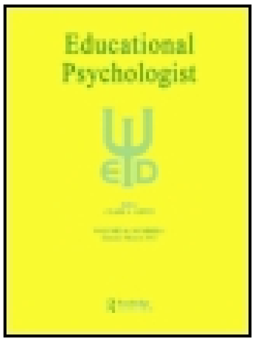

Educational Psychologist

ISSN: 0046-1520 (Print) 1532-6985 (Online) Journal homepage: http://www.tandfonline.com/loi/hedp20

How Software Technologies Can Improve Research on Learning and Bolster School Reform

Philip H. Winne

To cite this article: Philip H. Winne (2006) How Software Technologies Can Improve Research on Learning and Bolster School Reform, Educational Psychologist, 41:1, 5-17, DOI: 10.1207/s15326985ep4101_3

To link to this article: https://doi.org/10.1207/s15326985ep4101_3

EDUCATIONAL PSYCHOLOGIST, *41*(1), 5–17

How Software Technologies Can Improve Research on Learning and Bolster School Reform

Philip H. Winne
Faculty of Education
Simon Fraser University

Realizing the promise of software technologies in education requires thinking differently about how software simultaneously can serve research and contribute to learning. This article examines 3 axioms underlying contemporary educational psychology: Learners construct knowledge, learners are agents, and data include lots of randomness. By drawing out corollaries of these axioms, this research uncovers significant challenges researchers face in using classical forms of experimental research to build a basis for school reform and for testing school reforms using randomized field trials. This article describes a software system, gStudy, that is designed to address these challenges by gathering finer grained data that better support theorizing about the processes of learning and self-regulated learning. This research illustrates how this can be realized and suggests 10 ways that using software like gStudy can help pull up research by its bootstraps and bolster searches for what works.

Although it is clearly debatable, Patrick Suppes probably changed the course of North American education in the latter half of the 1960s. In an article in *Scientific American*, he made a bold conjecture: “One can predict that in a few more years, millions of school children will have access to what Philip of Macedon enjoyed as a royal prerogative: the personal services of a tutor as well-informed and responsive as Aristotle” (p. 206). Three years later, he and a colleague published a breakthrough report in *Science* demonstrating that sophisticated computer software capable of adapting to each learner had strong potential to enhance elementary students’ learning of arithmetic (Suppes & Morningstar, 1969).

Perhaps one of the most appealing ideas for using software technologies to revamp educational practices was proposed approximately a decade later by Seymour Papert. In his best-selling book, *Mindstorms: Children, Computers and Powerful Ideas* (1980), Papert advanced the notion that a computer could be “a *carrier* of cultural ’germs’ or ’seeds’ whose intellectual products will not need technological support once they take root in an actively growing mind” (p. 9). The seeds of which he spoke were sowed by inviting children to devise programs in the LOGO language so that a “turtle” would trace particular paths. Papert theorized that by programming, children could establish “an intimate contact with some of the deepest ideas from science, from mathematics, and from the art of intellectual model building” (p. 5).

In a different sector of the computing science community, another trail was blazed in efforts to develop intelligent tutoring systems. Whereas Suppes’s software applied algorithms based on the accumulating successes and failures of each learner as the basis for selecting which exercise to pose next, those working on intelligent tutoring systems sought to model a learner’s knowledge, its qualities and how these could be developed. In 1982, Derek Sleeman and John Seely Brown edited a landmark book on these topics. These systems sought, in very confined domains of knowledge, to realize the conjecture Suppes had made about a tutor.

Approximately a decade later, Barbara Means edited a book entitled *Technology and Education Reform* (1994) that surveyed the juncture of “two of the most significant trends in education”(p. xi) of the day, school reform and software technologies. The general conclusion of contributors to this book painted a checkerboard of some moderate successes laced with substantial challenges. As Means and Olson (1994) put it succinctly: “The complexity of the school experiences described here precludes such simple prescriptions as ’computers improve learning’ or ’give technology to teachers first.’ There is no right answer …” (p. 220).

More recently, Alan Lesgold (2000) sought to summarize this considerable activity. His description of the “ritual that has been under way for close to four decades” (p. 399) merits repeating:

Correspondence should be addressed to Philip H. Winne, Faculty of Education, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6. E-mail: winne@sfu.ca

> The ritual has two parts. In the first part, visionary authors examine new tools from the cognitive and information sciences, notice the substantial and ever-increasing penetration of the home market by computers, and propose revolutionary approaches to learning that take advantage of the new affordances of the information age. In the second part, reflective authors point out how little these visionary ideas have been implemented and show how the visions have failed to take account of the ways in which educational practice is taught, evaluated, and paid for. Both are mostly on target in their assertions. (p. 399)

There has been a great volume of research and much ink spilt in the scholarly and popular presses about whether and how software technologies might realize promises foreseen by its pioneers and advocates. Evaluations range widely, and the question is still open about whether and how software technologies can enhance learners' knowledge and contribute to national goals for education.

I begin my attack on these problems by exposing corollaries of three widely accepted "truths"—axioms—that I perceive to bedevil contemporary research in educational psychology, particularly research on learning and self-regulated learning (SRL) where learners modulate their approaches to learning based on monitoring how well their goals are met (Winne, 2005; Winne & Hadwin, 1998). Next, I briefly document that school reforms have not yet reached their goals, although I explicitly acknowledge that the purview of this article precludes developing a full warrant for this claim. Then, I summarize properties of a software system colleagues and I are developing. I try to describe why it has strong potential to ameliorate the problems I previously identified as plaguing today's research in educational psychology. Finally, I conjecture how software systems like the one I describe can bolster efforts to improve schools while, hopefully, escaping the ritual Lesgold (2000) described.

If lessons have been learned about posing models of learning and proposing school reforms, one is that critique is inherent. Critique should be invited. When engaged vigorously, skillfully, and constructively, critique advances the field toward understandings that are "the best that [can] be achieved by the methods of observation and analysis which are [presently] acceptable in scientific and scholarly communities" (Shils, 1983, p. 4). In hope of promoting this ethic, I fashion the next section as a formal argument, not because this format is valued for its own sake but because it may most readily invite and facilitate critique.

AXIOMS AND COROLLARIES

At least three axioms underlie a great deal of contemporary theorizing and research in contemporary educational psychology. These axioms merit explicit statement and some unfolding of their corollaries because their natural consequences have powerful implications for researching learning, especially SRL. These corollaries also have import for considerations about reforming learning in schools, universities, and lifelong education in general.

Axiom 1: Learners Construct Knowledge

Five facets can be identified when learners construct knowledge. Specifically, in a particular context for learning, learners:

1. use tools—cognitive operations and physical devices such as highlighting pens or software technologies …
2. to operate on raw materials—information in any of its diverse formats, such as text, diagrams, photographs and videos, charts, tables, mathematical expressions, and so on—that are available in the environment or retrieved from memory …
3. to construct a product, first as a form of information, that later can be retrieved from memory …
4. which is evaluated in a formative way or summatively with respect to …
5. standards of sociocultural kinds, such as being in accord with widely accepted "fact" or as being justifiable using accepted forms of argument such as modus ponens.

Learners are eminently capable of constructing knowledge without assistance. But one purpose of education and of designs for instruction is to enhance what learners can and will do on their own. Thus, software technologies (and other manifestations of instructional designs) that support learners in constructing knowledge should address each of the five features just noted. What follows from this partial analysis of the axiom that learners construct knowledge?

Corollaries of Axiom 1

Because of space considerations, I examine just two of the five facets involved in constructing knowledge, using tools and raw materials.

Tools. Learners use tools to operate on information. Examples of physical tools common to many academic tasks include the index of a book, the bolding of words in text, the underlining or coloring of hyperlinks on Web sites, glossaries, advance organizers, instructional objectives, and templates for taking notes about a subject being studied. Tools also can be cognitive. Examples are methods learners use to identify information to be labeled with highlighting, reading comprehension strategies, heuristics for writing valid and convincing arguments, and various study tactics and learning strategies such as Cornell note taking or survey, question, read, recite, and review (SQ3R) methods.

Teachers, developers of software learning objects, textbook authors, and so on provide tools for particular learners to use within a particular context. However, no tool's use is evident in an intrinsic sense. People must learn about tools and purpose(s) they can serve. (See Norman, 1988, for compelling arguments and entertaining illustrations regarding this claim.) Tools may empower their users but only under particular conditions. For example, novice Web surfers need to learn whether listing two words in a search engine's input field applies the Boolean operator *OR* as opposed to *AND*. In research on how tools may be involved when learners construct knowledge as they participate in instructional activities, tools make mischief when researchers try to interpret whether learning is affected by the tool because using a tool entails at least four considerations that, in concert, are rarely given due respect in research (Winne, 1982).

1. Learners must recognize or attend to occasions where a tool can be used. If learners miss the signal(s) an instructional developer embeds in content about when a learner should use a tool for learning, the tool cannot have any effect on learning. The learner is sometimes described as (or blamed for) a production deficiency (Flavell, 1970).
2. If the signal is observed, learners must map that context to one or a few tools that will "work." Choosing a tool that is ill-matched to the task can undermine knowledge construction, an instance of negative transfer. Choosing a tool different than intended by the content developer or researcher may generate different consequences than the content developer or researcher intended. Achievement and theory are both jeopardized in this instance. Achievement may suffer because the knowledge constructed does not correspond to objectives the content developer set out. Theorists misrepresent how knowledge was constructed because the tool they name as cause is not the tool the learner used.
3. If the preceding two conditions are satisfied, learners must be capable of using the tool skillfully. This means using the tool to produce particular results—those intended by the content developer—with good chances of success—that is, efficaciously and without suffering either the *mediation deficiency* (being unable to assemble bridging information between tools and to-be-learned information; Reese, 1962) or the *utilization deficiency* (omitting steps or failing to complete steps in a cognitive operatioin; Miller & Seier, 1994). Skillful use of a tool also means that the learner uses it without intruding too much on other appropriate ongoing cognitive and behavioral activities—that is, avoiding extraneous cognitive load while optimizing intrinsic and germane cognitive load (see Paas, Renkl, & Sweller, 2003).
4. If all three preceding requisites are satisfied, the learner must be motivated to spend effort required to use the tool skillfully, to monitor and control its use under conditions of the task (i.e., exercise volition), to accept the degrees and kinds of risks that attend the attempt to construct knowledge this way, and to acknowledge emotions (e.g., responsibility, pride, shame; see Hareli & Weiner, 2002; Weiner, 1986) that likely arise in relation to consequences that materialize within mind and in the external world.

Each of these conditions individually and all four together have to be documented when researchers develop an account that a particular tool causes or helps learners to construct knowledge better than they might have without it because by using that tool, learners engage in particular cognitive activities (e.g., see Pressley & Harris, in press; Winne, 1983b, 1997, 2005).

Raw materials. Notwithstanding arguments that some "bits" of knowledge or some forms of knowledge may be innate (e.g., Pinker, 2002), there is overwhelming empirical evidence that raw material, in the forms of a learner's prior knowledge and other kinds of information in long-term memory, exerts powerful influences on whether and how new knowledge is constructed, and whether that constructed knowledge is correct or valid. (For a succinct summary, see Alexander, 2006, pp. 72–73.) Two examples illustrate that raw materials affect how knowledge is constructed.

After assessing junior college students' beliefs about facets of what knowledge is, how knowledge is constructed, and what influences processes used in constructing knowledge, Schommer (1990, Experiment 2) asked students to read a passage that introduced a controversy. The passages were incomplete; they lacked a concluding paragraph resolving the controversy. Students wrote that concluding paragraph. Schommer's regression analyses showed that the more students believed knowledge is certain and the more they believed processes of constructing knowledge should yield results quickly or would not succeed, the simpler the paragraphs they wrote.

Van der Meij (1990) assessed fifth-graders' knowledge of vocabulary. Then the students were invited to use questions of different types, about which they had received prior instruction, as tools for learning the meaning of challenging new words. Compared with students with more vocabulary knowledge, those with lower prior scores on vocabulary asked fewer questions. Of the questions asked by students with lower vocabulary knowledge, most returned general information not useful for discovering the meaning of words rather than being questions that could return information specifically helpful to learning the new vocabulary word.

Schommer's study can be interpreted as meaning that learners' beliefs influence products they create using whatever tools they choose for constructing knowledge in an essay task. Van der Meij's study supports an interpretation that learners' prior knowledge correlates with their choices of tools for constructing knowledge by asking questions. I claim without proof that a fuller review of the research literature about prior knowledge would demonstrate its strong influence on the form(s) of information a product takes, the evaluations a learner generates and seeks, and the particular standards a learner uses to evaluate products.

Axiom 2: Learners Are Agents

Agency is an ontological concept that refers to, among other things, the capability to exercise choice in reference to preferences. Learner-agents (hereinafter, I use *learners* in this way) who are successful in complex environments such as schools and workplaces are agents. They act with purpose. That is, they anticipate outcome(s) of engaging in activities and of engaging in a particular activity in one or another way, and they make plans to achieve outcomes they favor by using methods they prefer. Even trial-and-error behavior is a choice to investigate what happens by exercising diverse options.

With respect to perhaps the most frequent kind of activity in education, namely, activities intended to help learners construct knowledge, learners make choices about each of the five facets I previously listed as constituents in constructing knowledge. They choose tools to use in fashioning products. For example, learners may choose simple learning tactics that yield shallow knowledge or more effortful and complex tactics that are recommended by research (Winne & Jamieson-Noel, 2003). Learners do not access raw materials willy-nilly or without bounds. They do so selectively, for example, by choosing to attend to surface features of statistics word problems or more meaningful structural features (Quilici & Mayer, 1996). To use a statistical metaphor, they sample. Learners decide which form to give to information in products they construct, such as choosing whether goals they develop at the start of study sessions describe specifics about what will be learned versus time to spend studying (Morgan, 1985). As they proceed to construct knowledge, learners choose whether and when to monitor their comprehension of the information they are studying, immediately or after some delay (e.g., Theide, Anderson, & Therriault, 2003). And, they decide which standards to apply to their products, for example, whether products are good enough (satisificing; Simon, 1953) or optimized. They also decide which standards describe optimal operations to use in creating products (e.g., Rabinowitz, Freeman, & Cohen, 1992).

Educators and instructional designers who recognize learners as agents, including those who design software technologies to help learners construct knowledge, face the practical challenge of guiding learners to make successful choices without running afoul of overly restricting choice. Failing at the former means that instruction is no more effective than what learners can do on their own. Failing at the latter is blameworthy for "stamping out spirit" and inhibiting each learner's opportunity to develop as an individual within bounds the learner and her or his community deems acceptable.

Corollaries of Axiom 2

Amid celebrations of choice in learning, it must be acknowledged that choice can have drawbacks. (For an interesting and often amusing account of the perils of choices and choosing optimization as a standard, see Schwartz, 2004.) Some choices learners (and educators) make are less beneficial than others. Warranting this claim is easy. Thousands of experiments on constructing knowledge have been published in research journals. The overwhelming majority demonstrate that learners left to express agency without guidance—operationalized as the "control" group versus a treatment group—store information in less retrievable forms of knowledge, retrieve less knowledge often in less sophisticated forms, solve fewer problems, and construct knowledge that is fragile when they try to apply it in new situations. Learners in treatment groups, where guidance about how to construct knowledge is provided, do relatively better. This is good news. There are many methods for guiding an average learner to construct knowledge better than she or he can if left alone. However, these descriptions of what works beg for explanations about how and why they work, that is, for theories. There are two significant challenges to theorizing validly about how and why learners use tools provided by an instructional design to benefit knowledge construction.

What tools do learners choose and how are tools used? Because learners can and do make choices about tools used in constructing knowledge, researchers are obliged to gather data about which choices each learner makes in a treatment condition and, if possible, to identify learners' bases for these choices as well as to model mechanisms they used to make choices. Merely observing a benefit to achievement, a product measured after multiple unknown cognitive operations are completed, and that a treatment was implemented as intended, as recommended by D. T. Campbell and Stanley (1963) and Cook and Campbell (1979), does not generate a logically sufficient basis for an account about how knowledge was constructed (see Borsboom, Mellenbergh, & van Heerden, 2003, 2004). Researchers who strive to develop "accounts of how factors of instructional design 'cause' learning have to 'get inside' the time period between the independent variable and the dependent variable" (Winne, in press) by collecting traces of learners' using tools (Winne, 1982, 1983b; Winne, Jamieson-Noel, & Muis, 2002; Winne & Perry, 2000).

One approach to meeting this requirement has been to ask learners about these matters while they engage in learning. These are concurrent think-aloud protocols. Another method is to ask learners before or after they participate in an experiment to describe how they study in their own words or using self-report inventories. Both tacks make requirements that likely compromise their use.

In the case of concurrent think-aloud methods, learners are invited to attend to and report on contents of their thoughts. By making these targets of attention and because associated information in long-term memory is activated automatically, learners may explicitly consider one or more of the five constituents of constructing knowledge that otherwise would not have been attended to. If this happens, thinking aloud induces metacognitive monitoring into the ongoing learning activity that otherwise would not have occurred. Among agents, instances of metacognitive monitoring establish a pivot point where they may change one or more features of how they construct knowledge. If metacognitive control is exercised, think-aloud data are confounded because they trigger SRL that would not have occurred otherwise.

Ericsson and Simon (1993) were concerned about this issue. They distinguished three levels of verbalization in thinking aloud. Levels 1, 2, and 3 were, respectively: "simply the vocalization of covert articulatory or oral encodings," "explication of the thought content [in ways that] do not bring new information into the focus of the subject's attention, "and "verbalization [that] requires the subject to explain his thought processes or thoughts" (p. 79). Ericsson and Simon's review of research led them to the interpretation that only if participants in experiments were instructed to think aloud in a way conforming to Level 1 or 2, "the studies gave no evidence that verbalization changes the course or structure of the thought processes" (p. 106). Although I trust this interpretation relative to the studies they reviewed, it is important to notice that almost all these studies were not at all like learning in school. Participants engaged in tasks such as judging whether to parole a criminal by selecting evidence, choosing coffee makers or selecting other consumer items, narrating imagery relating to modeling assertive behavior, solving various experimental problems (e.g., Duncker's candle problem, anagrams, block problems, insight problems). Only one study was plausibly close to situations learners face in schools, solving algebra problems (Flaherty, 1974), but the participants were not asked to learn anything while solving the problems. On this basis, whether thinking aloud alters how learners construct knowledge in school-like learning activities is indeterminate.

In the case of inviting free descriptions or using self-report surveys to generate data about how learners construct knowledge, research raises concern about these data as bases for theorizing about facets of knowledge construction (Winne, Jamieson-Noel, & Muis, 2002; Winne & Perry, 2000). Self-report inventories ask learners to report about facets of constructing knowledge in relation to very large-grain conditions such as "this course" (e.g., Motivated Strategies for Learning Questionnaire; Pintrich, Smith, Garcia, & McKeachie, 1991) or unspecified conditions "about learning and study practices" (Learning and Study Strategies Inventory; Weinstein, Schulte, & Palmer, 1987). However, Hadwin, Winne, Stockley, Nesbit, and Woszczyna (2001) found that learners within a single undergraduate course self-report differently about facets of SRL as a function of the task immediately before them—say, studying for an examination versus writing a paper. Wolters and Pintrich (1998) reported that the same learners' self-reports about facets of constructing knowledge differed when the subject matter of their class varied from mathematics to English to social studies. These studies substantiate that learners see themselves as agents who make choices about how to construct knowledge. The studies also show that conditions affect choices. This is important because, although models of SRL such as Hadwin's and mine (Winne, 2001; Winne & Hadwin, 1998) acknowledge conditions as a key category of variables, principles rarely do. The question remains open about how to manage generalizations over conditions in giving accounts of knowledge construction.

Using self-reported data to reflect how learners construct knowledge is further challenged when the question is raised, How accurately do learners report on facets that characterize constructing knowledge? In other words, how calibrated are learners' self-reports in relation to objective data? Maki's (1998) review of 25 studies yielded an average correlation of .27 between learners' prediction of comprehension and performance. More recently, Dunning, Heath, and Suls (2005) reviewed studies of people's calibration in health, education, and the workplace. They concluded, "People are not very good at assessing their comprehension of written materials" (p. 87) and, in general,

> The views people have of themselves are often flawed. The correlation between those views and their objective behavior is often meager to modest, and people often claim to have valuable skills and desirable attributes to a degree that they do not. (p. 98)

Not only are learners inaccurate but they are biased. Learners who report greater confidence in their comprehension tend to score lower, whereas learners with less confidence tend to score higher than predicted. The same inaccuracy and bias were observed when learners self-reported about operations they carried out during learning (Winne & Jamieson-Noel, 2002; Winne, Jamieson-Noel, & Muis, 2005).

In short, self-report methods—concurrent think-aloud protocols, free descriptions, and questionnaires—have a particular and restricted use in building theories to account for how learners construct knowledge. What learners self-report is the information they use to regulate learning. It is critical to consider that information in developing accounts about how learners construct knowledge because it inherently affects how learners self-regulate learning. However, when learners report what they know—the raw material that is prior knowledge—or what they do to construct knowledge—the tools and corresponding cognitive operations they use—theorists cannot take these reports at face value. Data complementing learners' self-reports are required to reveal reliably how learners actually construct knowledge.

Why do learners choose a tool? One of the reasons learners may not follow the guidance teachers and instructional designers offer is lack of motivation. Motivation concerns three issues in relation to constructing knowledge (Winne & Marx, 1989). First, when there is a choice of tools to use in constructing knowledge, which tool(s) does a learner choose? For, example, to integrate new material in a textbook chapter, does a learner choose massed maintenance rehearsal or concept mapping? Second, having chosen a tool, what qualities characterize use of the tool? Is rehearsing or concept mapping used "intensely" or "casually," with great attention to detail or at a surface level in representing information? Third, how dedicated is the learner to the chosen tool? Does the learner persist in using a tool to construct knowledge or shift quickly to other tools? In this arena, learners' self-reports are important data. But, these data must be interpreted carefully with respect to calibration and factors that affect self-reports (see Menon & Yorkston, 2000; Tourangeau, Rips, & Rasinski, 2000) and should be coordinated with data that trace what learners do (Winne, Jamieson-Noel, & Muis, 2002).

Relative to studies of learners' perceptions about how well tools work, there have been few studies examining why learners choose tools. One was Rabinowitz et al.'s (1992). They investigated whether undergraduates encouraged to use a common study tool—clustering—in a first session would use it in a second session depending on whether their first experience using clustering was relatively easy or effortful. Using both self-reports and measures of clustering, these researchers found learners were likely to abandon clustering if it had proved difficult to use at first, even though they recognized its continuing usefulness in the second session. Presumably, effort was modulated. Speculatively, learners in this experiment generated a judgment of utility: What are my returns for different investments of effort in constructing knowledge? In this study, whatever tool(s) learners substituted for clustering served them as well as clustering in terms of amount recalled.

Rabinowitz et al.'s (1992) study implies three important needs for theorizing about how learners construct knowledge when teachers and instructional designers try to influence their choices. First, effort is likely one of the factors that affects whether learners use a tool. Second, learners can shift tools inside what an experimenter considers a single, homogeneous treatment activity. Third, researchers may not be able to use scores of products (recall in Rabinowitz et al.'s study) to infer reliably which tools learners use. To do that, data are required that directly trace which tools learners used (Winne, 1982; Winne & Jamieson-Noel, 2002).

Axiom 3: There Is Randomness in Data, but a Central Tendency Prevails

A nearly ubiquitous assumption in the research on educational psychology is that indicators—scores—of most variables relating to learning, achievement tests, attitudes, motivation, and the like are distributed as a normal curve. Under this assumption, the mean and the variance of scores have a particular interpretation. It is that in a population, scores differ from the mean only because random factors cause these deviations. There are two possible accounts of this randomness according to Bennett (1998).

One account is that randomness is ignorance about causal factors. If a random factor was known and measured or if it was controlled by setting it to a single value, a more precise specification for describing the data could be forged. This specification would divide the original heterogeneous population—heterogeneous because scores vary around the mean due to previously unknown but knowable factors—into two or more populations. In each new population, because scores have the same value on this previously unspecified factor, variance due to that factor is eliminated. When researchers model the data as "due to" the factor labeled by the mean but factors not specified in defining the population cause deviations from the mean, the model is technically labeled as misspecified. Under a view that variance can be reduced by taking account of a previously unaccounted-for variable, it is perhaps more accurate to say the model is underspecified.[1]

The second account is that randomness arises due to "impenetrable uncertainty." In this case, it is impossible, no matter how hard researchers try, to remove variance exhibited by scores because no factor(s) can be identified and possibly controlled to reduce variance.

Corollaries of Axiom 3

I assume educational psychologists believe impenetrable uncertainty is a significantly less prevalent cause of variance in scores. Thus, it is critical to model data with a minimum of misspecification or underspecification. This raises big challenges for two separate cases.

The case of one score. Stanley (1971) tabled approximately 50 factors (some of his lists end with "etc.") that reduce the reliability of a single score that indicates something about a single learner responding to a single test item (task) administered at a single time and scored once. To these named factors, I add the variability attributable to the five factors affecting knowledge construction and the variability latent in a learner's choice to exercise agency in a particular way selected from an array of choices. The ubiquitous approach to addressing this issue is to aggregate over scores, for example, to represent recall of studied materials by summing a count of propositions in a learner's essay that can be located in materials the learner studied. To do this requires several assumptions, two of which I spotlight.

First, responding to one item must not affect the response to another item. This is the assumption of local independence. This assumption is dubious whenever human memory is involved. Studies of priming in tasks like learning lists or using schemas to recall information in essays show the opposite of local independence. So does research on how people answer self-report items, where a violation of the assumption of local independence is called the *framing effect*. Second, the scores added up must be based on homogeneous items or tasks. This assumption is moot unless one has a valid description of the operations that generate the response plus data that can be used to show the operations were similar (Borsboom, Mellenbergh, & van Heerden, 2004).

[1]I thank John Nesbit for this suggestion.

The case of scores in a group. In this instance, variance around the mean is attributed to unreliability in the scores per se plus factors that affect the aggregate score. Consider research on the effects of early childhood home interventions. This topic was recently reviewed by Bakermans-Kranenburg, van Ijzendoorn, and Bradley (2005). They concluded six factors moderate effectiveness of an early childhood intervention in the home: mother's home country (United States, other), percentage of adolescent mothers (>70%, 30%–70%, <30%), child's age at start of intervention (prenatal, 0–6 months, >6 months), location of intervention (home, center), number of sessions (0–4, 5–16, >17), and cognitive focus (yes, no). In light of this knowledge, an experiment on the effectiveness of a home intervention for young children should take these factors into account. Empirical investigations, including the oft-regarded "gold standard" of randomized field trials (see Coalition for Evidence-Based Policy, 2002), thereby become extraordinarily hard to do.

Considering the six factors Bakermans-Kranenburg et al. (2005) documented, one randomized field trial in which randomness is mathematically able to reduce bias in estimates of effects must be designed minimally as a 2 (intervention, comparison) × 2 × 3 × 3 × 2 × 3 × 2 factorial experiment. When there is insufficient basis for estimating statistical power to detect higher order interaction effects, researchers might retreat to an "old saw" and populate each of these 432 cells with 30 children. This experiment requires $N = 12{,}960$. Omitting any one of the research-validated factors raises the specter of model underspecification and increases bias in estimating whether an effect is detected. Using an insufficient sample size jeopardizes the opportunity to detect real effects. Moreover, if the world really is stochastic (involving randomness), this one large, golden experiment should be replicated many times before it is suitably safe to draw strong inferences about whether this particular intervention works. Somewhere in the course of this series of experiments, it is important to reach a judgment about whether the size of the effect merits spending more resources to shore up evidence about it. As I suggest later, this judgment is, likely, "Don't."

To the extent that effects reported in other areas of research—for example, how to teach reading, boost problem solving in mathematics and science, reduce bullying, or enhance self-concept—are moderated by five or six variables, very few well-controlled programs of experimental research have the resources to reliably identify which factors afford learners' constructions of knowledge(s) and how much each factor matters (e.g., as measured by its effect size).

SUMMARY

Because cognitive operations are proximal causes in constructing knowledge, experimental designs should verify as thoroughly as possible that what learners do during experiments corresponds to the operations a theorist hypothesizes to take place. Experimental designs are then fashioned to manipulate; control; and, to the extent possible, record data that describe each of the other facets involved in constructing knowledge. I observe few laboratory and small-scale experiments collect all the requisite kinds of data. To my knowledge, no large-scale randomized field trials do.

By acknowledging that learners are agents, educational psychologists cannot assume all learners make identical choices "inside" the package of factors labeled as a *treatment* or an *instructional design*. Researchers are obliged to gather data inside a treatment or instructional experience that can document choices learners make about which operations they apply to which raw material, evaluations they receive about products created, and standards learners use to regulate how their engagement unfolds over time. In other words, under the assumption that learners are agents, warrants for "treatment implementation validity" necessarily entail validating how the learner exercises agency as well as how the experimenter or teacher attempts to guide learners' exercise of agency. Again, I observe very few laboratory or small-scale field experiments that do this. To my knowledge, no large scale randomized field trials do.

The take-home message is straightforward. Educational psychologists and learners both need data about "what's going on." The educational psychologist strains accounts of constructing knowledge without these data. The learner is in a poor position to self-regulate learning effectively without these data. Whenever such data fall beyond the purview of researchers or learners, neither is on solid ground when trying to infer what works because the what—the cognitive operations that construct knowledge—is speculative rather than recorded as traces.

IMPLICATIONS FOR SCHOOL REFORMS INTENDED TO ENHANCE LEARNING

As Baker and O'Neil (2002) observed, despite decades of effort, learners in the United States still do not meet goals for academic achievement when they are gauged in relation to national standards (J. R. Campbell, Hombo, & Mazzeo, 2000) or when judged normatively by their performance in international studies of achievement (Third International Mathematics and Science Study, 1998). Thus, calls for reforming schools to be more effective, more efficient, or less costly are justified. Similar concerns are raised in other countries such as Canada.

In some instances, experiments about school reforms are horse races: Is this better than that? Horse-race studies can answer plain questions about effectiveness, efficiency, and cost. However, finding such answers has proven to be difficult. When none of the horses fare well enough, everyone involved in education looks for a better horse. But what qualities should be sought? What will work better than the horse we're currently riding?

If one has really no idea what will work, statistically speaking, finding an answer to what works will likely require many trials plus stamina to suffer many errors. Even if an early trial is successful, this one experiment's results should be reinvestigated (replicated) because there is no assurance the effect will reappear in a successive trial. This is clear by understanding that the probability of a Type I error such as the (sadly) revered $p < .05$ describes the probability a sample is mistakenly ascribed as belonging to the population where the treatment doesn't work (well enough). The Type I error rate does not describe whether the sample belongs to any other particular population, including one in which the treatment works as intended. That crucial matter is judged by other standards that lie outside statistical measures of errors that attend deciding in which population a sample may belong.

If one has an idea about what works, one has a hypothesis. To have some credibility, the hypothesis should rest on the top of a pyramid of various kinds of "supporting" research (Levin, 2004). But just how supportive is the support? There at least are two kinds of replies.

One reply concerns whether prior empirical studies and theorizing that arises from them justifies a particular hypothesis relative to competing hypotheses. In other words, is the hypothesis at hand—a model founded on the supporting research—not misspecified (or underspecified)? On the basis of the argument I developed earlier, I predict the hypothesis probably is misspecified. To the degree that my prediction holds, the "what" in what works is indefinite. It is not clear enough what happened when an effect is observed that learners did construct knowledge. Manipulating features in the world to re-create those operations is chancy and, in this sense, support for the hypothesis is fragile.

Assume, for sake of argument, that the foregoing analysis is flawed and it is clear enough what works. A second reply to the question of how much support for the hypothesis is provided by underlying research addresses whether treatment effects are "big enough." There is widespread agreement that this should be gauged by effect size statistics. Slightly more than a decade ago, Lipsey and Wilson (1993) surveyed a wide variety of meta-analyses, many of which examined the effects of treatments on targets for school reform, such as early intervention programs for learners who were disadvantaged or had a disability, remedial language programs and bilingual instruction, instruction in science and math, and class size. The mean effect size of 181 meta-analyses they identified in education was $d = 0.44$ ($SD = 0.29$), or, expressed as a correlation, $r = .22$ (Rosenthal, 1994, Equations 16–24). Suppose a woman taking the Scholastic Aptitude Test–Math in 2005 would have scored at the mean of 508, except that she participated in a program to improve her score and it caused an effect of this magnitude ($d = 0.44$). Her score probably would rise to 553, a gain of 49 points. The College Board (2005) advises test takers, "Most of the time, your score would fall in a range about 30 to 40 points above or below your true ability." Thus, probabilistically speaking, this is a genuine gain. The correlation also can be interpreted as the intervention accounting for 5% of the variance of the outcome variable.

Although effects near the size of $d = 0.44$ may well represent improvements, I do not consider them "big enough" to achieve real reform in schools. Why might effect sizes reported about experiments be this small? I suggest five reasons. First, treatments are very brief. Very rarely do treatments span time equal to that which a learner spends on a chapter in a textbook or a unit of study in schools. If learners use tools to construct knowledge, using tools skillfully probably requires practice with feedback over time. Brief treatments do not allow for this. Second, quantitative analyses in many studies lack statistical power. This occurs in part because samples are smallish and in part because control over factors that introduce "random" variance is weak or difficult to achieve in the real world. Third, even if experimental control is effective and treatments are powerful, agency still introduces variance that is not yet accounted for in the experimental design or analyses of data. Fourth, because studies focus on one or a very limited set of treatments, it is unknown whether mixing treatments, as often happens in interventions proposed for school reform, generates interactions that depress or enhance achievement. Fifth, as shown in Cronbach and Snow's (1977) seminal review of aptitude–treatment interaction research, tools that help some learners construct knowledge sometimes make it more difficult for other learners to construct knowledge.

The upshots of these propositions are as follows:

1. Previously credible accounts of why a treatment works to help learners construct knowledge—theories—are challenged because (among several other reasons) data almost always are missing that are necessary to help validate such accounts.
2. Even if a strong account was available about how a treatment may help a learner construct knowledge, there would remain a challenge in forecasting accurately whether any particular learner will, in exercising agency, choose to construct knowledge as hypothesized.
3. Candidates for treatments that may reach some goals of school reforms are not faring well enough, perhaps in part because of the preceding two issues.

SOFTWARE AS A TOOL FOR DOING SCIENCE ABOUT CONSTRUCTING KNOWLEDGE

In this section, I very briefly describe a software system called gStudy that colleagues and I (Winne, Hadwin, Nesbit, Kumar, & Beaudoin, 2005) are developing in the Learning Kit Project.[2] gStudy is a tool for pursuing research on SRL

[2]For information about gStudy and the Learning Kit Project, visit http://www.learningkit.sfu.ca/

that learners apply in the service of constructing knowledge. Tools gStudy provides to learners and data it collects on the fly about tools' uses can lessen the aforementioned challenges to research, although I hasten to add there are caveats to be considered. I note these later.

gStudy and Its Tools

Four principles underlie the design philosophy of gStudy. First, learners should have access to a wide variety of tools that afford them options for exercising agency as they go about their work to construct knowledge. Second, these tools should be appropriate for frequent use in diverse subject matters, across varied instructional contexts, and over a very long term (i.e., years). Third, to provide raw materials each learner needs to evaluate which methods for constructing knowledge really "work," abundant data should be collected about more than just the amount of knowledge that is constructed. Data that traces how learners go about constructing knowledge are essential for them (and researchers) to discriminate which study tactics and learning strategies work. Fourth, techniques for analyzing trace data in concert with other kinds of data, primarily data about achievement but also about time spent, should be available to learners (and researchers). Our research group's hypothesis is that, if all four principles can be realized, software systems like gStudy can help each learner to pull up his or her learning by its bootstraps because the software supports constant, intense, long-term, programmatic, learner-driven research on how to learn better. In other words, one of the aims is to help learners develop a personal program of design experiments that are progressively more and more effective. So, how might gStudy accomplish these goals?

Learners use gStudy to engage with information in software-based learning kits. The information can be formatted as text, diagrams, photos, charts, tables, audio and video clips, and so forth—that is, the multimedia information formats found in libraries and on the Internet. Learning kits can be about almost any topic. As learners study in a learning kit, they can use gStudy's tools to create "information objects" and forge links between information objects. Kinds of information objects include notes, glossary entries, hierarchical (tree-structured) indexes, hierarchical labels applied to other information objects, entries in a table of contents, nodes and arcs and sets of nodes in concept maps, search queries, HTML documents, spreadsheet documents, documents that record chats learners generate in conversation with peers and with gStudy's software coach, and archives of Web sites.

For example, to make a note about content in a learning kit, a learner first selects information presented in a Web browser by clicking then dragging the cursor across the information. The selection can be a string of text, a rectangular region in a diagram or chart, or a frame in a video or audio clip. (Using another method, an entire information object can be selected.) The learner then uses a keystroke combination to pop up a contextual menu. It offers several options, one of which is to create a link between the selection and a new note. Selecting this option opens another window where the learner can choose a template for this note. Note templates are schemas for annotating content that an instructional designer (researcher or teacher) provided because the instructional designer hypothesized the templates would be useful to learners as they studied content in the learning kit. For example, a debate template we designed includes seven fields that students can fill in to annotate contentious information: issue, Position A, evidence for Position A, Position B, evidence for Position B, my position, justification. Other templates can be designed—for example, summary, critical detail, apply to practice, self-test, and so on. Learners can design their own templates for notes.

The methods for creating a note as guided by a template constitute a tool. Whenever a learner uses this tool to create a note, gStudy traces in very fine-grained detail all the events just described in creating a note: which content was selected and when, which option was chosen from the contextual menu, which note template the learner chose to use for annotating the selected content, which fields of the note template the learner filled in and what that information was, and when the learner closed the window used for making a note and returned to studying the main material. All these data are traces of the learner's engagements with raw materials.

On the basis of singular trace data and patterns across trace data, inferences are made about kinds of cognitive and motivational events the learner has chosen to carry out. Regarding the note just describe, the learner first metacognitively monitored content in the learning kit to determine that a particular element in it merited annotation. This is traced when the learner selects a portion of the information and chooses the option "link to new note." Second, the learner metacognitively monitored how the information selected could be classified. This is traced by the learner's selection of one among several templates available for recording notes. Third, if the learner fills in the slots of the schema that refer to Position B in the debate note's template, this traces that the learner was able to identify a counterargument in the learning kit's content, if it is there. The instructional designer knows if it is there and can tag it in a way that gStudy also "knows" it is there. Otherwise, the learner recalled or constructed a counterargument based on prior knowledge.

Suppose the learner next creates a link between this note and a term in the glossary using one of the several methods gStudy provides. As a result of building this link, one information object can "retrieve" the other with a simple click of the mouse. The trace of this event can be interpreted to indicate the learner recalled that a domain-specific term in the glossary was relevant to the debate recorded in the note. As well, the learner either predicted it would be helpful to be able to review that term when reexamining the note at a future time or, vice versa, that the note about the debate was a good illustration of the glossary term that might be handy to retrieve later. If we observe the learner to make these kinds of links repeatedly, this is evidence of planning.

Advantages of Trace Data in Researching How Learners Construct Knowledge

Trace data like those gStudy logs when learners work in learning kits create a time-stamped record of events. These records support grounded interpretations about how a learner constructs knowledge. Trace data reflect what learners do in ways that help researchers avoid the previously described four kinds of mischief that plague many studies intended to test whether a tool helps learners construct knowledge. Trace data also reveal more accurately, although not perfectly, whether, when, and how learners access prior knowledge. Trace data track a learner's choices, the ways in which they express agency. With sufficient samples of trace data, it may be possible to identify what standards learners use to make these decisions.

When trace data are complemented by the other forms of data, researchers can paint a much fuller and more detailed picture of each learner's actual engagement in learning. Researchers can significantly reduce under- or misspecification in their models with respect to constructs like metacognitive monitoring, elaborating, searching for information, and recall of prior knowledge in the midst learning. These events are theorized to be causes of variance in achievement. Moreover, rather than be forced to assume a priori that variance around a mean is random, trace data allow researchers to measure in situ many key sources of that variance. This provides a basis for blocking participants (e.g., see Kirk, 1982) a posteriori. Of importance, it allows researchers to avoid risky and very likely invalid interpretations growing out of statistical methods by which the variance of some measures is partialed from others. (For different accounts of how statistical control of variance can mislead theorists, see Borsboom, Mellenbergh, & van Heerden, 2003, regarding the case of latent variable models, and Winne, 1983a, concerning the case of multiple regression.)

HOW SOFTWARE TECHNOLOGIES CAN BOLSTER SCHOOL REFORMS FOR ENHANCING LEARNING

At the outset, I first must note four important caveats bearing on the possibility that software systems like gStudy can benefit school reform. One is that learners who use gStudy need to be readers. The second is that using gStudy requires each learner to have easy and regular access to a computer so learning kits are available at will. Third, although topics presented in learning kits can range across almost all the curriculum, some particular topics (e.g., physical skills) may be difficult to represent well in learning kits. Finally, not every kind of educational activity can take the form of a learning kit. Each qualification limits the range of using software systems such as gStudy in research and in school reforms.

In circumstances where none of these caveats apply, and if the emphasis of a reform is to help learners construct knowledge, I suggest that software systems like gStudy can make 10 important contributions to reforming schools for the better. (Hereinafter, for ease of expression, I use the label *gStudy* to refer to similar kinds of software systems.)

- Using gStudy and learning kits as a medium for research on how learners construct knowledge can provide more data and more informative data about how this process happens. This will enhance the base of the pyramid on which rest proposals for reforming some school practices. Consequently, it should be possible to choose better horses to race in randomized field trials that investigate the benefits of school reforms that aim to improve achievement by improving learning.
- Because gStudy's learning kits can present material in almost any curriculum area, reforms that concern relatively general skills for constructing knowledge can be investigated across subject matters as within-subject designs. This significantly enhances learners' opportunities to practice and generalize such skills. Simultaneously, it strengthens researchers' opportunities to identify factors that increase or limit generalization.
- Because gStudy is easy to distribute over the Internet, enhancements to gStudy that articulate well with a school reform can be made available quickly and with minimal or no cost relative to changing non-electronic resources such as books or replacing equipment. This makes it cheaper to investigate plausible reforms (except that computers are expensive and necessary in this research).
- Data generated as learners use gStudy to study learning kits can be analyzed, aggregated within and across episodes and learners, and reported back to all parties engaged in school reform with a very short delay. In the same way as my stock portfolio can be updated every 20 min, learners, teachers, and researchers can have data upon which to make on-the-spot adaptations. Design experiments can therefore evolve more rapidly to converge on what works.
- Because data are preserved electronically and are almost instantly and very easily searched and sorted, recovering precise information about what worked is made very much easier. Equally important, comparing what worked better is facilitated.
- gStudy data provides a bridge to join data about how learners study and learn solo and how they collaborate in these activities. This affords opportunities for learners to develop skills in evaluating their own work by practicing skills in evaluating others' work, and it helps researchers bring together theories about learning alone and together.
- Trace and other data gathered by gStudy provide raw materials for researchers to investigate how learners construct knowledge. Each learner can use the same data to pur-

sue a personal program of research on how to succeed in school by self-regulating learning. That is, as learners participate in new designs for learning, they double as researchers. Because learners' perceptions play a critical role in how they exercise agency, these data are important to them and to researchers.

- Because software can manage and analyze volumes of data that overwhelm people, and because volumes of data can be gathered, judging the effects of school reform can be somewhat liberated from perils of statistical inference because samples may include most of the population.
- Statistically testing whether a group mean on a final measurement moves from a lower to a higher location (inferred by comparing the treatment group mean to a comparison group mean) fails 49.99% of cases who fall below the group mean and ignores how the mean moved (or did not). Instead, using data that gStudy collects, it will be possible to ask different questions. How much has each learner accomplished? Is each learner's change for the better? Is every learner's accomplishment acceptable?
- Because gStudy makes available trace data about how learners construct knowledge, empirical investigations, including randomized field trials of school reforms about how to enhance learning, can be more than just horse-race studies. More studies can contribute to advancing theories that are sources for ideas about how to enhance support for learners who work to construct knowledge. That is, every study can simultaneously investigate what works, what does not work, and why.

In summary, except for the caveats previously noted, software systems like gStudy can bolster school reform because they satisfy a criterion proposed by Lowther, Gassoppo-Moyo, and Morrison (1998). That criterion is that school reform depends on learners' having opportunity to use software tools that genuinely support learning from instruction. To this criterion, I add that software tools must genuinely support learners, as well as researchers, in their respective tasks of researching how they learn (called *self-regulated learning* when learners research their learning). Systems such as gStudy can satisfy these criteria and can help learners and researchers alike to pull up their research by its own bootstraps (Winne, 1992). Rather than straining to close the loop between experimental research and school reforms, I propose using software systems like gStudy to fuse these practices into one coincident activity.

ACKNOWLEDGMENTS

Support for this work was provided by grants to Philip H. Winne from the Social Sciences and Humanities Research Council of Canada (410–2002–1787 and 512–2003–1012), the Canada Research Chair program, and Simon Fraser University. This article is partially based on my Presidential Address to Division 15 presented on August 9, 2003.

Thanks to Lyn Corno, John Nesbit, and Nancy Perry for their excellent reviews of drafts.

I dedicate this article to the memory of an exemplary scholar and my treasured friend, Paul Pintrich.

REFERENCES

Alexander, P. A. (2006). *Psychology in learning and instruction.* Upper Saddle River, NJ: Pearson Education.

Baker, E. L., & O'Neil, H. F. (2002). Measuring problem solving in computer environments: Current and future states. *Computers in Human Behavior*, *18*, 609–622.

Bakermans-Kranenburg, M. J., van Ijzendoorn, M. H., & Bradley, R. H. (2005). Those who have, receive: The Matthew effect in early childhood intervention in the home environment. *Review of Educational Research*, *75*, 1–26.

Bennett, D. J. (1998). *Randomness.* Cambridge, MA: Harvard University Press.

Borsbom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203–219.

Borsbom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Campbell, D. T, & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.

Campbell, J. R., Hombo, C. M., & Mazzeo, J. (2000). *NAEP 1999 trends in academic progress: Three decades of student performance* (Report No. NCES 2000–469). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Coalition for Evidence-Based Policy. (2002). *Bringing evidence-driven progress to education: A recommended strategy for the U.S. Department of Education.* Washington, DC: The Council for Excellence in Government.

College Board. (2005). *Score range.* Retrieved October 1, 2005, from College Board SAT Web site: http://www.collegeboard.com/student/testing/sat/scores/understanding/scorerange.html

Cook, T. J., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitude and instructional method.* New York: Irvington.

Dunning, D., Heath, C., & Suls, J. M. (2005). Flawed self-assessment. Implications for health, education, and the workplace [Special issue]. *Psychological Science in the Public Interest*, *5*(3).

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

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

Flavell, J. H. (1970). Developmental studies of mediated memory. In H. W. Reese & L. P. Lipsitt (Eds.), *Advances in child development and behavior* (Vol. 5, pp. 181–211). New York: Academic.

Hadwin, A. F., Winne, P. H., Stockley, D. B., Nesbit, J. C., & Woszczyna, C. (2001). Context moderates students' self-reports about how they study. *Journal of Educational Psychology*, *93*, 477–487.

Hareli, S., & Weiner, B. (2002). Social emotions and personality inferences: A scaffold for a new direction in the study of achievement motivation. *Educational Psychologist*, *37*, 183–193.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Monterey, CA: Brooks/Cole.

Lesgold, A. (2000). What are the tools for? Revolutionary change does not follow the usual norms. In S. Lajoie (Ed.), *Computers as cognitive tools:*

Vol. 2. No more walls (pp. 399–408). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Levin, J. R. (2004). Random thoughts on the (In)credibility of educational–psychological intervention research. *Educational Psychologist*, *39*, 173–184.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational and behavioral treatment. Confirmation from meta-analysis. *American Psychologist*, *48*, 1181–1209.

Lowther, D. L., Gassoppo-Moyo, T., & Morrison, G. R. (1998). Moving from computer literate to technologically competent: The next school reform. *Computers in Human Behavior*, *14*, 93–109.

Maki, R. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117–144). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Means, B. (Ed.). (1994). *Technology and education reform.* San Francisco: Jossey-Bass.

Means, B., & Olson, K. (1994). Tomorrow's schools: Technology and reform in partnership. In B. Means (Ed.), *Technology and school reform* (pp. 191–222). San Francisco: Jossey-Bass.

Menon, G., & Yorkston, E. A. (2000). The use of memory and contextual cues in the formation of behavioral frequency judgments. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 63–79). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Miller, P. H., & Seier, W. L. (1994). Strategy utilization deficiencies in children: When, where, and why. In H. W. Reese (Ed.), *Advances in child development and behavior* (Vol. 25, pp. 107–156). New York: Academic.

Morgan, M. (1985). Self-monitoring of attained subgoals in private study. *Journal of Educational Psychology*, *77*, 623–630.

Norman, D. A. (1988). *The psychology of everyday things.* New York: Basic Books.

Papert, S. (1980). *Mindstorms: Children, computers and powerful ideas.* New York: Basic Books.

Paas, F., Renkl, A., & Sweller, J. (Eds.). (2003). Cognitive load theory [Special issue]. *Educational Psychologist*, *38*(1).

Pinker, S. (2002). *The blank slate: The modern denial of human nature.* New York: Viking.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)* (Tech. Rep. No. 91-B-004). Ann Arbor: University of Michigan, School of Education.

Pressley, M., & Harris, K. (in press). Cognitive strategy instruction: From basic research to classroom instruction. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Quilici, J. H., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, *88*, 144–161.

Rabinowitz, M., Freeman, K., & Cohen, S. (1992). Use and maintenance of strategies: The influence of accessibility to knowledge. *Journal of Educational Psychology*, *84*, 211–218.

Reese, H. W. (1962). Verbal mediation as a function of age level. *Psychological Bulletin*, *59*, 502–509.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology*, *82*, 498–504.

Schwartz, B. (2004). *The paradox of choice: Why more is less.* New York: HarperCollins.

Simon, H. A. (1953). *Models of man.* New York: Wiley.

Shils, E. (1983). *The academic ethic*. Chicago: University of Chicago Press.

Sleeman, D., & Brown, J. S. (Eds.). (1982). *Intelligent tutoring systems.* London: Academic.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356–442). Washington, DC: American Council on Education.

Suppes, P. (1966). The uses of computers in education. *Scientific American*, *215*, 206–221.

Suppes, P., & Morningstar, M. (1969). Computer-assisted instruction. *Science*, *166*, 343–350.

Theide, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*, 66–73.

Third International Mathematics and Science Study. (1998). *Third International Mathematics and Science Study, Pursuing excellence: A study of U.S. twelfth-grade mathematics and science achievement in international context* (Report No. NCES 98–049). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, England: Cambridge University Press.

Van der Meij, H. (1990). Question asking: To know that you do not know is not enough. *Journal of Educational Psychology*, *82*, 505–512.

Weiner, B. (1986). *Attribution theory of motivation and emotion.* New York: Springer-Verlag.

Weinstein, C. E., Schulte, A., & Palmer, D. (1987). *LASSI: Learning and study strategies inventory.* Clearwater, FL: H & H.

Winne, P. H. (1982). Minimizing the black box problem to enhance the validity of theories about instructional effects. *Instructional Science*, *11*, 13–28.

Winne, P. H. (1983a). Distortions of construct validity in multiple regression analysis. *Canadian Journal of Behavioral Science*, *15*, 187–202.

Winne, P. H. (1983b). Training students to process text with adjunct aids. *Instructional Science*, *12*, 243–266.

Winne, P. H. (1992). State-of-the-art instructional computing systems that afford instruction and bootstrap research. In M. Jones & P. H. Winne (Eds.), *Adaptive learning environments: Foundations and frontiers* (pp. 349–380). Berlin, Germany: Springer-Verlag.

Winne, P. H. (1997). Experimenting to bootstrap self-regulated learning. *Journal of Educational Psychology*, *89*, 397–410.

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed., pp. 153–189). Mahwah, NJ: Lawrence Erlbaum Associates.

Winne, P. H. (2005). Key issues in modeling and applying research on self-regulated learning. *Applied Psychology: An International Review*, *54*, 232–238.

Winne, P. H. (in press). Meeting challenges to researching learning from instruction by increasing the complexity of research. In J. Elen & R. E. Clark (Eds.)*, Handling complexity in learning environments: Research and theory.* Amsterdam: Pergamon.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Winne, P. H., Hadwin, A. F., Nesbit, J. C., Kumar, V., & Beaudoin, L. (2005). gSTUDY: A toolkit for developing computer-supported tutorials and researching learning strategies and instruction (Version 2.0.3) [Computer program]. Burnaby, British Columbia, Canada: Simon Fraser University, Faculty of Education.

Winne, P. H., & Jamieson-Noel, D. L. (2002). Exploring students' calibration of self-reports about study tactics and achievement. *Contemporary Educational Psychology*, *27*, 551–572.

Winne, P. H., & Jamieson-Noel, D. L. (2003). Self-regulating studying by objectives for learning: Students' reports compared to a model. *Contemporary Educational Psychology*, *28*, 259–276.

Winne, P. H., Jamieson-Noel, D. L., & Muis, K. (2002). Methodological issues and advances in researching tactics, strategies, and self-regulated learning. In P. R. Pintrich & M. L. Maehr (Eds.), *Advances in motivation and achievement: New directions in measures and methods* (Vol. 12, pp. 121–155). Greenwich, CT: JAI Press.

Winne, P. H., Jamieson-Noel, D. L., & Muis, K. (2005). *Calibration of self-reports about study tactics and achievement: A replication.* Manuscript submitted for publication.

Winne, P. H., & Marx, R. W. (1989). A cognitive processing analysis of motivation within classroom tasks. In C. Ames & R. Ames (Eds.), *Research on motivation in education* (Vol. 3, pp. 223–257). Orlando, FL: Academic.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531–566). Orlando, FL: Academic.

Wolters, C. A., & Pintrich, P. R. (1998). Contextual differences in student motivation and self-regulated learning in mathematics, English, and social studies classrooms. *Instructional Science, 26,* 27–47.