derived facts, special tricks, and addition facts to be used on harder problems, for which retrieval might not be an effective strategy for all students. Of these three strategy categories, special tricks was reported most frequently, and thus this category was included in some of the analyses. As mentioned previously, there is no expected pattern of accuracy, latency, or use for the other category. The same is true of the guessing and unknown categories, although one might expect guessing to occur more frequently on difficult problems. However, guessing was reported so infrequently that there is the possibility of differences being obscured or overemphasized by floor effects. Thus, the counting and retrieval strategies are the two categories that were examined in detail. Special trick strategies were also included but were examined in less detail.

Counting was the first strategy to be analyzed. A higher incidence of reported counting use was expected on harder problems, that is, Type 3 (e.g., 14 - 8) and Type 4 (e.g., 15 - 12) problems than on easier problems, that is, Type 1 (e.g., 6 - 1) and Type 2 (14 - 1) problems. More specifically, Type 1 problems were considered to be the easiest of the four problem types whereas the Type 3 problems were considered the most difficult. As reported previously, a 2 (Report Condition: RR and CR) x 2 (Gender: male and female) x 4 (Problem Types: 1 through 4) analysis of variance was performed on the frequencies of reported counting use. Each grade was analyzed separately.

In Grade 1, no effects involving gender or report condition were found. Thus, the means for both report conditions were similar and only the main effect of type is presented. As expected, Type 3 problems had the highest reported use of counting but the reported use for Type 1 problems, the "easiest" problems, was not different from the Type 3 problems (see Table 9 and Figure 11). Type 2 and 4 problems had the lowest reported counting use although Type 3 and 4 problems were not different from each other. One factor that may at least partially explain the unexpected pattern of results is the higher proportion of cut-offs that were implemented on Type 3 and Type 4 problems for Grade 1 students. Although cut-offs were only used on 9.9% of the 36 trials, on Type 3 problems they were used on 28.2% of the trials and on Type 4 were used on 16.1% of the trials (cf. 0.2 and 2.8% on Types 1 and 2, respectively). On these cut-off trials, students were only sometimes able to report what strategy they had been attempting (14.2 and 10.9% for Types 3 and 4, respectively). The percentage of reported counting use on problem types 3 and 4 may thus be an underestimate due to the fairly high percentage of cut-offs on which students were unable to provide any report of what strategy they were trying to use.

Table 9

Strategy Use and Accuracy (in Percentages), and Latencies (in Seconds) of Counting, Retrieval, and Special Tricks for Retrospective (RR) and Concurrent Report (CR) Students as a Function of Grade and Problem Type

	Grade																	
	1						3						5					
	Report Condition																	
	RR			CR			RR			CR			RR			CR		
Strategy	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc
	Problem Type 1																	
Counting	54.5	9.2	85.7	45.6	8.7	79.3	25.1	3.2	95.3	6.4	4.3	90.9	21.1	1.9	89.5	8.3	2.7	100.0
Retrieval	22.2	3.4	92.5	18.9	3.7	97.1	29.8	1.6	97.1	28.1	2.5	100.0	35.6	1.1	98.4	32.2	1.5	96.6
Spec. Trick	6.7	3.8	100.0	8.3	8.3	80.0	18.1	2.5	100.0	9.4	3.2	93.8	11.1	1.6	100.0	8.3	6.0	73.3

table continues

	Grade																	
	1						3						5					
	Report Condition																	
	RR			CR			RR			CR			RR			CR		
Strategy	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc
Problem Type 2																		
Counting	40.0	8.1	59.7	43.3	4.1	61.5	24.0	3.8	95.1	17.5	5.4	93.3	23.3	2.2	92.9	8.9	3.1	87.5
Retrieval	19.4	4.1	97.1	15.0	3.4	88.9	12.3	1.8	100.0	16.4	2.8	96.4	26.1	1.5	89.4	26.1	1.8	95.7
Spec. Trick	17.8	5.1	65.6	17.2	10.3	93.5	57.3	2.9	92.9	28.7	5.4	91.8	33.9	1.8	93.4	16.7	4.9	90.0
Problem Type 3																		
Counting	43.9	13.1	38.0	60.6	19.3	42.2	32.2	7.0	81.8	20.5	7.0	94.3	25.6	3.1	91.3	20.0	6.3	88.9
Retrieval	3.9	3.3	92.9	8.3	4.4	66.7	12.3	1.8	90.5	13.5	2.9	100.0	15.6	1.5	92.9	20.0	1.8	88.9
Spec. Trick	6.7	14.9	33.3	6.1	15.3	81.8	26.7	2.9	95.7	13.5	9.0	87.0	18.9	2.4	94.1	9.4	3.9	100.0

table continues

	Grade																	
	1						3						5					
	Report Condition																	
	RR			CR			RR			CR			RR			CR		
Strategy	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc
Problem Type 4																		
Counting	34.4	15.4	38.7	45.6	20.5	53.7	12.9	8.9	77.3	17.5	8.3	77.1	22.2	2.3	85.0	9.4	4.2	94.1
Retrieval	6.7	3.3	75.0	10.6	3.5	73.7	11.1	1.8	100.0	10.5	2.8	100.0	20.0	1.3	97.2	29.4	1.3	90.6
Spec. Trick	13.9	7.4	60.0	17.2	10.4	58.1	65.5	3.1	83.0	35.1	4.8	32.2	37.2	1.9	91.0	17.8	5.7	90.6
All Problems																		
Counting	43.2	10.3	58.2	48.8	12.7	57.8	23.5	4.5	88.2	15.5	7.0	92.5	23.1	2.3	89.8	11.7	4.3	91.7
Retrieval	13.1	9.4	88.3	13.2	3.5	85.3	16.4	1.6	97.3	17.1	2.7	99.1	24.3	1.3	94.9	26.9	1.6	93.3
Spec. Trick	11.3	5.2	64.2	12.2	10.4	77.3	42.0	2.9	87.8	21.6	5.1	93.9	25.3	1.9	93.4	13.1	5.1	89.4

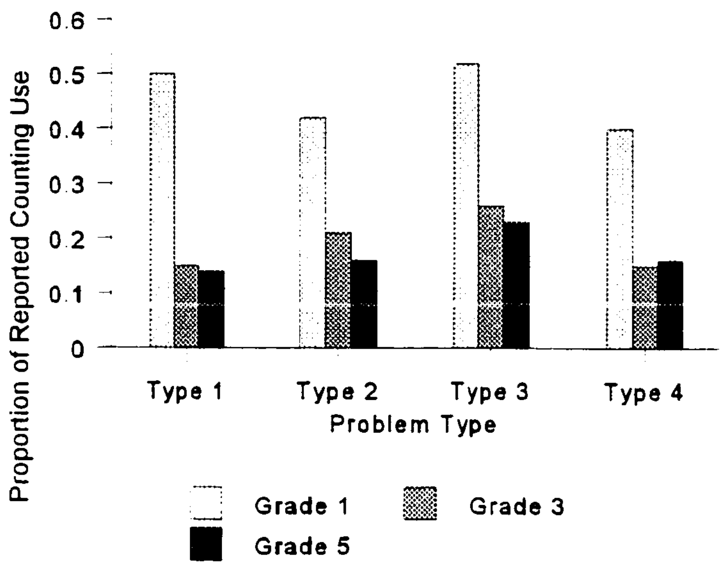

Figure 11. Proportions of Reported Counting Use in Grades 1, 3, and 5.

In Grade 3, no effects involving gender or report condition were found. Problem type was a significant factor. Although problem types 1, 2, and 4 were not different from each other, counting was reported less often on Type 1 and 4 problems. Type 3 problems had the most reported counting use, as was expected. In Grade 5, once again only a main effect of type was found. Type 3 problems were reported more frequently than on Type 1, 2, and 4 problems, which did not differ from each other. Thus, in all grades, as expected, Type 3 problems had the highest reported use of counting. Grade 1 students did not have the lowest reported counting use on Type 1 problems, although the results may have been at least partially affected by the high frequency of cut-offs on Type 3 and 4 problems. In Grades 3 and 5 reported counting use was lowest on Type 1 problems, as expected, though the means did not differ significantly from all other problem types. Overall, patterns of reported counting provided at least a fair match with expectations in all grades.

As will be discussed later, a third strategy category, not as well documented as counting and retrieval, may account for some of the results found above.

Although analyses were performed on each grade separately, a higher frequency of reported counting use overall was expected for Grade 1 students as counting is the most common approach to solving subtraction problems for younger children (Geary, 1994). Indeed, Grade 1 students reported using counting on 43.2 and 48.8 %, respectively, for the RR and CR groups whereas the Grade 3 and Grade 5 students reported using counting on only 23.5 and 15.5% for the Grade 3 RR and CR groups and on 23.1 and 11.7% of trials for the Grade 5 RR and CR groups. More support for the veridicality of verbal reports is established by the finding that patterns of frequency across grade followed expected patterns.

Next, counting trials were compared in terms of solution latencies. Latencies were expected to be longer for harder problems than for the easier problems because more counting would be involved. Type 3 and 4 problems both have large minuends and the subtrahends are larger on these problems than on Type 1 and 2 problems. The finding that solution times increase as problem size increases is a strong and pervasive finding in the literature (e.g., Ashcraft, 1982) and is known as the problem-size effect. Thus, for counting trials, the pattern expected for latencies would be longer latencies on Type 3 and Type 4 problems than on Type 1 and 2 problems.

One approach would be to compare the latencies of counting trials to retrieval trials for students who reported using both strategies. Presumably, if reports are veridical, solution latencies on counting trials would become longer as problem type became larger or more difficult whereas solution latencies on retrieval trials would remain fairly flat across problem type. However, this within-subject statistical comparison was not possible as few students reported using both strategies frequently enough within a problem type to permit an analysis of variance. Consequently, an examination of group means was conducted to determine whether they follow the expected patterns.

In all grades and both report conditions, the same general pattern was found (see Figure 12). In Grade 5, solution latencies on all problem types were fast and differences were small, therefore comparisons must be interpreted cautiously. Type 1 and 2 problems had shorter latencies than Type 3 and 4 problems. Overall, median latencies in all grades and report conditions followed expected general patterns and therefore provide more evidence for the veridicality of the students' verbal reports.

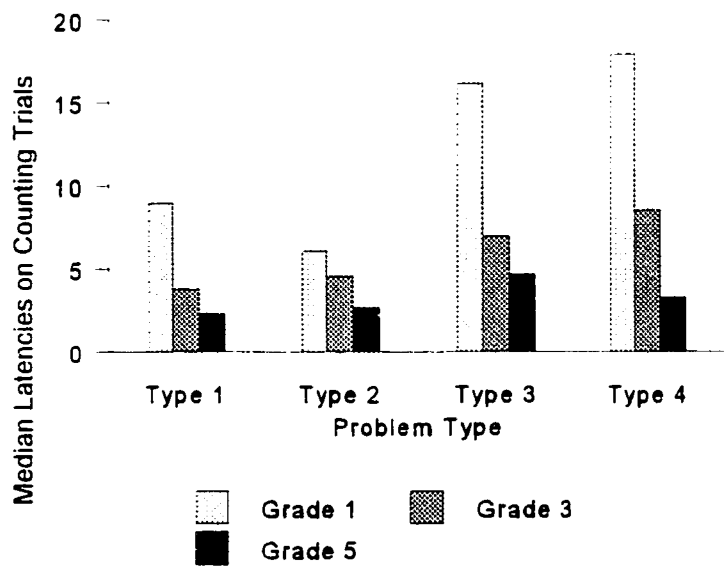

Figure 12. Median Latencies on Counting Trials across Problem Types in Grades 1, 3, and 5.

Finally, on counting trials, more errors were expected on more difficult problems than on easier problems. Counting requires more time and effort and thus, more opportunities for errors also occur. Type 1 and 2 problems should have fewer errors than Type 3 and 4 problems because the former problem types require less counting than the latter problem types. One caveat is that Grade 3 and 5 students were quite close to ceiling and thus expected differences may be obscured by ceiling effects. In both Grade 1 and 3 report conditions, the expected patterns of accuracy were found (see Figure 13). Type 1 and 2 had the highest percentages of accurate counting and Types 3 and 4 had lower percentages of accuracy, with Type 3 having the lowest percentage. In Grade 5, the RR condition had the highest accuracy on Type 2 and 3 problems and in the CR condition, Type 2 problems had the lowest accuracy. Overall, though, in Grade 5, the distribution of accuracy was quite flat and very close to ceiling on all problem types but in Grades 1 and 3 patterns of accuracy supported the veridicality of the verbal reports.

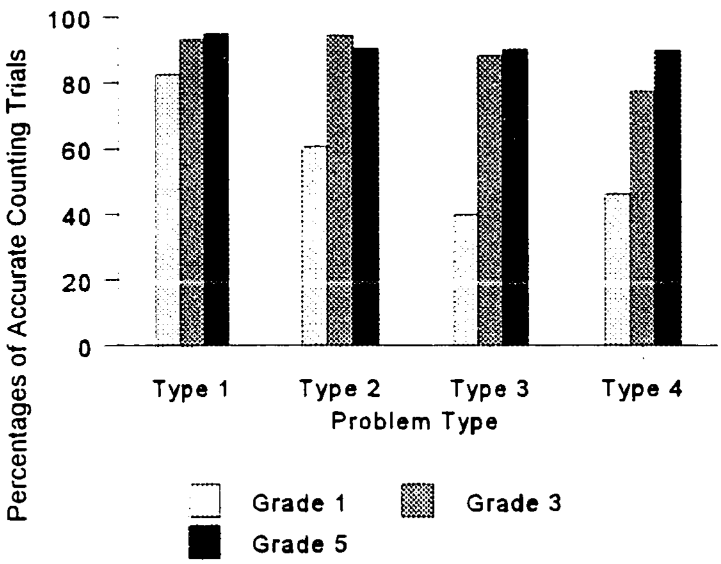

Figure 13. Percentages of Accurate Counting Trials across Problem Types in Grades 1, 3, and 5.

To summarize, on reported counting trials, analyses and comparisons are generally consistent with the view that verbal reports counting are veridical. In Grade 1, counting was reported most frequently when expected. On the problems for which counting was expected to be least frequent, however, Grade 1 students had frequencies of reported counting use that did not differ from the most difficult problems. Possible reasons for this finding include the large number of cut-offs on Type 3 and 4 problems, thus frequency on these problems could be an underestimate of counting use and may have obscured otherwise significant differences. As well, students in Grade 1 may also have used less well-known backup strategies than counting to solve certain problem types. This hypothesis will be addressed later. More evidence for the veridicality of verbal reports in Grade 1 was that expected problem-size effects were found in the accuracy and solution latencies of reported counting uses with more errors and slower solution times on more difficult problems.

Grade 3 students reported counting most frequently and least often when expected and predicted problem-size effects on accuracy and latency data were also found. In Grade 5, accuracy percentages were very close to ceiling and flat across problem type. As well, solution latencies were very fast across all problem types. Thus, for the most part, establishing the veridicality of reported counting use was most difficult to establish in the oldest students. However, Grade 5 students did report using counting most frequently on the problems which were expected to yield the most counting use and least frequently on problems which were expected to yield the least counting use. Additionally, reported counting decreased with grade, as expected. Finally, no Report Condition differences were found. Thus, with few exceptions, the evidence for the veridicality of concurrent and retrospective reports was supported, especially in Grade 3. Next, retrieval analyses and comparisons are presented.

Retrieval trials were expected to have higher frequencies of reported use on easier problems (Types 1 and 2) than on harder problems (Types 3 and 4). More specifically, Type 1 problems were considered the easiest of the four problem types and Type 3 the most difficult. As reported previously, a 2 (Report Condition: RR and CR) x 2 (Gender: male and female) x 4 (Problem Types: 1 through 4) analysis of variance was performed on the frequencies of reported retrieval use for each grade. A main effect of type was found in all grades. In Grade 1, there was no effect of report condition. Retrieval was reported more frequently on Type 1 problems, as predicted, than on the remaining problem types.

Type 2 problems had higher reported use than Type 3 and 4 problems which did not differ from each other (see Figure 14). In Grade 3, Type 1 problems also had the highest reported frequencies but no other differences were found. In Grade 5, once again Type 1 problems had the highest reported use. Type 2 and 4 problems did not differ from each other but had higher reported retrieval use than Type 3 problems. An interaction between report condition and gender was also found in Grade 5 which has been reported previously (see Figure 5). Overall, results matched expected patterns very closely, lending support to the veridicality of reported retrieval use for all grades and report conditions.

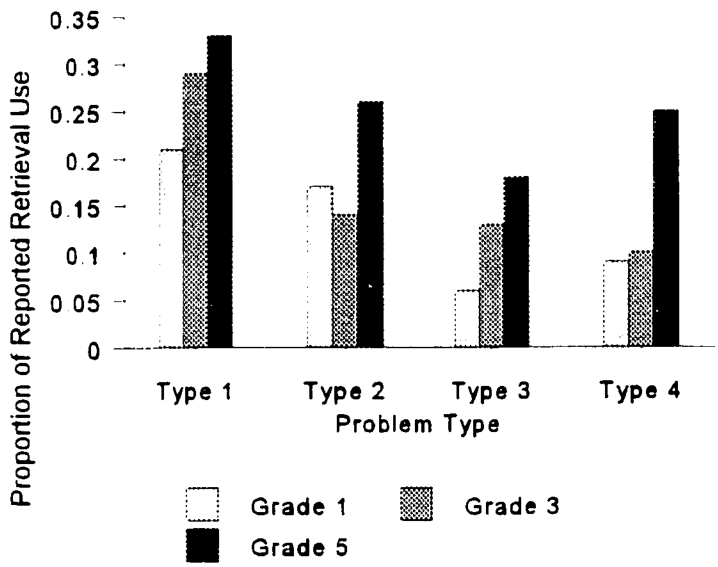

Figure 14. Proportions of Reported Retrieval Use in Grades 1, 3, and 5.

A higher frequency of reported retrieval use was expected for the older students (Geary, 1994; Lemaire, Barrett, Fayol, & Abdi, 1994). The overall means for each grade reported in Table 9 matched this expectation, though analyses were conducted separately for each grade. Students in Grade 5 reported using retrieval on 24.3 and 26.9% of all trials for the RR and CR groups, respectively. Students in Grade 3 had slightly lower reported use of retrieval with 16.4 and 17.1% for the RR and CR conditions while Grade 1 students had the lowest reported frequencies of retrieval use with 13.1 and 13.2% for the RR and CR groups.

Retrieval trials were then compared to determine whether there were differences in solution latencies for different problem types. Retrieval tends to be fast and may be automatic. Therefore, unlike counting, a strong problem size effect is not expected for different problem types although even with retrieval, latencies are sometimes slower on harder problems (LeFevre, Sadesky, & Bisanz, 1996). Report condition differences were minimal and therefore both report groups are shown together in Figure 15. The difference in both report conditions for each grade was, at most, one second. Overall, median latencies differed very little across problem type in all grades, as expected, regardless of report condition.

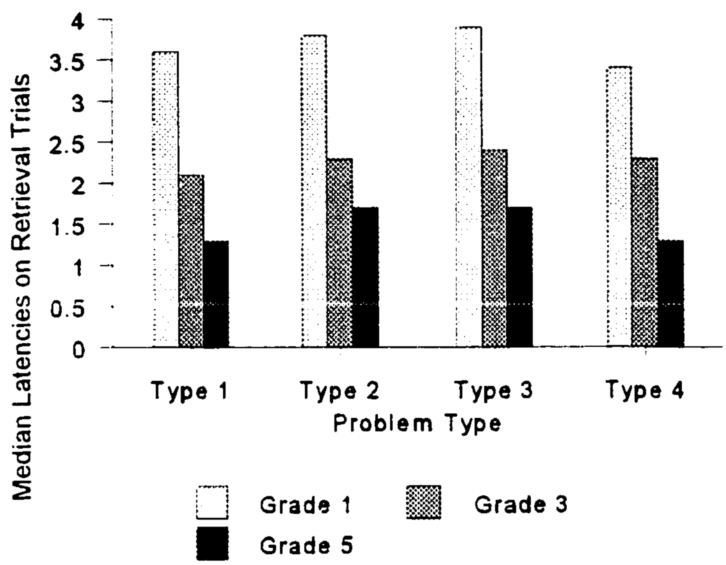

Figure 15. Median Latencies on Retrieval Trials across Problem Types in Grades 1, 3, and 5.

Retrieval not only tends to be fast, it also tends to be accurate. Therefore, as with the median latencies, no large problem type differences are expected although the harder problems may have lower percentages of accuracy than the easier problems. Once again, ceiling effects must be taken into consideration, especially in the older grades. Unlike median latencies, more marked differences were found in accuracy across problem types in Grade 1. However, the differences between the most accurate and least accurate problem type were much smaller for reported retrieval than for counting. For example, the problem type difference in accuracy for counting in the Grade 1 RR group was 47.7% and the difference for retrieval was only 22.1%. Therefore, even though there is a problem type effect on retrieval trials, it is a much smaller effect than that found on counting trials. In both report conditions, accuracy was higher on the easier problems than on the harder problems (see Figure 16). In Grade 3, accuracy was at or very near ceiling for all problem types in both report conditions. Grade 5 students were not quite as accurate as Grade 3 students, overall, but in both report conditions Type 1 problems had very slightly higher accuracies than the other problem types. Generally, accuracies matched expected patterns quite well but problem type differences were much smaller than the problem type differences found on counting trials. To summarize the retrieval findings, patterns were an even better fit for the reported retrieval data than for the reported counting data. Thus, evidence supporting the veridicality of both types of verbal reports continues to be positive.

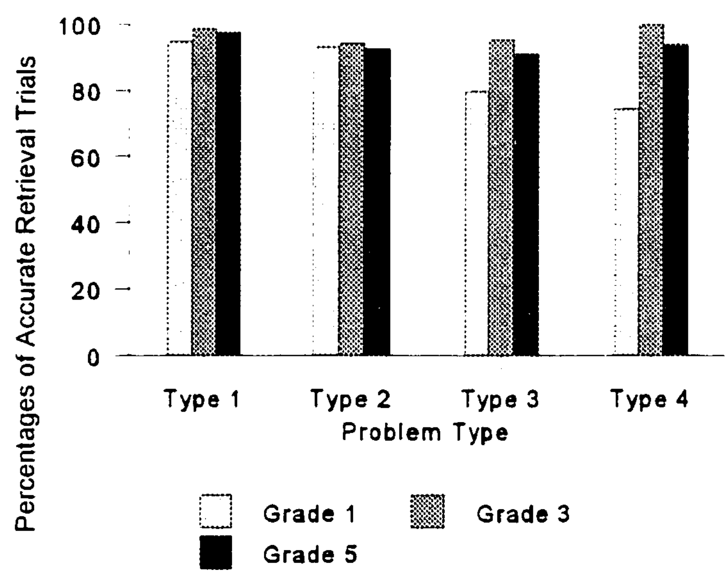

Figure 16. Percentages of Accurate Retrieval Trials across Problem Types in Grades 1, 3, and 5

Special tricks might tend to be used on trials in which the use of retrieval may be too difficult or impossible. Special tricks could also be considered an alternative to counting as a backup strategy and thus might be used on problems that are not difficult enough to require the use of a counting strategy. As a group, special tricks are strategies that are more sophisticated than counting and yet are not as efficient as retrieval. Thus, special tricks were expected to be reported more frequently on problems with "in-between" difficulty, although in general, reported use would be expected to be higher on harder problems than on easier problems.

A 2 (Report Condition: RR and CR) x 2 (Gender: male and female) x 4 (Problem Types: 1 through 4) analysis of variance was performed on the frequencies of reported special tricks use for each grade, as reported previously. In all grades, special tricks were reported more frequently on Type 2 and 4 problems than on Type 1 and 3 problems (see Figure 17). Type 2 and 4 problems consistently had mid-range levels of errors and medium solution times compared to Type 1 and 3 problems and thus, the more frequent special trick use on these medium-difficulty problems lends at least partial support for the veridicality of the strategy reports. As well, the consistency of frequencies across grade itself lends support to the hypothesis that students are reporting veridically. That is, if students were reporting the use of special tricks in a non-veridical fashion, their reported use might be expected to be randomly spread across all problems. Also, remember that retrieval is most likely to be reported on Type 1 problems and counting is most likely to be reported on Type 3 problems. Thus, this tendency for other strategies to be used on Type 1 and 3 problems may at least partially account for the lower reported usage of special tricks on these problems. A main effect of report condition was found in Grade 3 with, as reported earlier, students in the RR condition reporting more special tricks use than the

CR condition students. Overall, because expectations about special tricks were quite speculative, the results across problem type and grade provide only partial support of the veridicality of verbal reports, whether concurrent or retrospective. However, students were so consistent across grade and report condition on what problem types they reported using special tricks that this consistency itself supports the veridicality of both kinds of verbal reports.

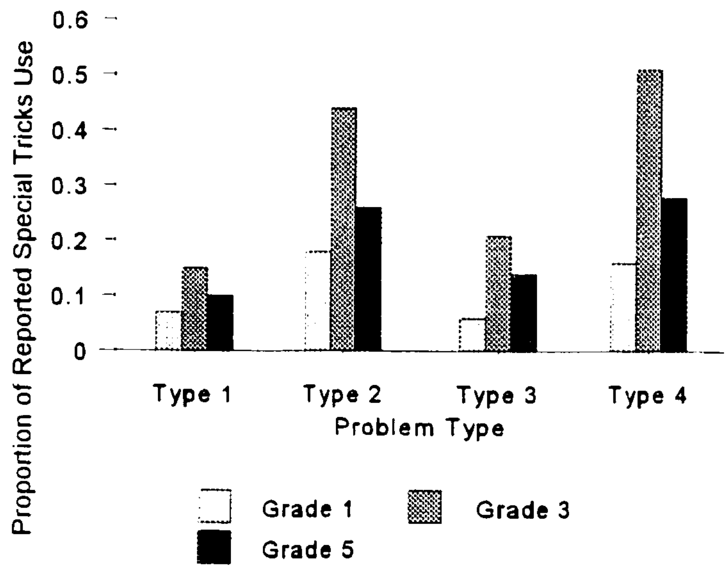

Figure 17. Proportions of Reported Special Tricks Use in Grades 1, 3, and 5.

Although there is little basis for predicting patterns of accuracy and latency data on different problem types for special tricks, how strategy use changes across development has been postulated. Geary (1994) states that counting tends to be the predominant strategy for students who are learning to subtract, later followed by derived facts or special tricks (decomposition), and then finally retrieval tends to become the dominant strategy. Grade 1 students had the lowest frequency of reported special tricks use (see

Table 9). Grade 3 had the highest frequency of reported use but as noted just above, students in the RR condition reported using special tricks more frequently than CR condition students (42.0 and 21.6% for RR and CR, respectively). Finally, Grade 5 had "medium" frequencies of reported use. Despite a significant difference in report condition in Grade 3, reported usage was still higher in the CR condition in Grade 3 than in Grades 1 or 5. Thus, frequency patterns across grade matched expectations well.

Overall, the analyses and examinations of the data associated with reported counting, retrieval, and special tricks provided generally supportive evidence for the veridicality of verbal reports, regardless of verbal report type. Few exceptions were found for any reported strategy category, for any grade, and for any report conditions. Thus, the evidence presented here indicates that both types of verbal reports are veridical sources of data. As a further test of veridicality, regression analyses were conducted using established predictors to account for solution times overall and for counting, retrieval, and special trick trials. Though there were sound theoretical reasons for these analyses, in reality these analyses yielded intractable and contradictory results.

Regression analyses were performed to determine whether predictors of solution times matched reported strategies in either or both report conditions. Predictors of performance were taken from earlier subtraction studies and are considered possible models of children's subtraction (Siegler, 1989; Woods, Resnick, & Groen, 1975). For problems of the form $\underline{a} - \underline{b}$, the predictors consisted of the minuend (the larger number, $\underline{a}$), the subtrahend (the smaller number, $\underline{b}$), the difference between the minuend and subtrahend ($\underline{d}$), the sum of the minuend and subtrahend ($\underline{a} + \underline{b}$), the prediction of the smaller-count model (small count), and whether the subtrahend was larger than 10 (b > 10). The smaller-count model postulates that students will either count up or count down, whichever requires the smallest number of counts. For example, the smaller count for 12 - 1 is 1 and for 12 - 9 is 3. That is, less effort is required to count down 1 from the 12 on the first problem and to count up from 9 to the 12 on the second problem. All predictors, though, are measures of problem size effect, and as such, if students reported counting, the latencies should be related to at least some of the structural predictors. On retrieval trials the predictors should not account for as much of the variance because problem-size effects are not as strong on retrieval trials (e.g., LeFevre, Sadesky, & Bisanz, 1996). On special trick trials, expected effects of the structural predictors were not known. Siegler reported that analyses of strategies such as retrieval, deleting 10s (which was part of the special tricks category in this study), and addition facts "lent considerable convergent validity to the strategy assessments" (p.501), but no specific expectations or data were presented.

Regression analyses of median latencies were done on all trials, counting trials, retrieval trials, and special trick trials (see Appendix L, Table L1). Each grade was analyzed separately as was each report condition. Of special note was the use of two criteria for the calculation of median latencies of each problem. Traditionally, regression analyses involving median latencies are based on correct trials only (e.g., LeFevre, Sadesky, & Bisanz, 1996). Woods, Resnick, and Groen (1975) used correct responses only to calculate latencies as "there is no reason to believe that subjects used any of the proposed models [of subtraction] when they generated an incorrect answer" (p.19).

However, Siegler (1989) used both correct and incorrect responses to calculate the median latencies "because children erred on a relatively high percentage of trials (16%), because the errors were generated by the same strategies as the correct answers, and because removing trials with errors would seriously bias estimates of solution times on the most difficult problems" (p.501). There are merits to both approaches, therefore analyses were done both ways. Analyses were first done using correct responses only (C/O) and then using both correct and incorrect responses (C/I). In the analyses involving specific strategy categories, the strategy had to be reported on a problem by at least three students for the problem to be included. Depending on the criteria for inclusion in the calculation of the median latency (correct trials only or both correct and incorrect trials), the number of problems in the analysis sometimes differed.

The results of the analyses were mixed. Structural predictors should account for more variance on counting trials than on retrieval trials. In Grade 1, this expectation was met in the CR condition but not the RR condition. In Grade 3, the RR condition conformed to expectations but the CR condition did not. In Grade 5, both RR and CR conditions met expectations, but only weakly. As well, results sometimes diverged dramatically depending on which criterion was used for the analysis, correct responses only (C/O) or both correct and incorrect responses (C/I). For example, in the Grade 1 RR condition, no predictors accounted for a significant amount of variance in the median latencies on counting trials when C/O responses were used and yet all six predictors were significant when C/I responses were used. Lastly, most of the predictors are based on the assumption that students who are counting are using counting up or counting down

(Siegler, 1989b). However, in this study counting up was reported rarely and thus, unlike Siegler's study, comparisons between reported counting up and counting down trials were not conducted (see Appendix L, Table L2 for the frequencies of each of these two strategies). Indeed, the counting category consisted of 20 separate counting strategies thereby making predictions about which predictors should account for more variance difficult.

Thus, the results of the regression analyses were neither straightforward or compelling. Results were inconsistent across grade, unlike the previous tests of veridicality. Results sometimes varied dramatically depending on the criteria used for inclusion in the calculation of median latencies with no theoretical reason to favour one criteria over the other. Counting up and counting down trials were too infrequent to conduct regression analyses on leading to the aggregation of many counting strategies. This large number of counting strategies in the counting category made predictions difficult and results difficult to interpret. And finally, in some cases only a few of the 36 problems were included in the regression analyses because the few students in some grade and report conditions rarely reported using some of the strategies. For example, only 11 problems were included in the Grade 5 CR group regression analysis of counting trials. Overall, regression analyses were a theoretically sound approach to testing veridicality but in practice led to ambiguous results that were difficult to interpret.

The issue of veridicality was investigated using analyses of variance and examinations of data patterns. A summary of veridicality results is shown in Table 10. Each result expected if verbal reports are veridical is presented in the table. For example, in the first two lines of the counting results, if verbal reports are veridical then the most counting would be expected on the hardest problems, i.e. Type 3 problems, and conversely, the least counting would be expected on the easiest problems, i.e., Type 1 problems. In Grade 1, however, this expectation was not met because although students reported the most counting on Type 3 problems, as expected, Type 2 and 4 problems had less reported counting use than Type 1 problems, the easiest problems. Thus, though theoretically the results for both lines (and other pairs of results in Table 10) should be the same, in practice they sometimes were not.

Table 10

Summary of Veridicality for the Retrospective (RR) and Concurrent (CR) Report Conditions

	Grade					
	1		3		5	
	Report Condition					
Expected Results if Reports Veridical	RR	CR	RR	CR	RR	CR
Counting						
More counting on hard problems	Yes	Yes	Yes	Yes	Yes	Yes
Less counting on easy problems	No	No	Yes	Yes	Yes	Yes
Counting decreases across grade	43.2%	48.8%	23.5%	15.5%	23.1%	11.7%
Longer counting latencies on hard problems	Yes	Yes	Yes	Yes	Yes	Yes
Shorter counting latencies on easy problems	Yes	Yes	Yes	Yes	Yes	Yes
More counting errors on hard problems	Yes	Yes	Yes	Yes	No[a]	No[a]
Fewer counting errors on easy problems	Yes	Yes	Yes	Yes	No[a]	No[a]
Retrieval						
More retrieval on easy problems	Yes	Yes	Yes	Yes	Yes	Yes
Less retrieval on hard problems	Yes	Yes	No	No	Yes	Yes
Retrieval increases across grade	13.1%	13.2%	16.4%	17.1%	24.3%	26.9%
Almost equal latencies on all problems	Yes	Yes	Yes	Yes	Yes	Yes
Almost equal accuracies on all problems	No	No	Yes[a]	Yes[a]	Yes[a]	Yes[a]
Special Tricks						
More special tricks on "medium" problems	Yes	Yes	Yes	Yes	Yes	Yes
Special tricks most frequent in middle grade	11.3%	12.2%	42.0%	21.6%	25.3%	13.1%

[a] ceiling effects

Overall, the notion that students were reporting veridically was supported. Students, with very few exceptions, consistently reported using counting and retrieval on expected types of problems, with expected patterns of accuracy and latency. There were fewer predictions made about special tricks except regarding their prevalence in the students who no longer relied on counting and yet were not primarily retrieving yet. As well, the problems on which special tricks were reported was exactly the same for all grades and all report conditions. Very few results inconsistent with the veridicality of verbal reports were found in any of the results for counting, retrieval, and special tricks. Therefore, whether students report concurrently or retrospectively does not appear to affect the veridicality of their reports. Finally, students in all grades had veridical reports.

Strategy report overview. As with accuracy and latency data, findings regarding grade, gender, and problem type were mostly expected. Grade 1 students relied most heavily on counting, Grade 3 students reported using special tricks most frequently, and Grade 5 students retrieved. Few gender differences were found. None were found in Grade 1, but in Grade 3 and 5 gender interacted with both report condition and problem type on a few strategy categories. No consistent explanation or pattern emerged from these interactions. Problem type was a significant factor in all grades and in all report conditions.

The reactivity results were problematic. Strategy report frequencies were compared for the retrospective and concurrent report condition. In Grade 1, students in both conditions reported using each of the analyzed strategy categories equally frequently. In Grade 3, three of the five analyzed categories had main effects or interactions involving report condition. In Grade 5, two of the five analyzed categories had report condition main effects or interactions. The data were re-analyzed with the three report conditions, retrospective, concurrent, and retrospective in concurrent, to determine whether the previous results were due to students failing to follow task instructions. These new analyses provided only slight modifications of the previous results. Thus, in Grades 3 and 5 students in each report condition did not report the same strategies with equal frequency.

Problem type effects found previously were examined in detail in the veridicality section to see whether students reported using counting, retrieval, and special tricks more or less frequently on certain problem types. A good match was found between hypothesized and actual results. Patterns of accuracy and latency for each strategy category also matched. Overall, the results for veridicality were positive, regardless of report condition or grade.

Subtraction Summary

Three measures were used to investigate grade, gender, problem type, reactivity, and veridicality issues on the subtraction problem solving task. Students in Grade 1 were slower and less accurate than the older students and relied more heavily on counting to solve the subtraction problems. Grade 3 and 5 students made few errors, and Grade 5 students were faster than the Grade 3 students. In Grade 3, students used special tricks most often to solve the subtraction problems and the Grade 5 students reported using retrieval most often. Few gender differences were found. Problem type effects were found in almost all grades and on all measures. Generally, on the more difficult problems, more errors were made and solution latencies were slower and more reliance was placed on either counting or other non-retrieval strategies such as special tricks.

No reactive effects were found on the accuracy data in any report condition or grade. Solution latencies were expected to be slower in the concurrent report condition than in the retrospective report condition. However, only in Grade 3 was there a significant slowing down for the concurrent report condition group. Because students in the concurrent report condition often had difficulty providing a concurrent report and often reported retrospectively instead, analyses were re-done with three report conditions: retrospective reporters, concurrent reporters, and retrospective reporting in the concurrent report condition. In Grades 3 and 5, "pure" concurrent reporters had slower solution latencies. However, in Grade 1, the concurrent reporters were still no slower. This grade effect may be due to the younger students being generally slow and often spontaneously talking or thinking aloud while they solved the problems. Reactivity effects were harder to assess for the strategy reports because any difference found in frequencies of reported strategy use could either reflect reactivity or non-veridicality. Grade 1 students reported each of the analyzed strategies with equal frequency, regardless of report condition. Grade 3 and 5 students, however, did not have equal frequencies in both report conditions for all of the analyzed strategy categories. Therefore, for strategy reports, the validity of one or both of the report conditions is at least somewhat affected in the older students.

Veridicality was assessed in the strategy report section but many of the analyses depended on the accuracy and latency data. The results for veridicality were generally positive in all grades and all report conditions. Consistently, frequency of strategy use, patterns of accuracy and latency for each strategy, and the effect of predictors on median latencies for each strategy category matched expectations. Therefore, in the subtraction problem solving task, the only problematic finding for the validity of verbal reports was that when frequencies of strategy reports were compared in Grades 3 and 5, students in the RR and CR conditions did not report using all strategies equally often. However, because there was strong converging evidence for the veridicality of verbal reports of either type, the discrepancy in frequencies was most likely due to reactivity.

Procedure Recognition Task Data

The procedure recognition task consisted of two parts: recognition of a legitimate or illegitimate procedure, and justification of why the procedure was considered legitimate or illegitimate. For the recognition component, the proportion of correct responses divided by total responses was calculated. For the justification component, the proportion of appropriate justifications divided by the total was calculated. Note that the students did not have to get the recognition portion "correct" to have an appropriate justification of their response. To reflect this independence between the recognition and justification scores, a third measure was included: the number of trials on which recognition responses and justifications were both correct and appropriate divided by the total (see Appendix M for examples of both correct and incorrect recognition responses and appropriate and inappropriate justifications).

Reliability for the classification of appropriate and inappropriate responses on the justification of the procedure recognition task was calculated because classification involved a judgment on the part of the coder. Reliability was calculated as the number of agreements divided by the total number of judgments. Reliability for the justification component was 93.5%. Disagreements were resolved by discussion between the two coders.

<u>**Grade and Gender**</u>

Proportions of correct recognition, appropriate justification, and both correct recognition and appropriate justification, were each subjected to a 3 (Grade: 1, 3, and 5) x 2 (Gender: male and female) analysis of variance. On each of the three measures, there was a main effect of grade, $F(2, 160) = 27.22$, $p < .001$, for correct recognition, $F(2, 160) = 175.50$, $p < .00,1$ for appropriate justification, and $F(2, 160) = 152.33$, $p < .001$, for both correct recognition and appropriate justification. On all three measures, tests of simple effects showed that Grade 1 students had fewer correct responses than the Grade 3 students, who had fewer than the Grade 5 students (see Table 11). No main effect or interaction involving gender was found on any of the three measures. In all grades, students performed better on the recognition component than the justification component. However, Grade 1 students' performance on the justification component dropped dramatically, indeed, almost to floor level. Thus, not surprisingly, students became better at correctly recognizing and at appropriately justifying the use of subtraction solution procedures as they got older. Providing appropriate justifications was especially difficult for the younger students.

Table 11

Procedure Recognition Task Scores (in proportions) as a Function of Grade

	Correct Recognition		Appropriate Justification		Both Correct Recognition **and** Appropriate Justification	
Grade	M	SD	M	SD	M	SD
1	.64	.17	.14	.18	.12	.16
3	.76	.18	.62	.23	.57	.23
5	.87	.14	.77	.18	.73	.20

Reactivity

The procedure recognition task immediately followed completion of the subtraction problem solving task. If verbal reporting does have a reactive effect on performance of the subtraction problem solving task, it might also affect performance on subsequent tasks. The proportion of (a) correct recognition of solution procedures, (b) appropriate justification of procedures, and (c) both correct recognition and appropriate justification of procedures were each subjected to a 3 (Report Condition: NR, RR, CR) x 2 (Gender: male and female) analysis of variance. Analyses for each grade were performed separately.

In the first set of analyses, on the proportion of correct recognition, there were no significant effects for either the Grade 1 or Grade 3 students. However, there was a main effect of Report Condition in Grade 5, $F(2, 59) = 3.89$, $p = .026$ (see Table 12). Students in the CR condition had higher proportions of correct recognitions than the NR students (.93 vs. .80). Students in the RR condition did not differ from either the NR or CR students (.88).

Table 12

Procedure Recognition Task Scores (in proportions) as a Function of Grade and Report Condition

Grade		Correct Recognition		Appropriate Justification		Both Correct Recognition **and** Appropriate Justification	
		M	SD	M	SD	M	SD
1							
	NR	.65	.19	.19	.19	.18	.17
	RR	.63	.16	.09	.15	.08	.14
	CR	.64	.17	.13	.19	.12	.16
	Mean	.64	.17	.14	.18	.12	.16
3							
	NR	.75	.16	.63	.26	.57	.24
	RR	.76	.15	.59	.24	.53	.26
	CR	.76	.22	.63	.18	.61	.19
	Mean	.76	.18	.62	.23	.57	.23
5							
	NR	.80	.15	.75	.15	.68	.17
	RR	.88	.15	.79	.19	.75	.22
	CR	.93	.10	.77	.19	.74	.21
	Mean	.87	.14	.77	.18	.73	.20

In the second set of analyses, on the proportion of appropriate justifications, there were no effects for either the Grade 1 or Grade 5 students. However, in Grade 3, there was an interaction between report condition and gender, $F(2, 57) = 3.48$, $p = .038$, and a main effect of gender, $F(1, 57) = 5.08$, $p = .028$. Overall, males had a higher proportion of appropriate justifications than females (.68 vs. .55), a difference that was evident in the NR condition but not in the other conditions (see Figure 18). No other significant differences were found. Males in the NR condition had better results than males in the RR condition. Females in the NR condition had lower results than females in the RR and CR conditions. From this interaction, no clear pattern is evident involving the potential reactivity of verbal reporting in the previous task. Indeed, for males and females the pattern of performance is almost completely opposite, suggesting that if there is reactivity, it may be affecting males and females differently.

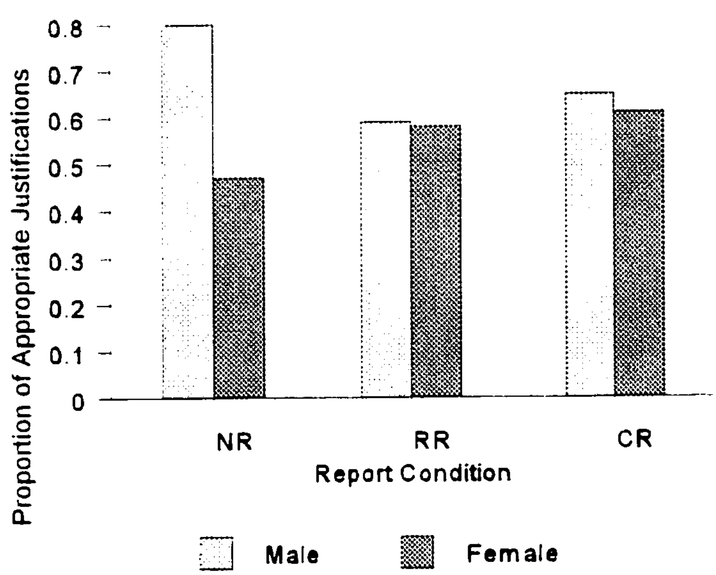

Figure 18. Interaction between Report Condition and Gender in the Recognition

Component of the Procedure Recognition Task.

For the last set of analyses on the proportion of both correct recognition and appropriate justifications, there were no significant effects in any of the grades. Analyses were repeated for all three measures of the task using the four report conditions (NR, RR, CR, and RR in CR). No main effects or interactions were significant although the previously significant effects involving Report Condition in Grade 5 on the recognition component and in Grade 3 on the justification component were close to significance (p = .058 and .080, respectively).

Overall, in Grade 1 there was no evidence that there are residual reactive effects of verbal reporting. In Grades 3 and 5, however, some reactive effects were found. In Grade 3, Report Condition and Gender interacted with males in the NR condition performing better on the justification component than males in the RR condition. Females in the RR and CR conditions, however, performed better than the females in the NR condition. In Grade 5, students in the CR condition performed better on the justification component than students in the NR condition. The reactive effects in Grades 3 and 5 followed no clear pattern but, taken together, indicate that effects of verbal reporting on a previous task may affect performance on later tasks. Therefore, if more than one task is included in a study, especially if it follows the verbal reporting task immediately, the inclusion of tests for longer term reactive effects may be warranted, especially for older students.

Procedure Recognition Task Summary

Analyses of the three measures of performance for the Procedure Recognition Task yielded results consistent with expectations: Performance on both the recognition and justification components improved with age, especially on the latter component. There was moderate evidence of a reactive effect of previous verbal reporting for the older students on the procedure recognition task. The procedure recognition task was not specifically designed as a test of reactivity and therefore results should be interpreted cautiously. However, a test of later reactive effects may be appropriate in studies that include tasks that immediately follow tasks utilizing verbal reports.

Fluency Task Data

Analyses of the fluency task involved three measures: the number of subtraction problems attempted, the number of subtraction problems correct, and the proportion of correct answers divided by the total number of problems attempted. As described in the methods section, the Grade 1 students only attempted the first part of the fluency task, which consisted of 45 single-digit subtraction problems. No Grade 1 students attempted more than 41 of the 45 problems (the average was 14.9). Grade 3 and 5 students were asked to work on both parts of the task and so were asked to solve as many of the additional 90 problems as possible, for a maximum total of 135.

Grade and Gender

The number of problems attempted, problems correct, and the proportion of attempted problems solved correctly were subjected to separate 3 (Grade: 1 through 5) x 2 (Gender: male and female) analysis of variance. There was a main effect of grade on all three measures, $F(2, 155) = 197.17$, $p < .001$, for number of problems attempted, $F(2, 155) = 179.54$, $p < .001$, for number of problems correct, and $F(2, 155) = 5.77$, $p = .004$, for the proportion of attempted problems solved correctly. Obviously, Grade 1 students could not have as many problems attempted or correct as the older students. Indeed, on the first two measures, Grade 1 students had fewer correct responses than the Grade 3 students who had fewer than the Grade 5 students (see Table 13). For the proportion of the number of attempted problems solved correctly, Grade 3 students did not differ from either Grade 1 or 5 students, although Grade 1 students were less accurate than Grade 5 students. Proportions were quite close to ceiling, however, for all grades. No main effect or interaction involving gender was found on any of the three measures.

Table 13

Fluency Task Scores as a Function of Grade

	Problems Attempted		Problems Correct		Proportion Correct	
Grade	M	SD	M	SD	M	SD
1	14.9	7.1	13.6	7.3	.89	.14
3	49.9	15.0	46.5	15.8	.93	.13
5	63.7	16.2	61.3	16.2	.96	.04

Note. The highest possible number of problems attempted was 45 for the Grade 1 students and 135 for the Grade 3 and 5 students.

Reactivity

A final test of the potential reactive effects of verbal reporting was conducted on the fluency task data that were collected in a later session than the subtraction problem solving task. Testing the fluency task data for reactivity is a crude measure of whether or not reactive effects can appear in, or last until, a later time when the students were performing a similar type of task, solving subtraction problems. The number of problems attempted, the number of problems correctly solved, and the proportion of the number of correctly solved problems divided by the number of problems attempted were each subjected to a 3 (Report Condition: NR, RR, and CR) x 2 (Gender: male and female) analysis of variance. Analyses for each grade were performed separately. No effects were found involving Report Condition (see Table 14). Analyses were repeated using the four report conditions (NR, RR, CR, and RR in CR) and again Report Condition did not have any effects. Thus, there appear to be no long lasting effects of providing verbal reports on subsequent performance on subtraction problems. Conversely, this task may not have been suitable for measuring possible long-term reactive effects. The task was not designed for this purpose.

Table 14

Fluency Task Scores as a Function of Grade and Report Condition

Grade	Report	Problems Attempted		Problems Correct		Proportion Correct	
		M	SD	M	SD	M	SD
1							
	NR	14.4	7.1	13.4	7.4	.90	.13
	RR	13.1	5.3	11.21	5.6	.85	.17
	CR	17.2	8.2	16.1	9.2	.03	.01
	Mean	14.9	7.1	13.6	7.3	.89	.14
3							
	NR	48.4	18.1	46.3	18.0	.95	.04
	RR	52.3	14.5	59.5	14.0	.95	.04
	CR	48.9	12.2	43.5	15.5	.90	.21
	Mean	49.9	15.0	46.5	15.8	.93	.13
5							
	NR	62.2	16.8	59.4	16.7	.95	.04
	RR	61.0	15.9	58.2	16.1	.96	.04
	CR	68.3	15.8	66.3	15.4	.97	.03
	Mean	63.7	16.2	61.3	16.2	.96	.04

Note. The highest possible number of problems attempted was 45 for the Grade 1 students and 135 for the Grade 3 and 5 students.

Fluency Task Summary

As with the Procedure Recognition Task, analyses of the three measures of fluency yielded results consistent with expectations. The number of problems students attempted to solve and the number of problems students solved correctly increased with age (remember that Grade 1 students only had the opportunity to solve a maximum of 45 problems versus 135 problems for the Grade 3 and 5 students). A more equivalent measure, then, was the proportion of the number of problems solved correctly divided by the number of problems attempted. In the analysis of this measure, all students performed near ceiling and Grade 1 students only had lower proportions than the Grade 5 students. Grade 3 students' proportions did not differ from the other two grades. No evidence was found for longer term reactive effects of verbal reporting on fluency task performance.

The Development of Subtraction Skills

In this section, both old and new analyses are presented on how accuracy, latency, and strategies change across development. As well, relations are presented between the three forms of mathematical knowledge, procedural, factual, and conceptual as measured by the subtraction problem solving task, the fluency task, and the procedure recognition task, respectively.

Subtraction Solution Strategies

Based on analyses presented in the previous section, as students get older, they get faster and more accurate. This finding corresponds closely to the existing literature. The findings on how the reported use of strategies changes across development, however, adds to the existing literature. Siegler (1989b) found in his study that students in Grades 2 and 4 predominantly used two strategies: counting and retrieval (see also Siegler, 1987b). Grade 2 students, however, relied more heavily on counting while Grade 4 students used retrieval more frequently. While Siegler's results paralleled the findings in this study to the extent that the youngest students, in Grade 1, relied more heavily on counting while the oldest students, in Grade 5, used retrieval more frequently, the inclusion of Grade 3 students yielded interesting results. That is, Grade 3 students in this study relied most heavily on a third strategy type, special tricks. Recall that special tricks included strategies such as deleting tens or using related subtraction facts to help make solving a problem easier. The only difference between special tricks and derived facts was that derived facts are applicable to all problems but special tricks were often usable on only specific problems. As well, derived facts were not reported frequently in this study by students in any grade.

The findings with the Grade 3 students lend support to the hypothesis that the transition from counting to retrieval as the predominant solution strategies is a gradual one. Geary (1994) postulated that students tend to move from counting to decomposition to retrieval. The special tricks reported by the students in this study would correspond to the definition of decomposition strategies. However, Geary referred primarily to decomposition as the use of addition facts to help solve subtraction problems (see also Ilg & Ames, 1951 and Siegler, 1989c). In this study, addition facts were a separate category from special tricks and indeed, students in Grade 3 reported using them most frequently. However, the use of special tricks as the most predominant strategy for children who no longer rely heavily on counting and yet are not ready to use retrieval most of the time is a new finding that needs to be replicated. These special tricks strategies may be one of the methods that students use to facilitate recall of basic fact combinations (Carpenter & Moser, 1984; Putnam, deBettencourt, & Leinhardt, 1990; Sternberg, 1985) as they are moving from counting to retrieval as their preferred strategy or solution procedure.

Though students in each grade had a "preferred" strategy for solving subtraction problems, they rarely relied on one strategy exclusively. Siegler (1989b) has previously demonstrated the diversity of strategies that children in Grades 2 and 4 use on the same set of subtraction problems used in this study. Although Siegler did not note whether there were any grade differences in multiple strategy use between grades, in this study students used multiple strategies most frequently in Grade 3 (see Figure 19). Students in Grade 1 used an average of 3.23 strategies whereas Grade 3 and 5 students reported using an average of 4.39 and 3.58 strategies, respectively. Note that the unknown strategy category was not included so the maximum number of strategies was eight. This increased use of multiple strategies in Grade 3 may be related to the previous finding that Grade 3 students used special tricks as their predominant strategy category. That is, Grade 3 students no longer rely on counting and yet are still not skilled enough to be able to rely primarily on retrieval. Therefore, they are going through a transition that encourages the use of several strategies.

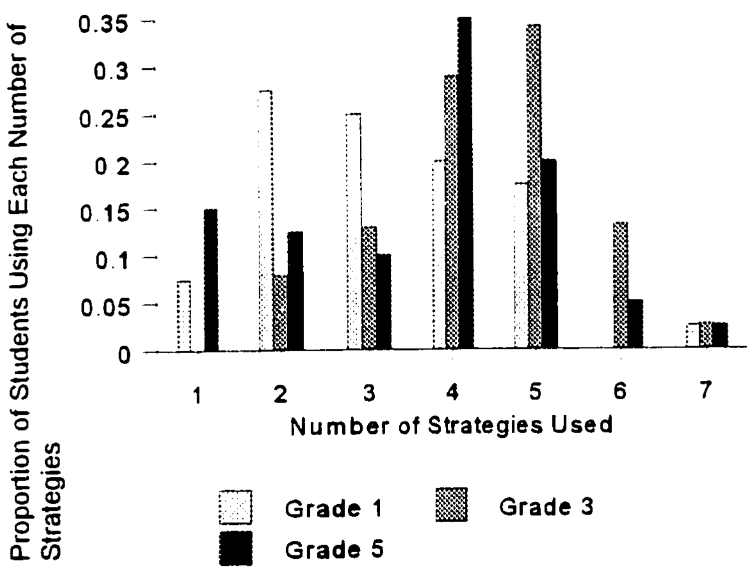

Figure 19. Proportions of students who used each number of strategies.

Overall, the findings regarding the solution strategies that students use to solve subtraction problems indicated that when students of several ages or grades were included in the study, that the picture of how strategies are used and develop was fairly complex. Grade 3 students appeared to be in a transition in their use of strategies. They did not rely on the two most common subtraction strategies of counting and retrieval, and they tended to use more of a variety of strategies. A study using students in Grades 2, 3, and 4 might yield more information about how students move from counting to retrieval.

Types of Mathematical Knowledge

Procedural knowledge of subtraction was assessed in the subtraction problem solving task, conceptual knowledge was tested in the procedure recognition task, and factual knowledge was measured in the fluency task. In this section, the relations between performance on each task, and therefore on each form of mathematical knowledge, are presented.

All three tasks have already been examined separately. In the subtraction problem solving task, younger students were slower, made more errors, and relied more heavily on counting than the older students. In the procedure recognition task, two components of conceptual knowledge were measured: the ability to recognize correct and incorrect strategies as well as the ability to justify whether or not the strategies are appropriate. As grade increased, performance on both these measures improved. However, Grade 1 students had marked difficulty with justifying whether or not the strategies were appropriate. Indeed, their performance was almost at floor. Thus, though recognition is poorer for younger students, justification is almost impossible for them. It could be argued that justification requires a greater ability to verbalize. However, Grade 1 students had little difficulty reporting their solution strategies in the subtraction problem solving task. Therefore, separate components of conceptual knowledge may develop at different rates. On the fluency task, as with the procedure recognition task, performance improved as grade increased. However, the proportion of correct responses was near ceiling for all grades but younger students were much slower and thus attempted fewer problems.

Relations among the three tasks, and thus the three forms of mathematical knowledge, were assessed by correlating measures from each task within each grade. One set of correlations was calculated for all three report conditions and compared to the overall accuracy and median latency on the subtraction problem solving task as well as to the three measures in each of the procedure recognition (correct recognition, appropriate justification, and correct recognition plus appropriate justification) and the fluency tasks (number of problems attempted, number of problems correct, proportion of correctly attempted problems). A second set of correlations was calculated for the RR and CR conditions and therefore included the strategy report measure from the subtraction task. Only significant correlations between measures from different tasks are discussed. See Tables 15 and 16 for a complete presentation of correlations.

In Grade 1, no significant correlations were found between subtraction task measures and procedure recognition measures (see Table 15). Overall accuracy on the subtraction task was positively correlated with all measures in the fluency task. That is, the higher the accuracy on the first task, the more problems that were attempted, the more problems that were solved correctly, and the higher the proportion of problems attempted that were solved correctly on the fluency task. There was also negative correlation between overall latency in the subtraction task and the number of problems attempted in the fluency task, indicating that the slower the students were in the former task, the fewer problems were attempted in the latter task. Therefore, students who were more skilled (i.e., who were faster and had fewer errors) on the subtraction task also tended to have higher scores on at least one of the fluency task measures.

All three measures of the procedure recognition task were positively correlated with the number of problems attempted and the number of problems solved correctly in the fluency task. Thus, students who were quick and accurate in the fluency task also tended to do well in both the recognition and justification components of the procedure recognition task. Thus, in Grade 1, performance on the fluency task was correlated with most of the measures in both the subtraction and procedure recognition tasks.

Grade 3 students had positive correlations between overall accuracy on the subtraction task and performance on the procedure recognition task. The more accurate students were on the first task, the higher their scores on the appropriate justification measure and on the correct recognition plus appropriate justification measure. On the latency measure of the subtraction task, as for the Grade 1 students, latencies were negatively correlated to the number of problems attempted as well as to the number of problems correctly solved. So, the faster the students were on the subtraction task, the higher the number of both problems attempted and problems correctly solved on the fluency task. There were no significant correlations between the procedure recognition and fluency tasks.

In Grade 5, there were no significant correlations between the subtraction and procedure recognition task, nor were there any significant correlations between the procedure recognition and fluency tasks. Accuracy on the subtraction task was positively correlated with the proportion of attempted problems correctly solved on the fluency task and latency was negatively correlated with the number of problems attempted and the number of problems correct on the fluency task.

To summarize, only in Grade 3 was the subtraction task correlated with the procedure recognition task. Students with high proportions of accurate responses in the subtraction task also tended to do well on the justification and recognition plus justification components of the procedure recognition task. In all grades, at least one measure of the subtraction task was correlated with the fluency task. In all grades, students who were faster in the subtraction task tended to attempt more problems on the fluency task, and in Grades 3 and 5, also tended to solve more problems correctly. Although in Grade 3 there were no significant correlations between accuracy on the subtraction task and the fluency task, in Grades 1 and 5 the students who had higher accuracies also tended to have a higher proportion of attempted problems correctly solved in the fluency task. As well, Grade 1 students with higher accuracies also tended to attempt and correctly solve more fluency task problems. Finally, in Grades 3 and 5 there were no significant correlations between the procedure recognition and fluency task measures but in Grade 1 all measures of the procedure recognition task were positively correlated with the number of problems attempted and the number of problems correct in the fluency task. Thus, there were more significant correlations between tasks for the younger students and the number of significant correlations decreased with age. The correlations may not have been as strong for the older students due to fast and accurate performance on all tasks by most of the older students.

Correlations between the subtraction problem solving task and the fluency task were found in all grades. These two tasks were the most similar to each other as they both involved problem solving. These tasks represented factual and procedural knowledge and they are intertwined forms of knowledge that operate and develop in conjunction with each other. Therefore, it is not surprising that performance on the tasks were related. The procedure recognition task represented a measure of conceptual knowledge, which is thought to be more closely linked to procedural knowledge than factual knowledge. Only in Grade 3 was this relation found between the two tasks. Students who were able to justify and to justify correctly recognized solution procedures tended to make fewer errors on the subtraction task. This finding, taken in conjunction with the finding that students in Grade 3 may be in a transitional state with their heavier reliance on special tricks to solve subtraction problems, suggests that further investigation of the changes and development of all types of mathematical knowledge in this grade is needed. Finally, the procedure recognition task and fluency task were only correlated in Grade 1. This finding is surprising, especially given that no correlations were found between the subtraction task and the procedure recognition task, and also indicates that further investigation of the relations between the different forms of mathematical knowledge and how they develop is warranted.

Table 15

Correlations Between Accuracy and Latency Measures on the Subtraction Task and all Measures of the Procedure Recognition and Fluency Tasks for Students in Grades 1, 3, and 5 (All Report Conditions)

Task	Measure	1	2	3	4	5	6	7	8
				Grade 1 (n = 60)					
Subtraction									
	1. Accuracy	---	-.24	.20	.14	.14	.48**	.55**	.44**
	2. Latency		---	-.11	-.15	-.14	-.31*	-.26	.10
Procedure Recognition									
	3. Recognition			---	.22	.27*	.46**	.44**	.15
	4. Justification				---	.97**	.44**	.46**	.16
	5. Recognition + Justification					---	.44**	.42**	.17
Fluency									
	6. Problems Attempted						---	.97**	.26*
	7. Problems Correct							---	.47**
	8. Problems Correct/ Problems Attempted								---
				Grade 3 (n = 58)					
Subtraction									
	1. Accuracy	---	.02	.08	.28*	.27*	.13	.16	.10
	2. Latency		---	.22	.01	.07	-.50**	-.49**	-.08

table continues

Task	Measure	1	2	3	4	5	6	7	8
			Grade 3						
Procedure Recognition									
	3. Recognition			---	33*	47**	-.04	-.09	-.14
	4. Justification				---	.91**	.13	.11	.01
	5. Recognition + Justification					---	.07	.03	-.03
Fluency									
	6. Problems Attempted						---	.89**	.01
	7. Problems Correct							---	.01
	8. Problems Correct/ Problems Attempted								---
			Grade 5 (n = 60)						
Subtraction									
	1. Accuracy	---	-.37**	.01	-.09	.04	.20	.24	.27*
	2. Latency		---	.12	-.04	-.02	-.43**	-.42**	-.03
Procedure Recognition									
	3. Recognition			---	.48**	.69**	-.16	-.13	.25
	4. Justification				---	.86**	-.07	-.06	.08
	5. Recognition + Justification					---	-.13	-.11	.13
Fluency									
	6. Problems Attempted						---	.99**	.20
	7. Problems Correct							---	.33*
	8. Problems Correct/ Problems Attempted								---

* $p < .05$

** $p < .01$

The second set of correlations was conducted on the RR and CR condition groups only. All correlations were calculated on all three measures of each of the three tasks, which included reported strategy categories from the subtraction task. Because there were nine strategy categories, each category was correlated with all of the other measures and these correlations are reported in Table 16.

Generally, there were few additional significant correlations between the strategy categories and the procedure recognition task and fluency task measures. Some of these correlations must be interpreted extremely cautiously due to very low frequencies for some of the reported strategies (e.g., guessing, derived fact, and the other category were each reported on less than 5 percent of problems in all grades). Though correlations involving these low-frequency strategies are reported in Table 16, they will not be discussed.

In Grade 1 there were positive correlations between the frequency of reported retrieval use on the subtraction task and measures from the other two tasks. Students who tended to retrieve more frequently also tended to have higher scores on the justification component of the procedure recognition task as well as having higher numbers of problems attempted and problems correct on the fluency task. The use of retrieval being related to performance on the fluency task is not unexpected (see LeFevre, Bisanz, & Sadesky, 1996), but the relation between retrieval use and being able to justify diverse solution procedures on the procedure recognition task is intriguing and suggests that retrievers are not only faster and more accurate (as evidenced by the correlation with the fluency task measures) but may also have a better conceptual understanding of subtraction.

In Grade 3, counting was negatively correlated to the other two tasks. The more counting students reported using, the lower their scores tended to be on the justification component and the correct recognition plus appropriate justification measure of the procedure recognition task. Also, more reported counting use was associated with fewer problems attempted in the fluency task. These correlations involving reported counting are not incompatible to the correlations with reported retrieval found for the Grade 1 students. Students who more frequently reported counting (rather than more sophisticated procedures such as retrieval or special tricks) were less likely to be able to justify why a solution procedure would be appropriate or not and also to correctly recognize and appropriately justify solution procedures. As well, students who used counting more frequently were not as fast as they attempted fewer problems on the fluency task.

In Grade 5, solution procedures that were categorized as suspected retrieval had a positive correlation with both the number of problems attempted and the number of problems correct on the fluency task. Suspected retrieval was an artificial category used by the experimenter to categorize otherwise unknown solution procedures that were fast and accurate. Thus, it is not surprising that a category that was based on fast and accurate performance is correlated to fast and accurate performance on another task, but it does provide some more support for the validity of the suspected retrieval category.

Thus, in Grade 1, students who retrieved tended to perform better on at least some components of the procedure recognition and fluency tasks. Grade 3 students who counted tended to perform more poorly on at last some of the components of the procedure recognition and fluency tasks. In Grade 5, no theoretically interesting correlations were found between strategies and measures of the procedure recognition and fluency tasks. The findings in Grades 1 and 3, however, suggest that young students using more sophisticated solution strategies also have better conceptual and factual knowledge and that the Grade 3 students who are least sophisticated in their solution strategies also have the weakest conceptual and factual knowledge. Thus, it may be that students, at least in Grades 1 and 3, who do well on one task will do well on all tasks while conversely, students who do not perform well on one task will not perform well on all tasks.

Overall, the general finding in all grades was that students who perform well on the subtraction task tend to also do well on the fluency task, which is not surprising given the close relationship between the two forms of knowledge measured by these tasks. In Grades 1 and 3, some correlations were found between the subtraction task and the procedure recognition task, indicating that these two forms of knowledge may develop in conjunction with each other. Correlations between the procedure recognition and fluency tasks were only found in Grade 1 and, taken with the other correlations, could mean that all three forms of knowledge are developing simultaneously when students are first learning about subtraction. Finally, the small number of significant correlations in Grade 5 may at least partially be due to ceiling effects on several measures of each task.

Table 16

Correlations Between Reported Strategies from the Subtraction Task and all Measures of the Procedure Recognition and Fluency Tasks for Students in the RR and CR Conditions in Grades 1, 3, and 5

Task	Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
								Grade 1 (n = 40)								
Subtraction																
	1. Counting	---	-.26	-.38*	-.15	-.43**	-.22	-.47**	-.14	.01	-.09	-.15	-.20	-.31	-.26	.14
	2. Retrieval		---	-.08	-.15	-.22	-.11	-.21	-.05	-.17	.14	.35*	.29	.60**	.55**	-.04
	3. Suspected Retrieval			---	-.06	-.15	-.03	-.04	-.03	-.11	.10	-.20	-.20	.11	.10	-.01
	4. Derived Fact				---	.42**	.15	-.05	-.10	-.01	.20	.43**	.51**	.16	.16	.07
	5. Special Trick					---	.34*	.09	-.15	.04	.21	.22	.30	.07	.11	.17
	6. Addition Fact						---	-.08	.15	.08	.07	.16	.21	.02	.05	.11
	7. Unknown							---	-.09	-.11	-.29	-.09	-.04	-.22	-.22	-.15
	8. Guessing								---	.09	-.06	-.23	-.22	-.07	-.18	-.41**
	9. Other									---	.04	-.07	-.06	-.25	-.23	.08
Procedure Recognition																
	10. Recognition										---	.20	.24	.44**	.42**	.10
	11. Justification											---	.97**	.51**	.53**	.22
	12. Recognition + Justification												---	.47**	.50**	.23

table continues

Task	Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
								Grade 1								
Fluency																
	13. Problems Attempted													---	.97**	.19
	14. Problems Correct														---	.42**
	15. Problems Correct/Problems Attempted															
								Grade 3 (n = 38)								
Subtraction																
	1. Counting	---	-.02	.20	-.32*	-.25	-.39*	-.28	-.33*	-.10	-.27	-.53**	-.48**	-.43**	-.27	.18
	2. Retrieval		---	-.11	-.29	-.43**	-.30	-.17	.08	-.23	.03	.07	.04	.04	.01	-.05
	3. Suspected Retrieval			---	-.31	-.32*	-.13	-.13	.07	-.02	.05	-.09	-.07	.00	-.13	-.22
	4. Derived Fact				---	.47**	.23	-.16	-.15	.08	-.16	-.08	-.12	.01	.09	.13
	5. Special Trick					---	-.10	-.28	-.24	.11	.11	.19	.12	.17	.20	.08
	6. Addition Fact						---	-.08	-.12	.04	.04	.07	.11	.09	.05	-.05
	7. Unknown							---	-.08	.02	.04	.20	.22	.04	.00	-.06
	8. Guessing								---	-.02	-.23	-.51**	-.43**	-.10	-.03	.11
	9. Other									---	.10	.09	.09	.09	.07	.00
Procedure Recognition																
	10. Recognition										---	.45**	.52**	.14	.04	-.13
	11. Justification											---	.94**	.20	.13	-.09
	12. Recognition + Justification												---	.19	.10	-.10

table continues

Task	Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
								Grade 3								
Fluency																
	13. Problems Attempted													---	.81**	-.08
	14. Problems Correct														---	.52**
	15. Problems Correct/Problems Attempted															---
								Grade 5 (n = 40)								
Subtraction																
	1. Counting	---	-.24	-.29	-.30	-.24	-.17	-.03	---	-.16	-.15	.04	-.01	-.28	-.28	-.12
	2. Retrieval		---	-.40*	.11	-.23	-.26	-.25	---	-.07	.07	.01	.04	.14	.15	.06
	3. Suspected Retrieval			---	-.17	-.30	-.17	-.02	---	.07	-.11	-.17	-.16	.41**	.41**	.11
	4. Derived Fact				---	.31	-.10	-.15	---	.08	.17	.01	.03	-.09	-.10	-.11
	5. Special Trick					---	-.13	.38*	---	.20	.05	-.04	.01	-.29	-.31	-.17
	6. Addition Fact						---	-.10	---	-.09	.15	.20	.15	-.09	-.08	.11
	7. Unknown							---	---	.01	.09	.10	.10	-.17	-.16	.07
	8. Guessing[a]								---	---	---	---	---	---	---	---
	9. Other									---	-.04	.09	.12	.08	.07	-.07
Procedure Recognition																
	10. Recognition										---	.56**	.69**	.01	.03	.25
	11. Justification											---	.93**	-.08	-.06	.20
	12. Recognition + Justification												---	-.13	-.11	.17

table continues

Task	Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
								Grade 5								
Fluency																
	13. Problems Attempted													---	.99**	.28
	14. Problems Correct														---	.38*
	15. Problems Correct/Problems Attempted															---

* $p < .05$

** $p < .01$

[a] no guessing was reported in Grade 5

GENERAL DISCUSSION

Verbal reports are becoming increasingly common tools for assessing children's strategies in the area of cognitive development. The trend in recent years has been to use verbal reports rather than relying exclusively on accuracy and latency data to infer the strategies that children are using. However, the use of verbal reports as data in many areas of psychology has been controversial and little research has been conducted on directly assessing the validity of verbal reports as an appropriate measure for the study of cognitive development. This study was designed to examine (a) the validity of verbal reports of children's solution strategies on subtraction problems and (b) the development of subtraction skills and strategies.

The Validity of Verbal Reports

Verbal reports were indeed found to be valid forms of data in this study. Careful study design and judicious use of verbal reports is, however, integral to the successful use of verbal reports. Tests of the validity of verbal reports are recommended, and concurrent and retrospective reports may be more appropriate on some tasks, and for children of certain ages. The advantages of verbal reports, however, supercede the inconvenience of integrating validity tests into the study design. As was found in this study, verbal reports revealed rich information regarding the development of children's subtraction skills and strategies that would otherwise not have been found. Thus, the significance of verbal reports as an important source of data must be underscored, so long as conditions of verbal report use are carefully considered.

Two forms of verbal report validity were examined in this study: reactivity and veridicality. As well, two more specific validity issues were also of interest in this study: First, whether the type of report, retrospective or concurrent, affected overall validity and second, whether verbal reports are appropriate forms of data when studying children.

Reactivity

On the whole, providing a verbal report had little effect on students' performance but three exceptions were found, affecting only Grade 3 and 5 students. First, Grade 3 students in the CR condition were slower than students in the other two report conditions. As well, when analyses were re-done by separating students in the CR condition into two subgroups, the "pure" CR group and the RR in CR group, the Grade 5 CR group had slower latencies. This slowing of performance, however, was expected a priori and was therefore not of major concern to the reactivity issue. Ericsson and Simon (1980, 1984, 1993) described several studies in which concurrent reporting slowed performance because of the instruction to verbalize but did not otherwise alter task performance.

A second finding that could imply reactivity was that in both Grades 3 and 5, strategies were not always reported with the same frequency in both report conditions. This finding, however, could be due to either a reactive effect of the instruction to provide a verbal report or, alternatively, could be due to non-veridical reports in either report condition. The comparison of the data yielded by two different types of verbal reports as a test of reactivity has never been done before. Thus, even if the normal tests of reactivity with accuracy and latency data yield non-reactive results, these tests may not be sufficient evidence of the non-reactive effects of verbal reports. A closer examination of frequencies of strategy reports might reveal reactive effects not discriminated by accuracy and latency data. Therefore, based on these discrepancies, tests of reactivity may still be necessary when studying similar tasks with verbal reports.

The final finding of reactive effects of verbal reporting was found in the procedure recognition task which was administered immediately following completion of the subtraction task. No reactive effects were found in the fluency task that was administered in a later session. For the procedure recognition task, in Grades 3 and 5, some of the report conditions performed significantly differently from each other. No clear patterns of these reactive effects emerged. However, based on these findings, the use of tests of longer-term reactivity are also recommended in studies that administer further tasks subsequent to the verbal reporting task.

Veridicality

Students in all grades were able to veridically report their solution strategies, either through the use of concurrent or retrospective reports. The analyses and examinations of the data associated with reported counting, retrieval, and special tricks were consistently supportive of the veridicality of verbal reports.

Students in all grades and report conditions counted most frequently on the difficult problems and retrieved most frequently on the easiest problems, as expected. More specifically, with few exceptions, counting occurred on the problems with the most errors and slowest solution times whereas retrieval occurred on the problems with the fewest errors and fastest solution times. In Grade 5, students were performing very close to ceiling and therefore significant differences may have been obscured. Further, previous research suggests that young students rely most heavily on counting whereas older students rely more on retrieval to solve subtraction problems. This pattern of results was found along with the intriguing new finding that Grade 3 students primarily use a different category of strategies altogether, that of special tricks. Overall, few exceptions were found for any reported strategy category, for any grade, or for any report condition.

Therefore, the previous finding that strategy frequencies differed in the concurrent and retrospective report conditions in Grades 3 and 5 is more likely to be due to reactive effects of verbal reporting than to students providing non-veridical reports of their strategy use. The question remains though of which, or whether perhaps both, types of verbal reports produce reactive effects on task performance for older students.

Verbal Report Validity and Type of Verbal Report

The validity of concurrent and retrospective verbal reports was assessed by using several measures. For the most part, accuracy and latency patterns were not altered by the task of providing a retrospective or concurrent verbal report. Frequencies of reported strategies, however, were more problematic for the reactivity issue. In Grades 3 and 5 some strategy categories were reported more often in either the retrospective or concurrent condition. Because evidence supporting the veridicality of students' strategy reports in both report conditions was found, the discrepancies in reported strategy frequencies were considered to be due to reactive effects of one, or both, of the report conditions.

Informal findings in this study suggest that concurrent reports are more suspect than retrospective reports. Many of the students, especially the older students, in the concurrent report condition had difficulty with following report instructions. Indeed, several students spontaneously told the experimenters that being asked to concurrently report interfered with solving the problem. Despite careful instructions and practice as well as reminders throughout the task, many students would repeatedly give the answer and *then* give their report of the solution strategy they had used. In essence, many of the students in the concurrent report condition spontaneously switched to retrospective reports suggesting that retrospective reporting on this task was much more natural or comfortable for the older students.

Further supporting the hypothesis that concurrent reporting was more difficult and more likely to change task performance, students in the concurrent report condition had much higher incidences of being unable to provide any report of their solution strategy. Therefore, based on the findings of this study, retrospective reports are recommended on similar tasks for students of all ages.

Concurrent reports, however, would also be appropriate for younger students especially if time or fatigue are factors of concern. No reactive effects were found for either report condition on any of the measures for Grade 1 students. Students in Grade 1 often spontaneously concurrently reported, or thought aloud, in all report conditions. Indeed, students in the concurrent report condition were probably not so much following the experimenter's instruction to concurrently report as they were solving the subtraction problems in their normal or natural fashion. As well, performance by the Grade 1 students was already slow enough that the concurrent reporting did not lengthen solution latencies any further. Finally, in the retrospective report condition, session lengths were particularly long as students often ended up giving verbal reports of their solution strategies twice, once concurrently and once retrospectively. Thus, concurrent reports are also recommended for younger students, especially on tasks that are lengthy or fatigue-inducing. On tasks that are more complex, concurrent reports might also be appropriate for older students although this suggestion would require further study.

Verbal Report Validity and Children

Previous studies have yielded conflicting evidence on the suitability of verbal reports as a measure with children. Results from this study provide more support for the appropriateness of verbal reports as a tool for studies involving children. Students in all grades were able to provide both retrospective and concurrent reports of their solution strategies on subtraction problems. However, as discussed in the previous section, though both types of reports yielded valid data, older students were more comfortable with retrospective reporting than concurrent reporting. In the concurrent report condition the older students had increased difficulties following instructions and had a higher incidence of problems for which they were unable to provide any strategy report of their solution process. This could be because the older students were quite skilled at subtraction and their quick and accurate performance was impeded by the requirement to concurrent report. Younger students performed well in both report conditions. Thus, verbal reports were valid sources of data for students in all grades, but concurrent reports may be more appropriate for younger children or on tasks that are more difficult.

The Development of Subtraction Skills

Not only were verbal reports found to be valid forms of data appropriate for use with children of all ages, verbal reports also yielded insight into the development of subtraction skills. Without verbal reports, insights into the complexity of how subtraction strategies change and develop across grade would not have been possible. As well, the use of verbal reports led to more detailed information about the development of the relations amongst the three types of mathematical knowledge. Thus, verbal reports helped fulfill the second goal of this study.

Subtraction Solution Strategies

In previous studies, younger children have been found to rely most heavily on counting while older students predominantly use retrieval to solve subtraction problems. In this study, evidence for the use of intermediate strategies was found. That is, Grade 3 students primarily used a third type of strategy, special tricks, to solve subtraction problems. These strategies present a middle ground between counting and retrieval as they are more sophisticated than counting because of their partial reliance on known information about subtraction and yet they are not as fast and automatic as retrieval.

Unlike Grade 1 students, the Grade 3 students were no longer dependent on counting as their primary backup strategy. However, because some of the problems were still too difficult for them to simply retrieve the answer, students may have implemented, due to their improving conceptual knowledge of subtraction, faster and easier backup strategies than counting that could be successfully employed to solve the problems. As experience with subtraction increased, however, students' factual knowledge increased sufficiently that the use of backup strategies such as counting or special tricks were usually no longer necessary. That is, Grade 5 students were able to rely primarily on retrieval. Therefore, the usage of special tricks by the Grade 3 student supports the hypothesis that these students are moving through an intermediate stage in the development of mathematical knowledge.

Further corroborating evidence for the hypothesis that Grade 3 students are in a transitional state in terms of their subtraction strategy development was that students in Grade 3 had the highest incidence of multiple strategy use. That is, students in Grade 3 tended to use more diverse strategies to help them solve the subtraction problems. Grade 3 students had more strategies at their disposal than the Grade 1 students and yet were not skilled enough to use retrieval as their primary solution procedure. A microgenetic study of Grade 3 students' subtraction strategies and how they change would provide the opportunity to examine this period of change more closely.

Types of Mathematical Knowledge

Three forms of mathematical knowledge were also measured in this study. A further finding in this study regarding the development of subtraction skills was that as development progresses, students' arithmetical knowledge steadily improves. The three forms of mathematical knowledge, procedural, factual, and conceptual were measured with the subtraction task, the fluency task, and the procedure recognition task, respectively.

Correlations between the subtraction task and fluency task were found in all grades. These tasks measured procedural and factual knowledge, respectively, which operate and develop together. Correlations between the procedure recognition task and the other two tasks were few, suggesting that conceptual knowledge is not as closely linked to procedural and factual knowledge. Only in Grade 1 were there correlations between conceptual knowledge and procedural knowledge, and only in Grade 3 were there correlations between conceptual knowledge and factual knowledge suggesting that the relations amongst types of mathematical knowledge may shift and change across development.

Correlations between strategies reported on the subtraction task and measures of the procedural recognition and fluency tasks were also found. Grade 1 students who reported using retrieval most frequently within each grade performed better on the fluency task which matches previous findings. Also, Grade 3 students who reported using the least sophisticated strategy, counting, most frequently performed more poorly on the fluency task. An intriguing positive correlation was also found between retrieval use in Grade 1 and counting use in Grade 3 to performance on the measure of conceptual knowledge. This finding suggests that Grade 1 retrievers are not only faster and more accurate than non-retrievers but also have a better conceptual understanding of subtraction in general. Conversely, Grade 3 counters were not only slower and less accurate, but also had poorer conceptual understanding of subtraction than their counterparts who used other strategies such as retrieval and special tricks. Few significant differences were found in Grade 5, probably due to performance at ceiling on most measures of each of the tasks. However, the Grade 1 and 3 findings suggest that mathematical knowledge should not only be investigated across development but that individual differences in conceptual knowledge would also add important insights into the development of arithmetic.

The results of this study compellingly indicate that verbal reports are not only useful components of the study of cognitive development in children's arithmetic, but that verbal reports provide rich and insightful data on such tasks under investigation. Suitable care must be taken to effectively monitor conditions for which verbal reports yield valid data, but the optimistic theoretical positions on verbal report validity taken by numerous researchers over the last decade or so is certainly warranted as is indicated by this study. Finally, in this study, through the use of verbal reports and measures on other tasks. a more complex picture emerged of how both strategies and relations amongst types of mathematical knowledge change and develop across time.

References

Anders, T. R. (1971). Retrospective reports of retrieval from short-term memory. Journal of Experimental Psychology, 90, 251-257.

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

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. Developmental Review, 2, 213-236.

Austin, J., & Delaney, P. F. (1998). Protocol analysis as a tool for behavior analysis. The Analysis of Verbal Behavior, 15, 41-56.

Bellezza, F. S. (1986)). Mental cues and verbal reports in learning. The Psychology of Learning and Motivation, 20, 237-273.

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

Biggs, S. F., Rosman, A. J., & Sergenian, G. K. (1993). Methodological issues in judgment and decision-making research: Concurrent verbal protocol validity and simultaneous traces of process. Journal of Behavioral Decision Making, 6, 187-206.

Bisanz, J., & LeFevre, J.-A. (1990). Strategic and non-strategic processing in he development of mathematical cognition. In D. F. Bjorklund (Ed.), Children's strategies: Contemporary views of cognitive development (pp. 213-244). Hillsdale, NJ: Erlbaum.

Boulton-Lewis, G. M. (1993). Young children's representations and strategies for subtraction. British Journal of Educational Psychology, 63, 441-456.

Bowers, V. A., & Snyder, H. L. (1990). Concurrent versus retrospective verbal protocols for comparing window usability. Proceedings of the Human Factors Society - 34th annual meeting (Vol. 2, pp. 1270-1274). Santa Monica, CA: The Human Factors Society.

Brewin, C. R., Andrews, B., & Gotlib, I. H. (1993). Psychopathology and early experience: A reappraisal of retrospective reports. Psychological Bulletin, 113, 82-98.

Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 1539-1553.

Carpenter, T. P., & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. Journal for Research in Mathematics Education, 15, 179-202.

Carr, M., & Jessup, D. L. (1997). Gender differences in first-grade mathematics strategy use: Social and metacognitive influences. Journal of Educational Psychology, 89, 318-328.

Christianson, S.-A. (1992). Emotional stress and eyewitness memory: A critical review. Psychological Bulletin, 112, 284-309.

Cooney, J. B., & Ladd, S. F. (1992). The influence of verbal protocol methods on children's mental computation. Learning and Individual Differences, 4, 237-257.

Cooney, J. B., & Swanson, H. L. (1990). Individual differences in memory for mathematical story problems: Memory span and problem perception. Journal of Educational Psychology, 82, 570-577.

Cooney, J. B., Swanson, H. L., & Ladd, S. F. (1988). Acquisition of mental multiplication skill: Evidence for the transition between counting and retrieval strategies. Cognition and Instruction, 5, 323-345.

Critchfield, T. S. (1993). Signal-detection properties of verbal self-reports. Journal of the Experimental Analysis of Behavior, 60, 495-514.

Critchfield, T. S. (1996). Differential latency and selective nondisclosure in verbal self-reports. The Analysis of Verbal Behavior, 13, 49-63.

Critchfield, T. S., & Epting, L. K. (1998). The trouble with babies and the value of bathwater: Complexities in the use of verbal reports as data. The Analysis of Verbal Behavior, 15, 65-74.

Crutcher, R. J. (1994). Telling what we know: The use of verbal report methodologies in psychological research. Psychological Science, 5, 241-244.

Denning, S., Hoiem, D., Simpson, M., & Sullivan, K. (1990). The value of thinking-aloud protocols in industry: A case study of Microsoft. Proceedings of the Human Factors Society - 34th annual meeting (Vol. 2, pp. 1285-1289). Santa Monica, CA: The Human Factors Society.

Ericsson, K. A. (1987). Theoretical implications from protocol analysis on testing and measurement. In R. R. Ronning, J. A. Glover, J. C. Conoley, & J. C. Witt (Eds.), The influence of cognitive psychology in testing (pp. 191-226). Hillsdale, NJ: Erlbaum.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. Psychological Review, 87, 215-251.

Ericsson, K. A., & Simon, H. A. (1984). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.

Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data (rev. ed.). Cambridge, MA: MIT Press.

Fidler, E. J. (1983). The reliability and validity of concurrent, retrospective, and interpretive verbal reports: An experimental study. In P. Humphreys, O. Svenson, & A. Vari (Eds.), Analysing and aiding decision processes (pp. 429-440). Amsterdam: Elsevier.

Flaherty, E. G. (1974). The thinking aloud technique and problem solving ability. Journal of Educational Research, 68, 223-225.

Fuson, K. C. (1984). More complexities in subtraction. Journal for Research in Mathematics Education, 15, 214-225.

Geary, D. C. (1994). Children's mathematical development: Research and practical applications. Washington, DC: American Psychological Association.

Ginsburg, H. P., Kossan, N. E., Schwartz, R., & Swanson, D. (1983). Protocol methods in research on mathematical thinking. In H. P. Ginsburg (Ed.), The development of mathematical thinking (pp. 8-49). New York: Academic Press.

Green, A. J. (1995). Verbal protocol analysis. The Psychologist, 8, 126-129.

Groen, G. J., & Parkman, J. M. (1972). A chronometric analysis of simple addition. Psychological Review, 79, 329-343.

Hamann, M. S., & Ashcraft, M. H. (1985). Simple and complex mental addition across development. Journal of Experimental Child Psychology, 40, 49-72.

Harel, G., Behr, M., Post, T., & Lesh, R. (1992). The blocks task: Comparative analyses of the task with other proportion tasks and qualitative reasoning skills of seventh-grade children in solving the task. Cognition and Instruction, 9, 45-96.

Hayes, S. C., White, D., & Bissett, R. T. (1998). Protocol analysis and the "silent dog" method of analyzing the impact of self-generated rules. Analysis of Verbal Behavior, 15, 57-63.

Houlihan, D. M., & Ginsburg, H. P. (1981). The addition methods of first- and second-grade children. Journal for Research in Mathematics Education, 12, 95-106.

Ilg, F., & Ames, L. B. (1951). Developmental trends in arithmetic. The Journal of Genetic Psychology, 74, 3-28.

Kerkman, D. D., & Siegler, R. S. (1997). Measuring individual differences in children's addition strategy choices. Learning and Individual Differences, 9, 1-18.

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. Cognitive Psychology, 25, 111-146.

Kliman, M. (1987). Children's learning about the balance scale. Instructional Science, 15, 307-340.

Laszlo, J., Meutsch, D., & Viehoff, R. (1988). Verbal reports as data in text comprehension research: An introduction. Text, 8, 283-294.

Lawson, M. J., & Rice, D. N. (1987). Thinking aloud: Analyzing students' mathematics performance. School Psychology International, 8, 233-243.

Leder, G. C. (1992). Mathematics and gender: Changing perspectives. In D. A. Grouws (Ed.), Handbook of research on mathematics teaching and learning (pp. 597-622). New York: Macmillan.

Leiken, R., & Zaslavsky, O. (1997). Facilitating student interactions in mathematics in a cooperative learning setting. Journal for Research in Mathematics

Education, 28, 331-354.

LeFevre, J.-A., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problem-size effect in adults. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 216-230.

Lemaire, P., Barrett, S. E., Fayol, M., & Abdi, H. (1994). Automatic activation of addition and multiplication facts in elementary school children. Journal of Experimental Child Psychology, 57, 224-258.

Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: Contributions to children's learning of multiplication. Journal of Experimental Psychology: General, 24, 83-97.

McGilly, K., & Siegler, R. S. (1989). How children choose among serial recall strategies. Child Development, 60, 172-182.

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

Norris, S. P. (1990). Effect of eliciting verbal reports on thinking on critical thinking test performance. Journal of Educational Measurement, 27, 41-58.

Norris, S. P. (1991). Informal reasoning assessment: Using verbal reports of thinking to improve multiple-choice test validity. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), Informal reasoning and education (pp. 451-472). Hillsdale, NJ: Erlbaum.

Payne, J. W. (1994). Thinking aloud: Insights into information processing. Psychological Science, 5, 241-248.

Poole, D. A., & White, L. T. (1991). Effects of question repetition on the eyewitness testimony of children and adults. Developmental Psychology, 27, 975-986.

Poole, D. A., & White, L. T. (1993). Two years later: Effects of question repetition and retention interval on the eyewitness testimony of children and adults. Developmental Psychology, 29, 844-853.

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

Putnam, R. T., deBettencourt, L. U., & Leinhardt, G. (1990). Understanding of derived-fact strategies in addition and subtraction. Cognition and Instruction, 7, 245-285.

Robertson, W. C. (1990). Detection of cognitive structure with protocol data: Predicting performance on physics transfer problems. Cognitive Science, 14, 253-280.

Robinson, K. M. (1993). The anatomy of a shortcut. Unpublished master's thesis. University of Alberta, Edmonton.

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

Secada, W. G., Fuson, K. C., & Hall, J. W. (1983). The transition from counting-all to counting-on in addition. Journal for Research in Mathematics Education, 14, 47-57.

Short, E. J., Cuddy, C. L., Friebert, S. E., & Schatschneider, C. W. (1990). The diagnostic and educational utility of thinking aloud during problem solving. In H. L. Swanson & B. Keogh (Eds.), Learning disabilities: Theoretical and research issues (pp. 93-109). Hillsdale, NJ: Erlbaum.

Short, E. J., Schatschneider, C., Cuddy, C. L., Evans, S. W., Dellick, D. M., & Basili, L. A. (1991). The effect of thinking aloud on the problem solving performance of bright, average, learning disabled, and developmentally handicapped students. Contemporary Educational Psychology, 16, 139-153.

Siegler, R. S. (1987a). The perils of averaging data over strategies: An example from children's addition. Journal of Experimental Psychology: General, 116, 250-264.

Siegler, R. S. (1987b). Strategy choices in subtraction. In J. A. Sloboda & D. Rogers (Eds.), Cognitive processes in mathematics (pp. 81-106). Oxford: Oxford University Press.

Siegler, R. S. (1988). Individual differences in strategy choices: Good students, not-so-good students, and perfectionists. Child Development, 59, 833-851.

Siegler, R. S. (1989a). Strategy diversity and cognitive assessment. Educational Researcher, 18, 15-20.

Siegler, R. S. (1989b). Hazards of mental chronometry: An example from children's subtraction. Journal of Educational Psychology, 81, 497-506.

Siegler, R. S. (1989c). Mechanisms of cognitive development. Annual Review of Psychology, 40, 353-379.

Siegler, R. S. (1996). Emerging minds: The process of change in children's thinking. New York: Oxford University Press.

Siegler, R. S. (1998). Children's thinking (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. American Psychologist, 46, 606-620.

Siegler, R. S., & Jenkins, E. (1989). How children discover new strategies.

Hillsdale, NJ: Erlbaum.

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In G. Halford & T. Simon (Eds.), Developing cognitive competence: New approaches to process modeling. Hillsdale, NJ: Erlbaum.

Siegler, R. S., & Shrager, J. (1984). Strategy choices in addition and subtraction: How do children know what to do? In C. Sophian (Ed.), Origins of cognitive skills (pp. 229-293). Hillsdale, NJ: Erlbaum.

Smith, E. R., & Miller, F. D. (1978). Limits on perception of cognitive processes: A reply to Nisbett and Wilson. Psychological Review, 85, 355-362.

Steffler, D. J., Varnhagen, C. K., Friesen, C. K., & Treiman, R. (1998). There's more to children's spelling than the errors they make: Strategic and automatic processes for one-syllable words. Journal of Educational Psychology, 90, 492-505.

Stinessen, L. (1985). The influence of verbalization on problem solving. Scandinavian Journal of Psychology, 26, 342-347.

Sternberg, R. (1985). Instruction on derived facts strategies in addition and subtraction. Journal for Research in Mathematics Education, 16, 337-355.

Svenson, O. (1975). Analysis of time required by children for simple additions. Acta Psychologica, 39, 289-302.

Svenson, O. (1989). Eliciting and analysing verbal protocols in process studies of judgment and decision making. In H. Montgomery, & O. Svenson (Eds.), Process and structure in human decision making (pp. 65-81). Chichester: Wiley & Sons.

Svenson, O, Hedenborg, M.-L. (1979). Strategies used by children when solving simple subtractions. Acta Psychologica, 59, 477-489.

Svenson, O., Hedenborg, M.-L., & Lingman, L. (1976). On children's heuristics for solving simple additions. Scandinavian Journal of Psychology, 20, 161-173.

Toward, J. L., & Kerr, R. (1994). Exploring relations among perception, action, and knowledge using retrospective verbal reports. Perceptual and Motor Skills, 78, 511-514.

Williams, D. M., & Santos-Williams, S. (1980). Method for exploring retrieval processes using verbal protocols. In R. S. Nickerson (Ed.), Attention and performance (Vol. 8, pp. 671-689). Hillsdale, NJ: Erlbaum.

Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. Psychological Science, 5, 249-252.

Woods, S. S., Resnick, L. B., & Groen, G. J. (1975). Experimental test of five process models for subtraction. Journal of Educational Psychology, 67, 17-21.

Wulfert, E., Dougher, M. J., & Greenway, D. E. (1991). Protocol analysis of the correspondence of verbal behavior and equivalence class formation. Journal of the Experimental Analysis of Behavior, 56, 489-504.

Appendix A

Subtraction Task Problems

Subtraction Problems for Problem Solving Task

Type 1 Small minuends, small subtrahends, medium differences

5	6	7	6	7	8	7	8	9
-1	-1	-1	-2	-2	-2	-3	-3	-3
--	--	--	--	--	--	--	--	--
4	5	6	4	5	6	4	5	6

Type 2 Large minuends, small subtrahends, large differences

13	14	15	14	15	16	15	16	17
-1	-1	-1	-2	-2	-2	-3	-3	-3
--	--	--	--	--	--	--	--	--
12	13	14	12	13	14	12	13	14

Type 3 Large minuends, medium subtrahends, medium differences

13	14	15	14	15	16	15	16	17
-8	-8	-8	-9	-9	-9	-10	-10	-10
--	--	--	--	--	--	--	--	--
5	6	7	5	6	7	5	6	7

Type 4 Large minuends, large subtrahends, small differences

13	14	15	14	15	16	15	16	17
-12	-12	-12	-13	-13	-13	-14	-14	-14
--	--	--	--	--	--	--	--	--
1	2	3	1	2	3	1	2	3

Appendix B

Procedure Recognition Task

I know this boy/girl who's learning how to add and subtract. I gave him/her some addition and subtraction problems to solve. He/she told me about different ways to solve the problems. I would like you to tell me if the ways the boy/girl tried to solve each problem would work. Some of the ways may be "good ways" of solving the problems and some of the ways may be "silly ways" of solving the problems.

<u>Demo</u>

I. "good way" 1 + 2, 3 + 5

I asked this boy/girl to solve 1 + 2 and he/she said that he/she started with the first number, 1, and that was one finger and then it was plus 2 so he/she added 2 fingers so the answer was 3. Then he/she solved 3 + 5. He/she started with 3 fingers and added 5 more so the answer was 8. Is this a good way of solving addition problems or is it a silly way? Why?

II. "silly way" 4 + 3, 5 + 6

I asked this boy/girl to solve 4 + 3 and he/she said that he/she knew that he/she was 7 years old so the answer was 7. Then he/she solved 5 + 6. He/she knew that his older sister was 11 so the answer was 11. Is this a good way of solving addition problems or is it a silly way? Why?

<u>Task</u>

A. **counting down** 12 - 3, 14 - 6

The boy/girl saw the first problem, 12 - 3, and said he/she started with the first number, 12, and counted down by 3, so start at 12 then 11, 10, 9 and so the answer was 9. Then he/she solved 14 - 6. He/she started with the number 14 and counted down by 6, so start at 14 then 13, 12, 11, 10, 9, 8 and so the answer was 8. Is this a good way of solving subtraction problems or is it a silly way? Why?

B. **counting up** 11 - 8, 13 - 7

The boy/girl saw the first problem, 11 - 8, and said he/she started with the second number, 8, and counted up to 11, so start at 8 then 9, 10, 11 and so the answer was 3. Then he/she solved 13 - 7. He/she started with the number 7 and counted up to 13 so start at 7 then 8, 9, 10, 11, 12, 13 and so the answer was 6. Is this a good way of solving subtraction problems or is it a silly way? Why?

C. **deleting 10s** 15 - 2, 16 - 4

The boy/girl saw the first problem, 15 - 2, and said he/she took away the 1 from the 15 and he/she knew that 5 - 2 was 3 and then he/she put the 1 back on so the answer was 13. Then he/she solved 16 - 4. He/she took away the 1 from the 16 and he/she knew that 6 - 4 was 2 and then he/she put the 1 back on so the answer was 12. Is this a good way of solving subtraction problems or is it a silly way? Why?

D. **addition fact** 14 - 7, 12 - 5

The boy/girl saw the first problem, 14 - 7, and said he/she remembered that 7 + 7 is 14 so 14 - 7 must be 7. Then he/she solved 12 - 5. He/she said he/she remembered that 5 + 7 is 12 so he/she knew that 12 - 5 must be 7. Is this a good way of solving subtraction problems or is it a silly way? Why?

E. **combining digits** 12 - 1, 14 - 4

The boy/girl saw the first problem, 12 - 1 and said that he/she saw that the number on the left was 1 and the number on the right was 1 so the answer was 11. Then he/she solved 14 - 4. He/she said that the number on the left was 1 and the number on the right was 4 so the answer was 14. Is this a good way of solving subtraction problems or is it a silly way? Why?

F. **direction** 13 - 2, 11 - 5

The boy/girl saw the first problem, 13 - 2 and said that he/she knew the answer was 15 because 13 + 2 is 15. Then he/she solved 11 - 5. He/she said he/she knew the answer was 16 because 11 + 5 is 16. Is this a good way of solving subtraction problems or is it a silly way? Why?

Appendix C

Fluency Task Problems

Set 1

9-7	8-8	8-6	9-1	7-3
2-2	3-1	7-2	6-2	2-1
4-4	8-1	3-3	6-1	5-1
5-3	5-5	9-3	7-5	7-7
6-5	7-1	5-4	7-6	4-1
8-2	9-4	9-2	9-8	4-2
1-1	9-6	8-3	4-3	6-3
8-7	7-4	6-4	9-5	8-5
8-4	6-6	5-2	3-2	9-9

Set 2

11-3	19-5	17-9
12-4	12-1	14-6
18-9	18-5	14-5
13-3	14-2	18-2
11-4	11-8	13-7
12-2	13-4	14-9
16-5	15-5	16-3
11-6	19-3	18-7
13-2	16-9	19-2
17-6	14-7	18-3
13-6	17-1	12-8
15-6	16-7	15-8
14-1	13-8	17-7
19-9	13-1	19-4
11-7	11-2	13-9
17-3	12-7	11-9
15-3	18-4	11-1
18-8	17-2	16-4
16-6	15-4	14-8
15-9	16-1	15-1
12-6	17-4	12-5
18-6	14-3	16-8
17-8	18-1	11-5
19-8	14-4	19-1
15-2	19-7	17-5
12-9	15-7	13-5
19-6	16-2	12-3

Appendix D

Table D1

Proportion of Correct Responses on all Problem Types as a Function of Grade and Report Condition

		Problem Type				
Grade	Report	Type 1	Type 2	Type 3	Type 4	Mean
1						
	NR	.900	.717	.478	.406	.625
	RR	.889	.667	.261	.311	.532
	CR	.900	.761	.428	.522	.653
	Mean	.896	.715	.389	.413	.603
3						
	NR	.961	.878	.883	.789	.878
	RR	.983	.936	.836	.842	.899
	CR	.965	.959	.877	.942	.936
	Mean	.969	.923	.866	.856	.904
5						
	NR	.972	.944	.878	.894	.922
	RR	.950	.911	.906	.917	.921
	CR	.956	.928	.917	.917	.930
	Mean	.959	.928	.900	.909	.924

Appendix E

Table E1

Median Latencies (in seconds) of Correct Responses on all Problem Types as a Function of Grade and Report Condition

		Problem Type				
Grade	Report	Type 1	Type 2	Type 3	Type 4	Mean
1						
	NR	4.4	6.3	13.2	8.5	8.1
	RR	7.4	6.8	19.6	11.5	11.3
	CR	8.1	9.0	18.0	14.7	12.4
	Mean	6.6	7.4	16.5	11.8	10.6
3						
	NR	2.1	2.6	3.7	2.9	2.8
	RR	2.3	3.4	4.3	4.4	3.6
	CR	4.2	5.0	6.4	5.3	5.2
	Mean	2.9	3.7	4.8	4.2	3.9
5						
	NR	1.4	1.7	2.2	1.8	1.8
	RR	1.4	1.8	2.7	2.3	2.0
	CR	2.2	2.6	3.5	2.3	2.6
	Mean	1.7	2.0	2.8	2.1	2.2

Appendix F

Table F1

Median Latencies (in seconds) of Correct Responses on all Problem Types as a Function of Grade and Report Condition (involving RR in CR)

Grade	Report	Problem Type				
		Type 1	Type 2	Type 3	Type 4	Mean
1						
	NR	4.4	6.3	13.2	8.5	8.1
	RR	7.4	6.8	19.6	11.5	11.3
	RR in CR	8.2	8.9	18.9	13.6	12.4
	CR	7.8	9.1	14.6	17.8	12.3
	Mean	6.6	7.4	16.5	11.8	10.6
3						
	NR	2.1	2.6	3.7	2.9	2.8
	RR	2.3	3.4	4.3	4.4	3.6
	RR in CR	3.1	4.1	5.2	4.6	4.3
	CR	6.5	7.0	8.9	6.7	7.3
	Mean	2.9	3.7	4.8	4.2	3.9
5						
	NR	1.4	1.7	2.2	1.8	1.8
	RR	1.4	1.8	2.7	2.3	2.0
	RR in CR	1.7	2.2	2.8	1.7	2.1
	CR	3.3	3.7	5.1	3.6	3.9
	Mean	1.7	2.0	2.8	2.1	2.2

Note. In Grades 1 and 5, ns = 20, 20, 6, and 14 for NR, RR, RR in CR, and CR, respectively. In Grade 3, ns = 20, 19, 6, and 13 for NR, RR, RR in CR, and CR, respectively.

Appendix G

Raw Subtraction Problem Strategies

Main Categories

1. Counting (CC) - includes all strategies that were predominantly counting
2. Retrieval (KK) - includes retrieval, knowing it
3. Suspected Retrieval (EK) - Ss didn't report using retrieval but fast and accurate so coders suspected the use of retrieval
4. Derived Fact (DF) - includes all strategies that predominantly used derived facts
5. Special Trick (ST) - includes all strategies that predominantly used special tricks
6. Addition Fact (AF) - includes all strategies that were predominantly using addition facts
7. Unknown (UN) - unknown strategies due to NR, not reporting in report conditions, and cut-offs
8. Guessing (GG) - includes guessing at the answer
9. Other (WW)- includes strange strategies, guessing, or very rare strategies (e.g., WF)

Raw strategies

1. Counting

   C1 - counting down method 1 (8-5: 8,7,6,5,4,3: =3) CD1

   C2 - counting down method 2 (8-5: 8,7,6,5:=3) CD2

   C3 - counting up (8-5:5,6,7,8: =3) CU

   C4 - counting strategy that doesn't belong in any of the other counting categories CX/CO

   C5 - counting (no more detailed info) C

C6 - count all CA

C7 - count all but represent first number CR

C8 - counting down (not sure if method 1 or 2)CD

F1 - counting down method 1 with fingers CDF1

F2 - counting down method 2 with fingers CDF2

F3 - counting up with fingers CUF

F4 - counting with fingers that doesn't belong in any other categorie CFX

F5 - counting with fingers CF

F6 - count all with fingers CA

F7 - count all with fingers but represent first number CRF

F8 - counting with fingers + guessing CFG

F9 - counting with fingers + adding CFA

CG1 - counting with fingers that doesn't belong with other categories + guessing CFX+G

T1 - counting fingers + knowing CF+K

T2 - counting down + knowing CD+K

2. Retrieval

K - knowing it, retrieval

3. Suspected Retrieval

EK - experimenter thinks subject retrieved although subject gave no strategy report

4. Derived Fact

D1 - decomposition down over 10 (14-6:14-10=4,6-4=2,10-2=8) D1

D2 - decomposition take from the 10 (14-6:10-6=4,14-10=4,4+4=8)

D4 - decomposition method 3 (14-8:14-4=10,10-4=6) DQ

DX - decomposition method 4 (14-8:(14-2)-(8+2)=12-10=2,8-2=6) DX

E1 - derived fact (no more details) DF

E2 - derived fact doubles (9-5:10-5=5-1=4) DF1

E3 - derived fact sharing (14-8:(7+1)-(7-1)=14,8+6=14) DF2

E4 - derived fact going thru 10 (13-9:10+3=13,9+1=10,3+1=4) DF3

E5 - derived fact maintain difference (15-7:15+1=16,7+1=8,16-8=8) DF4

5. Special Trick

D0 - delete tens + counting up DT+CU

D3 - delete tens DT

D5 - using fractions (14-12:DT then 2 is ½ of 4 so=2) DFF

D6 - using multiplication tables (e.g.?) DFM

D7 - delete tens + addition fact DT+AF

D8 - delete tens + related fact DT+RF

D9 - delete tens + counting down DT+CD/DT+CDF1/etc.

DS - special trick (see below) + derived fact ST+DT

DK - delete tens + retrieval DT+K

OA - ones rule type 1 (14-1=13 because 13 is just one less than 14)

OB - ones rule type 2 (14-13=1 because you just add one more to get 14)

ST - special trick not included in other DD codes (15-10, count by 5s, 5,10,15)

ST

AD - addition fact + derived fact AD

RF - related fact (8-2:8-3=5so8-2=1more=6)

AR - related fact + addition fact

T0 - delete tens + special trick DT+ST

T3 - counting down + related fact CDRF

T5 - addition fact + DQ AF+DQ

T7 - delete tens + using fractions DT+DFF

T8 - delete tens + using multiplication tables DT+DFM

T9 - special trick + counting down ST+CD/ST+CDF/ST+CD1/ST+CDF1/etc.

6. Addition Fact

AF - using an addition fact (8-5=3 because 5+3=8)

T4 - counting down + addition fact CDAF

T6 - counting up + addition fact CUAF

7. Unknown

U - unknown

8. Guessing

G - guessing G

9. Other

Q - coder can't figure out what strategy is Q

W - weird strategy W

A - addition (instead of subtraction) A

WF - solving problem by visualizing problem in written form (vertical rather than horizontal)

Appendix H

Descriptions of strategy categories

The counting category included almost all of the instances of counting strategies, including counting up (8 - 5: 6,7,8: = 3), counting down (8 - 5: 8,7, 6, 5, 4, 3: = 3), and counting all (8 - 5: 1,2,...,8 - 1,2,..,5: =3), with use of fingers or not. All of these counting strategies were lumped together to reflect that this family of strategies is the least sophisticated of all of the possible solution strategy groups and related to this, generally the slowest.

Retrieval, however, is generally considered to be the most sophisticated of the solution strategies. Indeed, it is even debatable whether this solution procedure is even a strategy as it usually occurs automatically and effortlessly (e.g., Bisanz & LeFevre, 1990). This solution procedure or strategy as it is labelled in this study, is generally fast and accurate. Students using retrieval often reported that "the answer just popped into my head" or "I just know it."

Suspected retrieval was used on trials where, although the students reported no strategy, the students appeared to be using retrieval (very fast solution times with few errors). Thus, this category is not based on the verbal reports of the students but on the judgment of the experimenter. Note that the reliability for this category was very high at 93.9%. The experimenter suspected the use of retrieval frequently enough to deem a separate category justifiable. The reports classified as suspected retrieval would otherwise have been classified as unknown. (Also note that the task of concurrently reporting retrieval can be extremely difficult. For example, the reader is invited to concurrently report her solution procedure while solving 2 + 2. This difficulty is reflected in the high rates of suspected retrieval in the concurrent report condition for the older students)

The derived fact category included strategies that involved breaking up a problem into smaller and easier components. For example, for the problem 14 - 6 a student might break the problem down into the easier problem of 14 - 4 = 10 and then subtract another 2 from the 10 to get the answer of 8. Another example, for the problem 9 - 5, would be to changed it into an easier problem involving doubles, 10 - 5 = 5, and because 9 is one less than 10, 1 would be subtracted from 5 thus the answer would be 4. Strategies in this category rely on known subtraction or related addition facts to help solve the problem.

The special trick category consisted of strategies that could only be used on very specific problems or types of problems (unlike derived facts that can be applied to all problems). However, as with the derived fact category, strategies in the special trick category were used by the students to make a problem easier to solve. For example, many students reported using the special trick of deleting tens on problems such as 15 - 12. In these problems, students would cancel out the ones (or tens) and would be left with, for example, 5 - 2 to solve. Students might then retrieve the answer or use counting. Solutions that included counting were not categorized as counting because before counting occurred, the students were using a more sophisticated and efficient strategy to simplify the problem.

The use of corresponding addition facts to solve subtraction problems is another common subtraction strategy (Geary, 1994). For example, on a problem such as 10 - 2, a child would determine what amount + 2 would give 10.

Unknown was the category used when the students were unable to report a solution strategy or if the cut-off protocol had been implemented before they had arrived at a solution. That is, often the younger students had great difficulty with the more difficult problems (e.g., 17 - 14). In these problems, counting often resulted in losing track of the numbers or the students realizing that they possessed insufficient fingers and did not know what other strategy to use instead.

Guessing is a fairly common procedure that children use when they are unable to arrive at a solution for a problem and occurred frequently enough, especially for the younger students, that it became its own category. The final category labelled other essentially grouped together all remaining strategies such as doing addition instead of subtraction but primarily consisted of strategy reports that were incomprehensible or bizarre. An interesting example of an other strategy was, for the problem 15 - 3, a student (Grade 1) reported that "the answer is 15 and the way I got that answer was that I covered over the 3 and so I took away the 3 and I was left with 15."

Appendix I

Table I1

Proportions of the Frequency of the Strategy Categories Reported by the Retrospective (RR) and Concurrent report (CR) students as a Function of Grade and Problem Type

		Strategy Category								
Grade	Report	CC	KK	EK	DF	ST	AF	UN	GG	WW
		Problem Type 1								
1[a]										
	RR	.54	.22	.09	.00	.07	.01	.05	.02	.01
	CR	.46	.19	.12	.00	.07	.02	.13	.01	.01
	Mean	.50	.21	.10	.00	.07	.01	.09	.01	.01
3[b]										
	RR	.25	.30	.06	.00	.18	.14	.05	.00	.01
	CR	.06	.28	.18	.02	.09	.25	.12	.00	.01
	Mean	.16	.29	.12	.01	.14	.19	.08	.00	.01
5[a]										
	RR	.21	.36	.18	.01	.11	.13	.00	.00	.01
	CR	.08	.32	.41	.00	.08	.08	.02	.00	.00
	Mean	.15	.34	.29	.01	.10	.10	.01	.00	.00
		Problem Type 2								
1										
	RR	.40	.19	.07	.00	.18	.00	.09	.04	.03
	CR	.43	.15	.10	.00	.17	.01	.11	.02	.02
	Mean	.42	.17	.09	.00	.18	.00	.10	.03	.02

table continues

		Strategy Category								
Grade	Report	CC	KK	EK	DF	ST	AF	UN	GG	WW
		Problem Type 2								
3										
	RR	.24	.12	.03	.01	.57	.01	.02	.00	.01
	CR	.18	.16	.11	.00	.29	.14	.12	.00	.01
	Mean	.21	.14	.07	.00	.43	.07	.07	.00	.01
5										
	RR	.23	.26	.06	.01	.34	.08	.00	.00	.02
	CR	.09	.26	.37	.00	.17	.09	.01	.00	.01
	Mean	.16	.26	.21	.00	.25	.09	01	.00	.01
		Problem Type 3								
1										
	RR	.44	.04	.05	.03	.07	.02	.24	.09	.03
	CR	.64	.08	.03	.02	.06	.00	.17	.02	.01
	Mean	.52	.06	.04	.02	.06	.01	.21	.06	.02
3										
	RR	.32	.12	.02	.07	.27	.09	.05	.01	.05
	CR	.20	.13	.05	.09	.13	.26	.10	.01	.02
	Mean	.26	.13	.03	.08	.20	.18	.07	.01	.04
5										
	RR	.26	.16	.04	.12	.19	.21	.02	.00	.01
	CR	.20	.20	.27	.11	.09	.11	.02	.00	.00
	Mean	.23	.18	.16	.12	.14	.16	.02	.00	.01
		Problem Type 4								
1										
	RR	.34	.07	.05	.01	.14	.01	.20	.14	.04
	CR	.46	.11	.06	.00	.17	.00	.17	.02	.02
	Mean	.40	.09	..06	.00	.16	.00	18	.08	.03

table continues

Grade	Report	Strategy Category CC	KK	EK	DF	ST	AF	UN	GG	WW
		Problem Type 4								
3										
	RR	.13	.11	.06	.01	.65	.00	.03	.01	.01
	CR	.18	.11	.10	.00	.35	.13	.11	.00	.03
	Mean	.15	.11	.08	.00	.50	.06	.07	.00	.02
5										
	RR	.22	.20	.08	.01	.37	.11	.01	.00	.00
	CR	.09	.29	.34	.01	.18	.07	.01	.00	.01
	Mean	.16	.25	.21	.01	.27	.09	.01	.00	.00
		All Problems								
1										
	RR	.43	.13	.07	.01	.11	.01	.15	.07	.03
	CR	.49	.13	.08	.00	.12	.01	.14	.02	.01
	Mean	.46	.13	.07	.01	.12	.01	.14	.04	.02
3										
	RR	.24	.16	.04	.02	.42	.06	.04	.00	.02
	CR	.15	.17	.11	.03	.22	.19	.11	.00	.01
	Mean	.20	.17	.08	.02	.32	.13	.07	.00	.02
5										
	RR	.23	.24	.09	.04	.25	.13	.01	.00	.01
	CR	.12	.27	.35	.03	.13	.09	.02	.00	.00
	Mean	.17	.26	.22	.03	.19	.11	.01	.00	.01

Note. The strategy category abbreviations are as follows: CC - counting, KK - retrieval, EK - suspected retrieval, DF - derived fact, ST - special trick, AF - addition fact, UN - unknown, GG - guessing, and WW - other.

[a]n = 20 in each report condition group. [b]n = 19 in each report condition group.

Appendix J

Table J1

Proportions of the Frequency of the Strategy Categories Reported by the Retrospective (RR), Concurrent (CR), and Retrospective in Concurrent (RR in CR) Report Condition students as a Function of Grade and Problem Type

		Strategy Category								
Grade	Report	CC	KK	EK	DF	ST	AF	UN	GG	WW
		Problem Type 1								
1[a]										
	RR	.54	.22	.09	.00	.07	.01	.05	.02	.01
	CR	.57	.20	.09	.00	.00	.00	.13	.00	.00
	RR in CR	.40	.18	.13	.00	.10	.03	.13	.02	.01
	Mean	.50	.21	.10	.00	.07	.01	.09	.01	.01
3[b]										
	RR	.25	.30	.06	.00	.18	.14	.05	.00	.01
	CR	.04	.09	.22	.02	.07	.37	.17	.00	.02
	RR in CR	.08	.37	.15	.02	.10	.19	.09	.00	.00
	Mean	.16	.29	.12	.01	.14	.19	.08	.00	.01
5[a]										
	RR	.21	.36	.18	.01	.11	.13	.00	.00	.01
	CR	.22	.30	.00	.00	.24	.17	.07	.00	.00
	RR in CR	.02	.33	.59	.00	.02	.04	.00	.00	.00
	Mean	.15	.34	.29	.01	.10	.10	.01	.00	.00

table continues

Grade	Report	CC	KK	EK	DF	ST	AF	UN	GG	WW
		Strategy Category								
		Problem Type 2								
1										
	RR	.40	.19	.07	.00	.18	.00	.09	.04	.03
	CR	.56	.19	.06	.00	.07	.00	.13	.00	.00
	RR in CR	.38	.13	.12	.00	.21	.01	.10	.02	.02
	Overall	.42	.17	.09	.00	.18	.00	.10	.03	.02
3										
	RR	.24	.12	.03	.01	.57	.01	.02	.00	.01
	CR	.19	.00	.11	.00	.22	.30	.17	.00	.02
	RR in CR	.17	.24	.11	.00	.32	.07	.09	.00	.00
	Overall	.21	.14	.07	.00	.43	.07	.07	.00	.01
5										
	RR	.23	.26	.06	.01	.34	.08	.00	.00	.02
	CR	.17	.31	.00	.00	.33	.17	.02	.00	.00
	RR in CR	.06	.24	.00	.10	.06	.01	.52	.00	.02
	Overall	.16	.26	.21	.00	.25	.09	.01	.00	.01
		Problem Type 3								
1										
	RR	.44	.04	.05	.03	.07	.02	.24	.09	.03
	CR	.67	.17	.00	.02	.02	.00	.13	.00	.00
	RR in CR	.58	.05	.04	.02	.08	.00	.19	.03	.02
	Overall	.52	.06	.04	.02	.06	.01	.21	.06	.02
3										
	RR	.32	.12	.02	.07	.27	.09	.05	.01	.05
	CR	.19	.02	.02	.07	.15	.35	.15	.00	.08
	RR in CR	.21	.19	.06	.09	.13	.22	.08	.02	.00
	Overall	.26	.13	.03	.08	.20	.18	.07	.01	.04

table continues

		Strategy Category								
Grade	Report	CC	KK	EK	DF	ST	AF	UN	GG	WW
		Problem Type 3								
5										
	RR	.44	.04	.05	.03	.07	.02	.24	.09	.03
	CR	.37	.30	.00	.02	.09	.19	.04	.00	.00
	RR in CR	.13	.16	.39	.15	.10	.07	.01	.00	.00
	Overall	.23	.18	.16	.12	.14	.16	.02	.00	.01
		Problem Type 4								
1										
	RR	.34	.07	.05	.01	.14	.01	.20	.14	.04
	CR	.50	.19	.04	.00	.09	.00	.19	.00	.00
	RR in CR	.44	.07	.07	.00	.21	.00	.16	.02	.03
	Overall	.40	.09	.06	.00	.16	.00	.18	.08	.03
3										
	RR	.13	.11	.06	.01	.65	.00	.03	.01	.01
	CR	.11	.00	.07	.00	.26	.31	.17	.00	.07
	RR in CR	.21	.15	.11	.00	.39	.04	.09	.00	.01
	Overall	.15	.11	.08	.00	.50	.06	.07	.00	.02
5										
	RR	.22	.20	.08	.01	.37	.11	.01	.00	.00
	CR	.17	.30	.00	.00	.33	.17	.04	.00	.00
	RR in CR	.06	.29	.48	.02	.11	.02	.00	.00	.01
	Overall	.16	.25	.21	.01	.27	.09	.01	.00	.00
		All Problems								
1										
	RR	.43	.13	.07	.01	.11	.01	.15	.07	.03
	CR	.57	.19	.05	.00	.05	.00	.14	.00	.00
	RR in CR	.45	.11	.09	.00	.15	.01	.14	.02	.02
	Overall	.46	.13	.07	.01	.12	.01	.14	.04	.02

table continues

		Strategy Category								
Grade	Report	CC	KK	EK	DF	ST	AF	UN	GG	WW
		All Problems								
3										
	RR	.24	.16	.04	.02	.42	.06	.04	.00	.02
	CR	.13	.03	.11	.02	.18	.33	.16	.00	.04
	RR in CR	.17	.24	.11	.03	.24	.13	.09	.00	.00
	Overall	.20	.17	.08	.02	.32	.13	.07	.00	.02
5										
	RR	.23	.24	.09	.04	.25	.13	.01	.00	.01
	CR	.23	.30	.00	.00	.25	.17	.04	.00	.00
	RR in CR	.07	.26	.50	.04	.08	.05	.00	.00	.01
	Overall	.17	.26	.22	.03	.19	.11	.01	.00	.01

Note. The strategy category abbreviations are as follows: CC - counting, KK - retrieval, EK - suspected retrieval, DF - derived fact, ST - special trick, AF - addition fact, UN - unknown, GG - guessing, and WW - other.

[a]n = 20 in each RR group, 6 in each CR group, and 14 in RR in each CR group. [b]n = 19 in the RR group, 6 in the CR group, and 13 in the RR in CR group.

Appendix K

Table K1

Strategy Use and Accuracy (in Percentages), and Latencies (in Seconds) of Each Strategy for Retrospective (RR) and Concurrent Report (CR) Students as a Function of Grade and Problem Type

	Grade																	
	1						3						5					
	Report Condition																	
	RR			CR			RR			CR			RR			CR		
Strategy	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc
	Problem Type 1																	
Counting	54.5	9.2	85.7	45.6	8.7	79.3	25.1	3.2	95.3	6.4	4.3	90.9	21.1	1.9	89.5	8.3	2.7	100.0
Retrieval	22.2	3.4	92.5	18.9	3.7	97.1	29.8	1.6	97.1	28.1	2.5	100.0	35.6	1.1	98.4	32.2	1.5	96.6

table continues

	Grade																	
	1						3						5					
	Report Condition																	
	RR			CR			RR			CR			RR			CR		
Strategy	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc
Problem Type 1																		
Susp. Ret	8.9	4.3	93.8	11.7	5.0	100.0	6.4	1.1	100.0	17.5	1.7	96.7	17.8	1.2	93.8	41.1	1.4	100.0
Deriv. Fact	0.0	---	0.0	0.0	---	0.0	0.0	---	0.0	1.8	12.5	100.0	1.1	1.7	100.0	0.0	---	0.0
Spec. Trick	6.7	3.8	100.0	8.3	8.3	80.0	18.1	2.5	100.0	9.4	3.2	93.8	11.1	1.6	100.0	8.3	6.0	73.3
Addit. Fact	0.6	7.3	100.0	2.2	6.5	100.0	14.0	3.2	100.0	24.6	4.0	97.6	12.3	1.4	91.3	7.8	2.7	100.0
Unknown	5.0	9.2	77.8	12.8	9.6	100.0	5.3	2.5	100.0	11.7	5.3	90.0	0.0	---	0.0	2.2	3.7	50.0
Guess	1.7	5.4	100.0	1.1	---	0.0	0.0	---	0.0	0.0	---	0.0	0.0	---	0.0	0.0	---	0.0
Other	0.6	8.2	100.0	0.6	46.2	100.0	1.2	11.0	100.0	0.6	38.4	100.0	0.6	1.1	100.0	0.0	---	0.0
Overall	100.0	6.0	88.9	100.0	6.6	88.3	100.0	2.1	98.2	100.0	3.2	96.5	100.0	1.3	95.0	100.0	1.6	95.6

table continues

	Grade																	
	1						3						5					
	Report Condition																	
	RR			CR			RR			CR			RR			CR		
Strategy	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc
Problem Type 2																		
Counting	40.0	8.1	59.7	43.3	4.1	61.5	24.0	3.8	95.1	17.5	5.4	93.3	23.3	2.2	92.9	8.9	3.1	87.5
Retrieval	19.4	4.1	97.1	15.0	3.4	88.9	12.3	1.8	100.0	16.4	2.8	96.4	26.1	1.5	89.4	26.1	1.8	95.7
Susp. Ret.	7.2	3.8	84.6	10.0	4.4	94.7	2.9	1.6	100.0	11.1	1.9	100.0	6.1	1.6	90.9	36.7	1.3	93.9
Deriv. Fact	0.0	---	0.0	0.0	---	0.0	0.6	---	0.0	0.0	---	0.0	0.6	2.4	100.0	0.0	---	0.0
Spec. Trick	17.8	5.1	65.6	17.2	10.3	93.5	57.3	2.9	92.9	28.7	5.4	91.8	33.9	1.8	93.4	16.7	4.9	90.0
Addit. Fact	0.0	---	0.0	0.6	16.8	100.0	0.6	1.7	100.0	14.0	5.5	100.0	8.3	2.3	86.7	9.4	3.1	88.2
Unknown	8.9	6.5	43.8	10.6	6.3	78.9	1.8	3.6	100.0	11.7	5.5	100.0	0.0	---	0.0	1.1	4.4	100.0
Guess	3.9	5.5	57.1	1.7	---	0.0	0.0	---	0.0	0.0	---	0.0	0.0	---	0.0	0.0	---	0.0
Other	1.3	---	0.0	1.7	6.5	100.0	0.6	---	0.0	0.6	4.6	100.0	1.7	1.5	66.7	1.1	4.8	100.0
Overall	100.0	5.2	66.7	100.0	6.5	76.1	100.0	2.1	93.6	100.0	4.4	95.9	100.0	1.3	91.1	100.0	1.9	92.8

table continues

	Grade																	
	1						3						5					
	Report Condition																	
	RR			CR			RR			CR			RR			CR		
Strategy	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc
Problem Type 3																		
Counting	43.9	13.1	38.0	60.6	19.3	42.2	32.2	7.0	81.8	20.5	7.0	94.3	25.6	3.1	91.3	20.0	6.3	88.9
Retrieval	3.9	3.3	92.9	8.3	4.4	66.7	12.3	1.8	90.5	13.5	2.9	100.0	15.6	1.5	92.9	20.0	1.8	88.9
Susp. Ret.	5.0	---	0.0	2.8	4.7	60.0	1.8	1.0	100.0	4.7	2.6	100.0	3.9	1.3	100.0	27.2	1.0	91.8
Deriv. Fact	2.8	13.6	80.0	1.7	8.8	100.0	7.0	7.1	91.7	8.8	4.6	73.3	12.2	2.7	68.2	11.1	4.0	95.0
Spec. Trick	6.7	14.9	33.3	6.1	15.3	81.8	26.7	2.9	95.7	13.5	9.0	87.0	18.9	2.4	94.1	9.4	3.9	100.0
Addit. Fact	1.7	3.0	100.0	0.0	---	0.0	9.4	2.9	87.5	26.3	4.6	86.7	20.6	2.4	100.0	10.6	3.8	89.5
Unknown	24.4	6.3	2.3	17.2	8.0	16.1	4.7	2.8	37.5	9.9	5.0	76.5	2.2	6.0	75.0	1.7	6.3	100.0
Guess	8.9	16.7	12.5	2.2	11.1	25.0	0.6	---	0.0	1.2	1.9	50.0	0.0	---	0.0	0.0	---	0.0
Other	2.8	---	0.0	1.1	---	0.0	5.3	3.3	44.4	1.8	12.6	66.7	1.1	8.0	50.0	0.0	---	0.0
Overall	100.0	12.9	26.1	100.0	15.1	42.8	100.0	3.3	83.6	100.0	5.4	87.7	100.0	2.4	90.6	100.0	2.7	91.7

table continues

	Grade																	
	1						3						5					
	Report Condition																	
	RR			CR			RR			CR			RR			CR		
Strategy	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc
Problem Type 4																		
Counting	34.4	15.4	38.7	45.6	20.5	53.7	12.9	8.9	77.3	17.5	8.3	77.1	22.2	2.3	85.0	9.4	4.2	94.1
Retrieval	6.7	3.3	75.0	10.6	3.5	73.7	11.1	1.8	100.0	10.5	2.8	100.0	20.0	1.3	97.2	29.4	1.3	90.6
Susp. Ret.	5.0	3.8	11.1	6.1	5.2	81.8	5.8	1.8	90.0	9.9	1.9	100.0	7.8	1.1	100.0	33.9	1.2	91.8
Deriv. Fact	0.6	8.4	100.0	0.0	---	0.0	0.6	1.9	100.0	0.0	---	0.0	0.6	1.9	100.0	1.1	2.5	100.0
Spec. Trick	13.9	7.4	60.0	17.2	10.4	58.1	65.5	3.1	83.0	35.1	4.8	32.2	37.2	1.9	91.0	17.8	5.7	90.6
Addit. Fact.	0.6	4.1	100.0	0.0	---	0.0	0.0	---	0.0	12.9	6.0	100.0	11.1	1.5	95.0	6.7	1.9	100.0
Unknown	20.0	6.8	2.8	16.7	7.6	23.3	2.9	2.0	80.0	11.1	5.4	94.7	1.1	6.0	50.0	1.1	6.0	50.0
Guess	14.4	16.7	11.5	1.7	---	0.0	0.6	---	0.0	0.0	---	0.0	0.0	---	0.0	0.0	---	0.0
Other	4.4	3.2	12.5	2.2	21.6	50.0	0.6	2.0	100.0	2.9	6.1	80.0	0.0	---	0.0	0.6	2.9	100.0
Overall	100.0	9.5	31.1	100.0	9.5	52.2	100.0	2.9	84.2	100.0	4.7	94.2	100.0	1.7	91.7	100.0	1.6	91.7

table continues

	Grade																	
	1						3						5					
	Report Condition																	
	RR			CR			RR			CR			RR			CR		
Strategy	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc	Use	Mdn	Acc
All Problems																		
Counting	43.2	10.3	58.2	48.8	12.7	57.8	23.5	4.5	88.2	15.5	7.0	92.5	23.1	2.3	89.8	11.7	4.3	91.7
Retrieval	13.1	9.4	88.3	13.2	3.5	85.3	16.4	1.6	97.3	17.1	2.7	99.1	24.3	1.3	94.9	26.9	1.6	93.3
Susp. Ret.	6.5	4.1	57.4	7.6	4.8	90.9	4.2	1.5	96.6	10.8	1.9	98.6	8.9	1.3	95.3	34.7	1.2	94.8
Deriv. Fact	0.8	13.0	83.3	0.4	8.8	100.0	2.0	6.3	85.7	2.6	7.5	77.8	3.6	2.5	73.1	3.1	3.3	95.5
Spec. Trick	11.3	5.2	64.2	12.2	10.4	77.3	42.0	2.9	87.8	21.6	5.1	93.9	25.3	1.9	93.4	13.1	5.1	89.4
Addit. Fact	0.7	3.6	100.0	0.7	6.6	100.0	6.0	3.7	95.1	19.4	4.6	94.7	13.2	1.7	94.7	8.6	2.9	93.5
Unknown	14.6	8.0	15.2	14.3	8.4	48.5	3.7	2.6	76.0	11.1	5.3	90.8	0.8	6.0	66.7	1.5	4.8	72.7
Guess	7.2	7.5	23.1	1.7	11.1	8.3	0.3	---	0.0	0.3	1.9	50.0	0.0	---	0.0	0.0	---	0.0
Other	2.6	5.7	10.5	1.4	12.6	60.0	1.9	2.0	53.8	1.5	8.9	80.0	0.8	1.6	66.7	0.4	4.3	100.0
Overall	100.0	6.8	53.2	100.0	8.0	64.9	100.0	2.4	88.9	100.0	4.4	94.2	100.0	1.7	92.1	100.0	1.9	92.9

Appendix L

Table L1

Regression Analyses of Median Latencies on Only Correct Trials and Both Correct and Incorrect Trials for the Different Reported Strategy Categories as a Function of Report Condition and Grade

Group	All trials				Counting trials				Retrieval trials				Special tricks trials			
	Correct only		Both		Correct only		Both		Correct only		Both		Correct only		Both	
Predictor	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b
								Grade 1								
RR																
a	.46**	.51	.52**	.52	.003	.10	.24**	.64	.000	-.002	.000	-.01	.07	.21	.21	.71
b	.59**	.50	.65**	.50	.05	.37	.53**	.72	.01	.03	.07	.11	.73**	.31	.14	.36
d	.28	-.27	.30	-.26	.04	-.35	.17**	-.48	.01	-.02	.05	-.07	.36	-.20	.01-	.08
a + b	.61**	.33	.67**	.33	.03	.17	.50**	.45	.002	.01	.02	.03	.56*	-.19	.24	.36
small count	.63**	1.33	.48**	.93	.06	.93	.36**	.89	.63**	1.64	.59**	1.42	.01	-.39	.57**	5.94
b > 10	.23	-2.16	.41*	-3.46	.01	-2.04	.26*	-5.63	.004	.31	.01	-.44	.70**	-3.41	.10	-3.34
n problems	36		36		30		36		12		14		9		12	
CR																
a	.43**	.66	.44**	.73	.28**	1.08	.33*	1.17	.01	-.02	.01	.03	.29	.65	.38*	1.27
b	.51**	.60	.41*	.52	.40**	.98	.54**	1.15	.05	.06	.001	.01	.03	-.14	.15	.46

table continues

Group	All trials				Counting trials				Retrieval trials				Special tricks trials			
	Correct only		Both		Correct only		Both		Correct only		Both		Correct only		Both	
Predictor	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b
								Grade 1								
CR																
d	.21	-.30	.09	-.13	.07	-.48	.12*	-.63	.06	-.04	.02	.03	.29	.43	.002	-.05
a + b	.53**	.41	.49**	.39	.45**	.66	.56**	.75	.000	.002	.002	.01	.01	.07	.31*	.48
small count	.65**	1.93	.63**	1.99	.44**	2.54	.29*	2.09	.16	.23	.09	.27	.18	1.16	.11	1.66
b > 10	.21	-2.88	.09	-1.17	.07	-4.47	.42*	-7.27	---[a]	---	.01	.28	.12	3.06	.10	-4.39
n problems	36		36		36		36		12		18		13		16	
								Grade 3								
RR																
a	.40*	.17	.42**	.16	.27**	.39	.34**	.40	.01	.01	.04	.02	.09	.12	.10	.12
b	.30	.10	.34*	.10	.65**	.50	.58**	.41	.01	.01	.04	.02	.10	.08	.15*	.10
d	.003	.01	.02	-.05	.15	-.25	.11	-.19	.001	.003	.001	-.004	.02	-.04	.05	-.06
a + b	.39*	.08	.42**	.08	.62**	.31	.63**	.28	.02	.01	.05	.01	.13*	.06	.17*	.07
small count	.59**	.46	.55**	.41	.15	.56	.19*	.57	.07	.05	.17	.09	.09	.18	.10	.19
b > 10	.01	.03	.05	-.15	.35**	-4.54	.24*	-3.08	.001	.03	.01	.10	.03	-.46	.04	-.56
n problems	36		36		23		24		20		21		32		32	
CR																
a	.43**	.16	.42**	.18	.18	.93	.03	.54	.23*	.09	.21*	.08	.11	.41	.12	.32
b	.30	.09	.29**	.09	.13	.20	.15	.31	.09	.06	.11	.06	.05	.16	.02	.07

table continues

Group	All trials				Counting trials				Retrieval trials				Special tricks trials			
	Correct only		Both		Correct only		Both		Correct only		Both		Correct only		Both	
Predictor	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b
								Grade 3								
CR																
d	.03	.01	.04	.02	.07	-.14	.12	-.28	.04	.04	.02	.02	.002	-.04	.003	.03
a + b	.40*	.07	.39**	.08	.18	.22	.16	.30	.20*	.05	.21*	.05	.09	.16	.06	.10
small count	.49**	.35	.53**	.42	.15	.52	.02	.27	.11	.15	.12	.15	.13	.84	.29**	.91
b > 10	.05	-.17	.01	.03	.03	-.91	.12	-2.66	.003	-.13	.01	-.18	.001	-.20	.02	.71
n problems	36		36		17		19		20		20		24		24	
								Grade 5								
RR																
a	.37*	.08	.35*	.08	.11	.08	.08	.08	.09	.03	.12	.04	.004	.03	.002	.02
b	.19	.03	.23	.04	.09	.06	.07	.06	.000	.001	.01	.001	.01	.02	.01	.03
d	.10	.02	.05	.09	.002	.01	.004	.02	.07	.03	.09	.03	.01	-.02	.01	-.02
a + b	.31	.03	.32	.03	.12	.05	.10	.04	.02	.01	.09	.03	.01	.02	.01	.02
small count	.62**	.26	.66**	.27	.48**	.36	.39**	.35	.11	.08	.13	.10	.24*	.28	.31**	.32
b > 10	.15	.27	.13	.33	.004	.14	.003	.13	.01	.06	.01	.13	.02	.28	.03	.35
n problems	36		36		29		30		26		29		25		26	
CR																
a	.28	.06	.23	.05	.06	.17	.04	.14	.01	.02	.01	.02	.000	-.003	.01	.09

table continues

Group	All trials				Counting trials				Retrieval trials				Special tricks trials			
	Correct only		Both		Correct only		Both		Correct only		Both		Correct only		Both	
Predictor	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b	r^2	b
							Grade 5									
CR																
b	.07	.01	.04	.06	.04	.11	.03	.09	.03	-.02	.02	-.02	.03	.07	.10	.16
d	.16	.03	.16	.03	.000	-.002	.000	-.004	.09	.04	.07	.04	.03	-.08	.07	-.15
a + b	.18	.02	.14	.01	.06	.08	.04	.07	.004	-.01	.002	-.004	.02	.04	.08	.11
small count	.65**	.27	.64**	.26	.49*	.75	.44*	.68	.06	.08	.16*	.13	.05	.25	.01	.13
b > 10	.26	.48	.30	.54	.11	2.39	.11	2.29	.08	.36	.14*	.54	.001	-.11	.04	-1.13
n problems	36		36		11		11		32		36		14		17	

Table L2

Frequencies (in percentages and number of trials) of Reported Counting Up and Counting Down as a Function of Grade and Report Condition

Grade	Report Condition	Counting Strategy	
		Counting Up	Counting Down
1			
	RR	0.8 (6)	14.4 (104)
	CR	0.3 (2)	14.3 (113)
	Mean	0.6	14.4
3			
	RR	2.9 (20)	14.0 (96)
	CR	5.9 (40)	8.0 (55)
	Mean	4.4	11.0
5			
	RR	6.3 (45)	12.5 (90)
	CR	3.6 (26)	7.1 (51)
	Mean	4.9	9.8

Note. Percentages of reported counting up use corresponded fairly well to Siegler's findings (1% and 5% for Grades 2 and 4, respectively). However, a frequency of 1% in Siegler's study corresponds to between 7 and 19 reported uses of counting up out of a possible 1332 trials (36 problems and 37 subjects) and in this study to 4 to 10 times out of a possible 720 trials (36 problems and 20 subjects). Analyses based on so few trials could yield seriously misleading results.

Appendix M

Procedure Recognition Task Sample Responses

Legitimate procedures: counting down, counting up, deleting 10s, and addition fact. The correct recognition response would be "good." Illegitimate procedures: combining digits and addition instead of subtraction. The correct recognition response would be "silly." See Appendix B for complete task. Below are examples of correct and incorrect recognition responses paired with appropriate or inappropriate justifications.

1. R(ecognition): Good, J(ustification): Appropriate

a. Addition fact: "It's good because using addition to help with subtracting." (Grade 5 student)

b. Combining digits: "He memorized [both] the problem and the answer." (Grade 3)

c. Counting down: "Sometimes don't always just know the answer so it helps to do it this way" (Grade 1)

2. R: Good, J: Inappropriate

a. Addition fact [why using 7 + 7 = 14 to solve 14 - 7 and using 5 + 7 = 12 to solve 12 - 7, respectively, is good]: "7 x 2 = 14 and 7 x 1 = 7." (Grade 5)

b. Counting up: "Because it's a neat way." (Grade 1)

c. Combining digits: "[If you use this way] you'll have the answer. But it doesn't really matter if you get a good mark because everyone will still love you. All that matters is if you know the answer." (Grade 1)

3. R: Silly, J: Appropriate

a. Deleting 10s: "If the second number is bigger than the first, the answer will be wrong [say had 15 - 12, if took off the 1 would have 5 - 12]." (Grade 5)

b. Addition: "Instead of subtracting, he added." (Grade 3)

4. R: Silly, J: Inappropriate

a. Combining digits: "It's confusing and it's not as fast as other ways." (Grade 5)

b. Counting up: "Counting up gives the wrong answer." (Grade 3)

c. Addition fact "She was adding instead of subtracting." (Grade 1)

5. R: Undecided, J: Appropriate

a. Addition fact: "She might not remember [addition fact] properly if she was doing the problem when she was hurried" (Grade 5)

b. Counting up, Gr3: "It's silly because it's slow but it's also good because it's accurate."

6. R: Undecided: J: Inappropriate

a. Addition fact: "It's good on 14 - 7 because he's using doubles but he's not using doubles on the second one (12 - 5) so it's silly." (Grade 3)

b. Counting down: "It's both because sometimes it [the answer] is right and sometimes it's wrong." (Grade 1)