EDUCATIONAL PSYCHOLOGIST, *45*(4), 267–276, 2010

ISSN: 0046-1520 print / 1532-6985 online
DOI: 10.1080/00461520.2010.517150

Routledge
Taylor & Francis Group

Improving Measurements of Self-Regulated Learning

Philip H. Winne
Faculty of Education
Simon Fraser University, Burnaby, Canada

Articles in this special issue present recent advances in using state-of-the-art software systems that gather data with which to examine and measure features of learning and particularly self-regulated learning (SRL). Despite important advances, there remain challenges. I examine key features of SRL and how they are measured using common tools. I advance the case that traces of cognition and metacognition offer critical information about SRL that other state-of-the-art measurements cannot.

Articles in this special issue report leading-edge work on gathering data and measuring constructs that comprise self-regulated learning (SRL). In this article, I strive to mark keys for future work relating to four claims: (a) SRL is contextual and this must be taken into account in measuring it. (b) SRL can be conceptualized as an event and two common forms of self-report data—responses to inventories and think aloud—inadequately measure it as such. (c) Traces—observable representations of cognitive, metacognitive and motivational events—are keys to more fully modeling SRL processes. In my conclusions, I conjecture that (d) widespread use of computer-based learning environments (CBLEs) is vital to significantly accelerating the science of learning, particularly regarding SRL, and applying its findings in education.

CHALLENGES TO MEASURING SRL

Research on SRL struggles with a problem that is nicely illustrated by Herbert Simon's account of an ant making its way across a beach (Simon, 1981, pp. 63ff.). From a few meters away, we perceive the ant's path to have a direction. An ant psychologist might attribute latent constructs to this ant—it is motivated to go somewhere in particular and using problem-solving skills to get there. Broadening our scope of empirical observation, we might notice that the ant is crossing a landscape of troughs formed by a light westerly breeze and persistent waves lapping at the beach. On yet closer inspection, we observe the ant swerves around larger grains of sand, sometimes bearing left and sometimes right. Because these obstacles are randomly positioned, a lot of variance in the ant's path might be predicted quite well by features of its environment rather than the ant's motivation and cognition. The ant may choose its path using a very simple rule: Take a path of least resistance—avoid uphill travel, thus remaining in a trough, and swerve around large sand grains. On this view, features in the ant's environment determine most of the variance in its path.

With no disrespect to the intelligence of learners, might it be similar for them? Educational psychologists theorize otherwise. They (and I) have proposed complex models of motivated SRL (Winne & Hadwin, 1998; Winne, in press). In the model Hadwin and I developed, SRL unfolds in four phases. In Phase 1, learners develop a perception of factors that bear on a task they choose or are assigned. Their sense of a task is constructed from memories about similar tasks tackled in the past, features of the current situation plus their knowledge about the task's domain and personal attributes that relate to it. In Phase 2, learners set goals for the task they perceive and forge plans for addressing it. Goals are defined by standards learners choose for characteristics of (a) cognitive operations (e.g., rate of progress; Metcalfe & Kornell, 2003) and (b) products of cognitive operations (e.g., qualities of learning; see Kornell & Bjork, 2007). In Phase 3, learners enact plans for studying and solving problems. Throughout Phases 1 to 3, self-regulating learners metacognitively monitor qualities of their work and exercise metacognitive control to make adjustments "on the fly" if they judge there is utility in adapting their work. In Phase 4, which Hadwin

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

and I regard as optional, learners pause, either midtask or upon completion, to reflect on features of Phases 1 to 3 and consider large-scale modifications to their sense of the task, goals and plans, and features of their engagement.

Our model is explicit that these phases are weakly sequenced, that is, learners have freedom to shift among phases at will. We also maintain that SRL is recursive such that products at any point feed forward to shape subsequent work. Events in each phase have a common architecture handily summarized by a first-letter acronym, COPES (Winne, 1997). External and internal *conditions* create a context in which cognitive and metacognitive *operations* generate *products*. Operations and their products are *evaluated* relative to *standards*. In the context of our model, Simon's metaphor of the ant affords two important insights. First, SRL is contextual. Second, context evolves as learners regulate learning.

SRL Is Contextual

As emphasized by Aleven, Roll, McLaren, and Koedinger (2010/this issue) and by Greene, Muis, and Pieschl (2010/this issue), SRL is inherently contextual. At a basic level, this is modeled in terms of an IF–THEN–ELSE production. THENS and ELSES the learner chooses to execute operationalize metacognitive control that is conditional on the result of metacognitively monitoring the task's current conditions, the IFS (Winne, in press). For example, IF sentences in a textbook contain a bold word or phrase, a learner almost automatically will THEN highlight them, ELSE another studying tactic, such as making a note, may be used when the meaning of the sentence is judged to contain information that might require further elaboration. Some THENS and ELSES are blocked when, for example, critical resources are unavailable in (a) the external environment, for example, there is no peer to ask for a second opinion; or (b) the internal environment, for example, the gist of a key principle lies "on the tip of the tongue" but just can't be recalled.

Whether mental activity flows seamlessly or meanders forward, learners choose their next operations conditioned by the environment, including their internal environment of knowledge, developed skills, beliefs, expectations, interests and other mental states. Thus, whether data used to measure features of SRL are (a) responses to self-report survey items administered outside the arena of in-the-moment action while learning is happening, (b) introspections a learner offers by thinking aloud approximately apace with cognition and motivation, or (c) products generated by cognitive and motivational events such as a solution to a problem or a correction to a step in a problem-solving episode, data incompletely represent SRL to the extent they do not explicitly include IFS that set a stage for action.

CBLEs are splendid tools for gathering extensive data about many factors in the learner's environment that may shape SRL. To conjecture boldly, as Sir Karl Popper would recommend, a CBLE that logs how a learner interacts with information provides the most complete operational definition available to a researcher about how a learner navigates a landscape of content and how the learner chooses to operate on information. What the CBLE (and every other instrument of which I know) cannot directly record are facets of the learner's internal environment that shape learners' SRL—the learner's thoughts, feelings, beliefs, and other key latent variables (Borsboom, Mellenburgh, & van Heerden, 2003; Winne & Hadwin, 1998) about which educational psychologists theorize. As I discuss later, methods proposed in this special issue help map these events.

The Context of SRL Evolves Because Learners Regulate Learning

Adopting the view that context shapes learning and SRL opens a Pandora's box that makes researching SRL particularly challenging. As learners regulate learning and problem solving, they may *change the context* (IFS) at choice points where they judge current approaches are subpar. In this updated context, learners may *use new operations*, adapting how they learn and solve problems. That is, SRL shapes not only what learners learn. Information generated by SRL feeds forward to change the context for learning and how learning unfolds.

This may be at least one reason why Simon's ant and learners are intrinsically different cognitive creatures. Although the behavior of both is intrinsically contextual, shaped by environmental factors at each choice point, learners can change their environment, sometimes rapidly and dramatically. In Simon's terms, the environment of a self-regulating learner is an artifact. That artifact has the potential to change every time a learner adapts. For example, learners trained to apply a particular learning strategy, even if they have demonstrated skill with it prior to the test of an intervention, may modify that method or reject it altogether as an experiment proceeds. What a researcher presumes to be a constant skill framed by preexperimental training may vary during an experimental session precisely because self-regulating learners regulate learning.

In sum, context is a critical feature of SRL. Data used to represent features of SRL only partially represent its IF–THEN–ELSE form when context is ignored or decoupled from what learners do in context. Moreover, all three terms in the IF–THEN–ELSE representation of SRL may change as learners self-regulate learning. In fact, this is expected in SRL.

VIEWS OF SRL CONSTRUCTS

SRL as Aptitude

A decade ago, Perry and I proposed that research on SRL "need(s) to wrestle with questions about measuring constructs associated with SRL, including components such as metacognition, motivation, and strategic action" (Winne & Perry, 2000, p. 531). We highlighted two different conceptualizations of SRL, as aptitude and as event. We conceptualized aptitudes like Richard Snow did:

> . . . individuals differ in their readiness to profit from a particular treatment at a particular time; aptitude constructs are theoretical concepts fashioned to interpret these observed differences in person-situation interaction terms . . . aptitude is not limited to intelligence or some fixed list of differential abilities but includes personality and motivational differences along with styles, attitudes, and beliefs. (Snow, 1991, p. 205)

Aptitudes are essential to researching SRL. When coding think aloud data, Moos and Azevedo (2008) sought to capture learners' accounts of plans, expectations, evaluations, feelings, and other factors the learners believed to shape how they learn and regulate learning (see Table 1 in Azevedo, Moos, Johnson, & Chauncey, 2010/this issue). Greene et al. (2010/this issue) reviewed research that supports the hypothesis that epistemic beliefs moderate SRL because beliefs shape what learners perceive about information they study, how they study it, and kinds of goals that they set. These constructs differ among learners and, in terms of an IF–THEN–ELSE model, they serve as standards learners use to metacognitively monitor, which sets a stage for metacognitive control for enacting particular THENS and ELSES. Aggregated over people and multiple events that unfold across the timeline of each learner's studying session, THENS and ELSES appear to be shaped by these aptitudes (Moos & Azevedo, 2008; Pieschl, Stahl, & Bromme, 2008).

Aleven et al. (2010/this issue) and Graesser and McNamara (2010/this issue) logged patterns of behavior that tutees adopt in everyday tutoring sessions and in CBLE-supported tutoring. They theorize that these behavioral patterns are grounded in SRL. Again, standards learners use in metacognitive monitoring shape choices from their catalog of THENS and ELSES—study tactics and learning strategies—within affordances and constraints of CBLE tutoring systems.

Most aptitudes bearing on SRL have an important property—they are malleable, as Azevedo et al. (2010/this issue) emphasize. Learners can be trained to "have" or "develop" aptitude as Aleven et al. (2010/this issue) described for help seeking. Indeed, education might be characterized as an enterprise that aims to develop aptitudes. What research on SRL adds to this view is an axiom that learners have agency, "a person's ability to control their actions and, through them, events in the external world" (Haggard & Tsakiris, 2009, p. 242). Agents have potential to develop aptitudes on their own.

Merging this claim with the earlier proposition that learners' update context as a learning episode unfolds leads to an important conclusion: Self-regulating learners may well transform an aptitude's level or even its nature over the course of a learning episode, as Greene et al. (2010/this issue) note. As noted earlier, this is significant when researching SRL. First, researchers should not presume aptitudes measured or self-reported prior to an intervention are constant throughout the intervention. Second, researchers need to gather data during interventions, on the fly, to document whether aptitudes change and, if so, in what ways. Treatments can not be assumed to be fixed when agents engage in SRL.

Aptitudes often are theorized in terms of categories. For example, a learner holds one form of epistemic belief or another. Real learners may be more flexible and more responsive to local conditions. In an experiment, Zhou and I provided learners with tags and hyperlinks that operationally represented all four categories of goal orientation (defined by combinations of mastery-performance by approach-avoidance dimensions; Zhou & Winne, 2010). Greene et al. (2010/this issue) recommend the method we used to trace motivation. Most learners in our study used tags and clicked hyperlinks that expressed all four categories of goal orientations in a single study session versus just one goal orientation, say, performance approach orientation. For aptitudes modeled as categories that play a role as standards in metacognitive monitoring, like epistemic beliefs, measures may be better operationalized as profiles across categories versus one-at-a-time category membership (e.g., evaluativist). With a very large set of data to mine, it may be possible to develop predictions about how the shape of a learner's aptitude profile guides the learner's metacognitive control. This would inform research about how IFS link to THENS and ELSES. Data about what learners do, gathered on the fly as learners work, are required for these investigations.

SRL as Event

Perry and I (Winne & Perry, 2000) also proposed that SRL can be conceptualized in terms of events. Events are the very actions learners perform rather than descriptions of those actions or of mental states that actions generate. We operationalized events at "three successively more complex levels: occurrence, contingency and patterned contingency" (p. 535).

An *occurrence* of an event is a tally representing that a researcher observed a state. For example, we could tally occurrences of an event, "Kim made a note," such as Azevedo et al. (2010/this issue) describe or "Paul reached the bottom out hint," such Aleven et al. (2010/this issue) describe. Because a tally or a sum of tallies provides no information about the state that preceded it, these measures omit information about context.

A *contingency* records that a subsequent state was preceded by a prior state. An example: "Kim made a note immediately after highlighting an italicized term." A contingency explicitly includes some features of context. Suppose Kim studied a text containing 100 italicized terms. She highlighted 50 of these and, of the 50 highlighted terms, Kim annotated 30. We can now describe a conditional probability that Kim metacognitively monitors italicized terms as worthy of highlighting, Pr[highlight|italicized term] = 50/100 or 0.50; and the conditional probability that a highlighted italicized term is metacognitively judged appropriate to be annotated, Pr[note|highlight] = 30/50 or 0.60.

A *patterned contingency* refers to regular, repeated arrangements of component contingencies. Patterned contingencies assemble several IF–THEN-ELSE productions like a decision tree or a learning strategy. An example: "Kim made a note immediately after highlighting an italicized term. Right after that, Kim linked the note to that term in the glossary. This pattern repeated for 12 of 30 italicized terms which Kim highlighted and annotated." Patterned contingencies include context and represent it in more complex forms than simple contingencies.

As I argued earlier, researchers cannot access cognitive operations that are root causes of transitions from a prior state to a subsequent state. What researchers may be able to access are products of these operations. For example, a highlighted phrase externalizes the learner's decision to mark that phrase in the presence of a context such as italicized text.

Events provide touch points for researchers to map information on which learners cognitively operate and to infer the cognitive operation applied. When a CBLE is carefully designed to offer learners choices among operations represented by various tools—for example, tagging (of which highlighting is a semantically null form), linking bundles of information, searching, and so forth—and to record traces of events, data about how learners use tools provide raw material for researchers to track aptitudes "in action" and how aptitudes may evolve as they work (see Winne & Hadwin, in press). And, when a CBLE is designed to record every observable state-to-state change in context over the span of a learning episode, the researcher is prepared to gather data needed to more fully model and investigate SRL.

LEARNERS' REPORTS ABOUT SRL

Perry and I observed

> that a measurement protocol is an intervention in an environment, disturbing it in a fashion that causes data to be generated. Using that data and a logic of causal inference, we infer properties and qualities of a target of measurement. Thus, measurement involves understandings about a target, its environment, and causal relations that connect the two . . . measurement is akin to model building and model testing (Cliff, 1982) and, thus, all measures of SRL are reflections of a model of SRL. (Winne & Perry, 2000, pp. 562–563)

Researchers have so far used two main kinds of interventions to generate data for measuring SRL as an aptitude, inventories and think aloud protocols.

Inventories

The first common method for generating data about SRL as an aptitude is asking a learner to imagine learning events. Sometimes context is described nonspecifically in instructions, for example, "When you study . . . "; at other times, the researcher describes a context that spans multiple episodes of studying, for example, "In this course . . . ." In this kind of context, the researcher asks the learner to estimate the frequency, intensity, likelihood, difficulty (effort required), or capability of carrying out a process or of generating a product by some unspecified cognitive operation. Typically, the learner's response options are constrained, for example, to a 1-to-7 Likert scale anchored by *not true of me* and *true of me*. This is relaxed in oral protocols, such as interviews, in which learners can respond as they wish.

Karabenick et al. (2007) documented that small changes in the wording of items matter (e.g., "In my science class . . . " vs. "My teacher . . . " at the beginning of an item asking if "it's okay to make mistakes as long as you are learning."). Colleagues and I previously identified other challenges when measuring SRL as aptitude using learners' self-reports (Hadwin, Winne, Stockley, Nesbit, & Wosczyna, 2001; Winne, Jamieson-Noel, & Muis, 2002; Winne & Perry, 2000).

How Do Learners Answer a Self-Report Item?

Suppose a learner's textbook is structured with different sections: an overview; summaries after each section; several graphs, tables and figures; a list of glossary terms; and a chapter checkup at the end of each chapter. Standards for metacognitively monitoring learning will quite likely differ as a function of these varying contexts (IFs). Operations a learner applies (THENs and ELSEs) to scan content, build comprehension, and repair comprehension failures will quite likely vary across these formats for information. For example, rehearsing a glossary term at the end of the chapter differs from assembling information in a figure with its related text, and both of those events differ from monitoring comprehension of data presented in a table. Operations also may vary within a format as the learner gains knowledge and adapts tactics to understand information conveyed by graphs. Standards and study tactics likely vary even more widely for different assignments related to studying a chapter, for example, preparing for an in-class discussion versus developing prospectus for a project versus reviewing for a test on the chapter.

When an inventory asks about learning in "this course" or "for exams," these variations make a single learner's answer ambiguous if the researcher can not answer these questions:

- What defines the population of experiences—IFs, THENs, and ELSEs—the learner samples from memory? Does the learner consider sampling error and the sampling fraction (the ratio of sample size to population size) in forming a response intended to represent the population of experiences?

- Given even mild heterogeneity of attributes across experiences in a sample—the various IFs, THENs, and ELSEs constituting SRL—how does a learner compress this variance to form a single response, for example, to respond 5 on a 7-point response scale or answer an interview question by saying, "I usually do X but sometimes I do Y." (And, do expressions this sort indicate the sample of experiences is not homogenous?) Are methods the learner uses inaccurate or biased in the same ways as other forms of cognition are distorted by heuristics and biases (e.g., see Baron, 2008).
- If several learners make identical responses—for example, all report 5 on a 7-point Likert scale or provide semantically equivalent paraphrases to an interview question—do the learners share identical IFs, THENs, and ELSEs? Might identical responses refer to various IFs, THENs, and ELSEs generated on the basis of differentially biased samples that arise from using different heuristics? Is it valid to apply quantitative operators to responses, for example, to equate similarly phrased responses and to form a single category such as summarizing or to average scores across Likert item responses to form a subscale of summarizing? These quantitative moves require that units be interchangeable.

For SRL events modeled in IF–THEN-ELSE form, ignorance about these matters renders self-reports difficult to interpret. Their reliability and meaning are suspect.

Think Aloud Protocols and Unstructured Interviews

A second prevalent intervention used to generate data for measuring SRL involves the learner in a particular learning task or problem-solving episode and, without further constraining when or what a learner might report, invites the learner to "think aloud" about mental states approximately concurrently with their occurrences. When the account is retrospective (recall after a session has ended), the protocol is an unstructured interview.

Think aloud self-reports differ fundamentally from inventory-generated self-report data. First, they are temporally proximal to cognitive events and mental states that learners describe. Second, conditions that prompt data to be generated are a learner's decisions to think aloud about a topic the learner chooses rather than a researcher's question about a particular topic. Despite these differences, think aloud data represent aptitudes in Snow's sense rather than events as Perry and I operationalized them (Winne & Perry, 2000; see also Winne, Zhou, & Egan, in press) because think aloud data do not instantiate an event "in action." Think aloud data are learners' interpretations of events.

Think aloud and retrospective unstructured interview data may appear to overcome challenges just described for self-report data. This is not the case. The researcher has no information about how a learner samples on-the-fly experiences and determines some of them to be topics of think aloud utterances. Prompting learners to "Remember to think aloud"—because it appears to the researcher that the sampling fraction is too low—confounds the problem. Data generated by "natural" sampling, that is, at the learner's discretion, cannot be treated as equivalent to data generated by prompts.

To my knowledge, no research using think aloud data has attempted to estimate the sampling frame (attributes a learner uses to monitor when a mental experience will be described) or the sampling fraction. Retrospective interviews and even experiments might help address this issue but will need to take note of limitations of retrospective interview data. But this issue might be addressed without seeking additional information from the learner.

For example, suppose a learner highlights 15 of 20 italicized terms in an assigned text and thinks aloud about 10 of the highlighted terms and two other terms not highlighted. Table 1 tallies these data.

The researcher must make some assumptions. First, assume the learner reliably highlights using a homogeneous set of IFs. Therefore, standards for metacognitively monitoring comprehension and choosing highlighting as the operationalized expression of metacognitive control (a) differ between the 15 items highlighted and the five not highlighted but are the same for (b) the 15 items highlighted and (c) the five items not highlighted. In a verbal response (think aloud or interview), a learner may cast some light on standards applied for the 10 highlighted items. The researcher can then investigate these data to test the assumption that standards are homogeneous (assuming the learner describes them). Post hoc inspection of items highlighted but not described in think aloud episodes may shed light on why the learner didn't think aloud about some items. Similar post hoc inspections of other cells may be revealing.

Without such investigation to tie together information about all of an SRL episode—its IFs, THENs, and ELSEs—the meaning of think aloud self-report data is suspect. Those data depend on strong assumptions that are difficult to validate.

TABLE 1
Tallies of Think Aloud and Highlights in a Learning Task

	Think Aloud Present	*Think Aloud Absent*	*Total*
Highlighted	10	5	15
Not highlighted	2	3	5
Total	12	8	20

Traces

Learners don't behave randomly (see Nickerson, 2002), so events CBLEs can log about how learners study, solve problems, and self-regulate learning are not random either. What accounts for systematic patterns in these logged data about learners' observable behavior? Researchers assume mental operations generate behavior, and those cognitive and metacognitive events are what theories seek to account for (except for advocates of Skinner's paradigm). Because mental events are not observable, I recommend that researchers create opportunities to *trace* them.

A trace is a datum generated by a learner that is approximately simultaneous with the cognitive operations the learner applies to information in working memory. As previously introduced, a very simple trace is a learner's highlight of words in a text. In the act of highlighting, a learner traces metacognitive monitoring and a particular tactic—highlighting—that expresses metacognitive control. Meaning the learner generated by studying the text or some perceptual feature of the text, such as italics, met standards the learner used to monitor work in relation to a task at hand. And, given these conditions, the learner chose to highlight text rather than, for example, make a note elaborating it with a personal example, searching the Internet for related information or asking a peer to explain it.

I proposed traces as data for measuring cognitive and metacognitive events in an attempt to more validly test theories about learners' mental activities while learning. My first idea was straightforward: train learners to reveal proximal products of cognitive operations that are hypothesized to be responsible for learning, and use the presence or frequency of traces as indicators of whether learners applied the theorized cognitive operations (Winne, 1982). In an experimental follow-up, I investigated how objectives and adjunct questions affect learning (Winne, 1983). For example, I trained learners to record in the margin of a text they studied whether their answers to adjunct questions were right, partly right, or wrong using a code: R, P, W. To determine the code they recorded, learners logically had to attempt to answer each adjunct question. The codes traced the product of cognitive operations theorized to improve learning, namely, answering adjunct questions. Such data validate the extent to which learners behaved in accord with a hypothesis that answering adjunct questions improves learning. Among other findings, one particularly startled me: "80 percent of these students did not use the (adjunct questions) to learn from text in the way that previous research has hypothesized" (p. 252).

Graesser and McNamara (2010/this issue) hypothesize that one facet of deep comprehension is building a mental model of a causal structure. To generate trace data for testing a hypothesis about learners creating mental models, they could be trained about what a mental model is and how to display one. For example, train learners to draw a flowchart. Then, in an experiment where it is theorized learners develop deep comprehension by working with a mental model, learners can trace cognitive events involved in developing mental models by successively elaborating a flowchart. If a learner can't draw a flowchart that adequately captures the causal system described in content studied to that point, the cognitive events theoretically attached to mental models as contributors to developing deep comprehension can't be used to explain achievement. If other kinds of traces suggest those deep comprehension processes were applied—for example, if learners took notes in which they expressed in semantic form a fragment of "deep comprehension"—it may be that learners can't organize fragments into "fully" deep comprehension.

Tracing cognitive events can be "unnatural" if participants need to be trained to do something they normally might not do. But, as Perry and I argued, *every* measurement is an intervention that generates data learners would not otherwise generate. The important issues are whether intervening to measure cognition and SRL (a) aligns well with theory, (b) generates data that support the validity of a researcher's interpretation, and (c) might be a good practice for learners to use anyway. In the few cases where learners "naturally" generate traces, an example being highlighting by most undergraduates, what was at first an "unnatural" occurrence has become a "natural" one. What is initially unnatural can become "au naturel" with practice.

Five Standards for Validating Researchers' Interpretations of Traces

A highlighted section of text unambiguously marks that a learner discriminated the highlighted information from information not highlighted. But other information beyond trace data are needed to infer standards the learner used to monitor text and how decisions were reached to exercise metacognitive control in the form of highlighting versus some other action.

Training learners to generate traces prepares them to operationalize a metacognitive skill. When learners use the skill as the researcher trained them, each learner's behavior aligns with a researcher's theoretical interpretation on the condition that five standards are met. These standards support inferences of the form: IF a condition for cognition or SRL is present in the learner's environment (the context), the learner will reliably THEN carry out theoretically specified cognitive operation(s) or ELSE perform theoretically appropriate alternative operation(s).

1. Alertness to IFs

First, the researcher must guarantee that learners are alert to IFs that frame a context for particular cognitive operations. Learners need to develop dependable skill in discriminating conditions that set a stage for particular cognitive, metacognitive, or self-regulating operations.

2. THENS Match IFS per the Theory

Second, as a result of training, learners must clearly and fully understand which particular cognitive operations to apply when specific conditions are perceived. For example, Graesser and McNamara (2010/this issue) describe a second indicator of deep comprehension to be that a learner has a mental model of trade-offs between variables. This might be revealed if learners drew a graph to depict this relation. In an experiment where the researcher's hypothetical cognitive engines for explaining achievement are (a) building mental models and (b) characterizing trade-offs between variables, context might be straightforward and distinctive: Insert two questions at appropriate locations in the text: "What is your mental model now?" and "What is the trade-off between variables now?" A pretest administered after training learners how to answer such questions could validate they know what to do when encountering a question, namely, draw a mental model as a flowchart and graph a trade-off of relations.

3. Capability to Operationalize THENS

Third, the researcher must ensure that the learner is capable of carrying out the underlying cognitive operations when the context is right. For example, using content the experimenter is certain learners deeply comprehend, ask them to draw a flowchart depicting a mental model, and graph a trade-off of variables.

Such training and explicit cuing of specific cognitive productions is rare in experiments on learning; it is likely nearly always absent in the "real world" of schools and the "wild" Internet. In real-world situations, learners themselves must succeed at the preceding three issues—noticing conditions, matching conditions to cognitive operations, and skillfully carrying out cognition—to develop knowledge (Winne & Nesbit, 2009). If they fail to notice contexts that cue particular cognitive operations, achievement is in jeopardy. If they notice a context but are unclear about which particular cognitive operation(s) to use, achievement is in jeopardy. If they notice contexts and accurately match cognitive operations to them but are unskilled in those routines, achievement is in jeopardy. But, even if learners succeed in these three ways, two other issues may interfere with gaining knowledge (Winne & Nesbit, 2009).

4. Motivation

A fourth issue is that learners may lack motivation to activate particular cognitive operations. Reasons for rejecting THENS or ELSES are as varied as theories of motivation. Examples include fear of having to attribute failure to low ability, insufficient incentive, low efficacy, and more (see Winne & Hadwin, 2008). Zhou and Winne's (2010) experiment shows that some kinds of motivation such as level of engagement and goal orientations can be traced.

5. Contextual Support

Fifth, if learners choose to engage when context affords action, they can be stymied if the environment lacks (a) a necessary resource (e.g., a dictionary is not available to look up a key term), (b) adds too much extraneous cognitive load (e.g., requires too many clicks through pages cluttered with irrelevant information to navigate to a web resource), or (c) allows too little time.

Achievement Is Insufficient to Validate Theory

Graesser and McNamara (2010/this issue) characterize an ideally self-regulating learner as one whose "'meta'-knowledge of cognition, emotions, communication, and social interaction is well honed, so SLR [self-regulated learning] is easy" (p. 234). They quickly note learners like this may be rare in most educational settings. Azevedo et al. (2010/this issue) claim that learning in hypermedia environments "typically involves the use of numerous self-regulatory processes" (p. 210). Singly, these descriptions are valid. But the literature is replete with an invalid inference synthesized from them.

Even when educators have data from a preintervention training session that proves learners know and can apply particular tactics and strategies for learning in particular contexts, agentic learners can adopt different goals. If the researchers' theory is valid, learners who pursue the "wrong" goals but use methods the researcher trained will have low achievement scores. Thus, low achievement is ambiguous. Is it due to methods that don't raise achievement (which is the purpose of testing those methods in an experiment) or due to learners adopting goals that don't foster gains in achievement?

Agentic learners trained to use particular tactics and strategies for learning add new tools to their kit. But, because they are agentic, they may choose old tools over new ones or adapt old tools to new circumstances during learning activities. Unless the researcher has prior data to know (a) what the old tools are and (b) that using old tools lowers achievement relative to using new tools, learners who revert to or adapt old tools may score well on achievement measures. Thus, high achievement is ambiguous about its cause.

Without trace data to document that agentic learners choose particular goals and apply particular cognitive and metacognitive operations, even if other experimental matters are well managed, validly ascribing achievement scores to anything in particular is a challenge (see Borsboom et al., 2003).

Traces au Naturel

To research learning with less contrivance than when learners are trained to behave "unnaturally," my concept of traces morphed to become "observable indicators about cognition that students create [without training] as they engage with a task" (Winne & Perry, 2000, p. 551). Suppose a learner has a tool for tagging selections in content they study in the Internet. For a learner preparing a term paper about a hypothesis of interest in the field of educational psychology, some tags might identify dimensions of the subject matter: "control variables" and "effect sizes." Other tags might mark rhetorical features: "find a rebuttal" and "interesting example." Many of today's learners are already using tags like this.

Regardless of whether the content a learner selects for tagging actually describes effect sizes or is an interesting example of a control variable, tagging offers sturdy ground for inferences about the learner's metacognition. When a learner—call her Jen—applies a tag, she successfully managed all five issues, just described, using *her* standards for metacognitive monitoring and *her* choice to exercise metacognitive control. She perceived that the context was appropriate to cognitively operate on content, interpreted that she should tag a text fragment with the particular tag she chose, discriminated particular text to be tagged, and tagged it. Tagging inherently indicates she was motivated to tag and that potential environmental obstacles did not prevent her from completing this metacognitive task (at least in this instance).

This trace documents aspects of Jen's cognition. But, was Jen self-regulating when she tagged sources for her term paper? In general, how can traces inform theorizing about SRL?

CHARACTERISTICS OF TRACES

Tracing cognitive and metacognitive operations that make up what happens in learning and SRL affords new opportunities for measuring and analyzing data beyond counts and other analyses of occurrences. To illustrate, here is a modest but realistic example in which a learner uses tools provided by a software environment called nStudy. (For an overview of nStudy, see Winne & Hadwin, in press).

> In her web browser, Lucy drags her cursor across text to select two sentences. She opens a contextual menu and chooses the option to make a note. When the note window opens, she first titles it. Then, Lucy clicks a dropdown menu in the note window to select one of several forms to use for her note. She chooses a form titled "Argument" that provides a web form with text fields labeled: Claim, Evidence, and My Analysis. Positioned between the text fields where Lucy enters information representing evidence and then her analysis of the argument is a slider labeled "Reliability." She can position it to rate the quality of evidence. nStudy automatically logs that text Lucy enters contains two of the terms she has previously defined in her lexicon for this topic. nStudy adds the titles of these terms to a panel in the note window that shows various information. Lucy double clicks one of those terms to open a window that provides its description.

During this studying segment, nStudy generates time stamps for each software event such as completing the selection of text in the browser (a "mouse up" event), choosing an option from a menu, opening a window, and so forth. nStudy also records traces of events that describe Lucy's learning activities:

1. The text selected in the web browser.
2. The choice to make a note (versus other tools available in the contextual menu).
3. The form selected for structuring information in the note.
4. Text entered into each of the fields in the note form.
5. The rating of reliability of evidence.
6. The choice to open a previously created term.

Each trace corresponds to a product of cognition or metacognition, a THEN in the IF–THEN–ELSE model of cognitive events. Determining IFs that triggered Lucy's actions requires inferences grounded in nStudy's log of traces. First, as acknowledged earlier, data are incomplete about the full context occupying Lucy's attention. Although nStudy's log captures details about a variety of features that are visible as Lucy works in nStudy's interface, there are no data that trace all of these conditions Lucy attends to. Data are missing entirely about information that Lucy retrieves from long-term memory—domain-relevant content, beliefs, motivations, and so on—or that she may construct on the spot using general or domain-specific heuristics and algorithms.

A partial remedy for filling this gap may be achieved because nStudy traces events across multiple interactions Lucy has with content. These multiple, serial records can support inferences in backward fashion along the timeline of traces. For example, as Lucy reads a web page in her browser, no traces are generated other than time passing since opening the page. But, as soon as she selects text—the information about which she will shortly make a note using a schema for an argument—a plausible inference can be drawn given that Lucy used a note form for arguments: Lucy metacognitively monitored a fragment of content and judged it matched standards for information used in an argument. In cognitive terms, because Lucy chose a particular form for the note after selecting text, the argument form, we can infer standards Lucy used to metacognitively monitor what she was reading before she selected the text and annotated it using the note form for an argument. She perceived the selected information fit the schema of an argument, a schema that is explicitly defined by the form she chose for her note.

More can be inferred from these traces. Lucy selected particular text and, in generating her note, she introduced two terms she had previously installed in her nStudy lexicon. Relatively simple semantic analysis of text in Lucy's note and the text she selected in her browser provides grounds for inferring the topic Lucy was investigating. As well, because Lucy included in her note particular terms that did not appear in the text she selected, it also can be inferred that Lucy assembled a "larger" cognitive structure by retrieving from memory her understanding of the information represented by terms because she metacognitively judged them relevant and included them in her note.

Finally, Lucy opened a window for one of the terms after completing her note. Although IFs that represent the full context are not explicit, it seems plausible to infer that, because Lucy reviewed a term—a THEN that nStudy traced—she had metacognitively monitored the state of her understanding about that term in the context of what she was reading and the note she created. She judged her understanding subpar in some way. Opening the term's window traces metacognitive control that manifests Lucy's plan to raise her judgment of learning about that term by rehearsing its meaning.

nStudy offers other tools Lucy could have used to access other information about that term. For example, nStudy's termnet window is a node-link display that uses a simple rule to show how terms relate to one another: IF one term is defined using another term, THEN the terms are related (linked) to one another. That Lucy chose to open a term's window reflects a plan she implemented for repairing incomplete memory about one particular term. It can be inferred that Lucy metacognitively judged reviewing the term in its window was a study tactic with high utility. But she did not open the termnet window. Either she didn't consider this as a THEN or she judged that action didn't have high utility. The data are not clear about which was the case.

This brief scenario invites three conclusions. First, tools learners use to operate on information in the medium of a CBLE like nStudy can be designed to gather useful trace data that operationally define quite a lot about *how* learners work. Second, although trace data are proximal to latent IF–THEN–ELSE productions that underlie how learners use a CBLE's tools and affordances, there are important conditions (IFs) that traces can not reveal. As well, operations (THENs) that the CBLE's tools are not designed to instantiate and operations the learner does not express in the medium of the CBLE are hidden. Third, although the full set of conditions that trigger actions almost always need to be inferred, it is plausible that multiple instances of patterned behavior—strategies and SRL—can be inferred in dependable ways by indexes computed using matrices that tally transitions from one SRL event to another (see Winne et al., 2002).

CONCLUSIONS

SRL is contextual. This is explicitly reflected in the first term of its IF–THEN–ELSE model. Measuring SRL as an aptitude by self-reports gathered before or after learning sessions casts an adaptive, on-the-fly behavior as a homogeneous, static state. This falls short of accounting for context, except in the case when productions are utterly rigid: IF X, THEN Y no matter what. As well, measurements based on such self-reports may have other significant shortcomings in representing variance in learners' intentions and experiences. Think aloud data, if they avoid pitfalls I described, are better measurements of SRL on the fly than are inventories administered before or after learning, but these data still suffer important drawbacks regarding what experiences are sampled and how heterogeneity is averaged over in a response. Neither inventory nor think aloud data are sufficient to represent SRL as the dynamic, contextual, and adaptive process it is theorized to be.

Trace data operationalize what learners do as they do it. Trace data avoid shortcomings of (a) asking learners what they believe they do and (b) asking learners to perform mental calculations of unknown kinds (c) using sample fractions of past or possible future experiences that have unknown size and biases. When traces are faithful operational definitions of theoretical cognitive and metacognitive operations, they provide sturdy grounds for testing theories about when, whether, and how SRL processes affect learning. On the downside, like every measurement procedure, gathering traces requires intervening in learning experiences to generate the data. If the interventions are too unnatural, their capacity to support valid inferences is undermined.

Measurements based on trace data are neither a panacea nor sufficient to test theories as Azevedo et al. note (2010/this issue). Learners typically need to be trained to trace, coloring trace data with a degree of inauthenticity. This potential drawback vanishes when traces are "au naturel," that is, learners automate an IF–THEN–ELSE form of action that instantiates cognition and metacognition. Using traces as data from which to index patterns of SRL is worth considering whenever an observable behavior can indicate a latent factor that affects cognition and metacognition (see Winne et al., 2002; Winne et al., in press).

Surveys and think aloud protocols can complement trace data when researchers are interested in learners' interpretations of and memories about IF–THEN–ELSE events. Azevedo et al.'s (2010/this issue) article exemplifies this tack. Two questions about self-report data juxtaposed with trace data need address. First, to what extent, under what learning conditions and *why* might these two kinds of data be interchangeable as bases for measuring SRL processes? This question about interchangeability is scientifically interesting in its own right and a necessary precursor to the second question: Should self-report data and trace data be integrated to generate a fuller account of SRL and, if yes, how?

CBLEs Are Key to Advancing Learning Science and Education

The research programs described in this special issue illustrate that CBLEs are tools at the "cutting-edge [of] theoretical and methodological advances regarding how to accurately model and measure learners' SRL processing" (Greene & Azevedo, 2010/this issue, p. 205). I am enthusiastic about these efforts. Elsewhere, I urged that software systems like these need to come out of the laboratory and gain widespread use throughout education (Winne, 2006).

Data the field has garnered so far about SRL and constituent cognitive and metacognitive processes are too few and insufficiently representative. There is still too "little information about measurement issues and uses of trace methods" (Winne & Perry, 2000, p. 553) and not enough understanding about how trace data can articulate with self-report data in researching SRL. If CBLEs become prevalent tools throughout education and if learners will share with researchers the data CBLEs log as learners use the tools CBLEs offer in everyday and experimental ways, I conjecture these limitations can be quickly surmounted. Alongside advances in research methods like those presented in this special issue, the voluminous and more natural data that could be warehoused with widespread use of CBLEs can support significant advances in learning science and substantial acceleration in applying its findings to improve learning.

ACKNOWLEDGMENTS

Support for this research was provided by grants to Philip H. Winne from the Social Sciences and Humanities Research Council of Canada (410-2007-1159), the Canada Research Chair Program and Simon Fraser University.

REFERENCES

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2010/this issue). Automated, unobtrusive, action-by-action assessment of self-regulation during learning with an intelligent tutoring system. *Educational Psychologist, 45,* 224–233.

Azevedo, R., Moos, D, C., Johnson, A. M., & Chauncey, A. D. (2010/this issue). Measuring cognitive and metacognitive regulatory processes during hypermedia learning: Issues and challenges. *Educational Psychologist, 45,* 210–223.

Baron, J. (2008). *Thinking and deciding* (4th ed.). New York, NY: Cambridge University Press.

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

Cliff, N. (1982). What is and isn't measurement. In G. Keren (Ed.), *Statistical and methodological issues in psychology and social sciences research* (pp. 59–93). Hillsdale, NJ: Erlbaum.

Graesser, A., & McNamara, D. (2010/this issue). Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist, 45,* 234–244.

Greene, J. A., & Azevedo, R. (2010/this issue). The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educational Psychologist, 45,* 203–209.

Greene, J. A., Muis, K. R., & Pieschl, S. (2010/this issue). The role of epistemic beliefs in students' self-regulated learning with computer-based learning environments: Conceptual and methodological issues. *Educational Psychologist, 45,* 245–257.

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

Haggard, P., & Tsakiris, M. (2009). The experience of agency: Feelings, judgments and responsibility. *Current Directions in Psychological Science, 18,* 242–246.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, B., Bonney, C. R., . . . Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist, 42,* 139–151

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14,* 219–224.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132,* 530–542.

Moos, D., & Azevedo, R. (2008). Self-regulated learning with hypermedia: The role of prior knowledge. *Contemporary Educational Psychology, 33,* 270–298.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review, 109,* 330–357.

Pieschl, S., Stahl, E., & Bromme, R. (2008). Epistemological beliefs and self-regulated learning with hypertext. *Metacognition and Learning, 1,* 17–37.

Simon, H. A. (1981). *The sciences of the artificial* (2nd ed.). Cambridge, MA: MIT Press.

Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology, 59,* 205–216.

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

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

Winne, P. H. (2006). How software technologies can improve research on learning and bolster school reform. *Educational Psychologist, 41,* 5–17.

Winne, P. H. (in press). A cognitive and metacognitive analysis of self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance*. New York, NY: Routledge.

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: Erlbaum.

Winne, P. H., & Hadwin, A. F. (2008). The weave of motivation and self-regulated learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 297–314). Mahwah, NJ: Erlbaum.

Winne, P. H., & Hadwin, A. F. (in press). nStudy: Tracing and supporting self-regulated learning in the Internet. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies.* New York, NY: Springer.

Winne, P. H., Jamieson-Noel, D. L., & Muis, K. R. (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., & Nesbit, J. C. (2009). Supporting self-regulated learning with cognitive tools. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 259–277). New York, NY: Routledge.

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 Press.

Winne, P. H., Zhou, M., & Egan, R. (in press). Assessing self-regulated learning skills. In G. Schraw (Ed.), *Assessment of higher-order thinking skills.* New York, NY: Routledge.

Zhou, M., & Winne, P. H. (2010). *Modeling academic achievement by self-reported versus traced goal orientation.* Manuscript submitted for publication.