COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 42 (2018) 1777–1832

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12626

Translating the ICAP Theory of Cognitive Engagement Into Practice

Michelene T. H. Chi,[a] Joshua Adams,[a] Emily B. Bogusch,[b] Christiana Bruchok,[a] Seokmin Kang,[c] Matthew Lancaster,[d] Roy Levy,[e] Na Li,[f] Katherine L. McEldoon,[g] Glenda S. Stump,[h] Ruth Wylie,[i] Dongchen Xu,[j] David L. Yaghmourian[k]

[a]*Mary Lou Fulton Teachers College, Arizona State University*
[b]*Phoenix Union School District*
[c]*School of Informatics and Decision Systems Engineering, Arizona State University*
[d]*Department of Psychology, Lourdes University*
[e]*Denny Sanford School of Social and Family Dynamics, Arizona State University*
[f]*Center for Human Applied Reasoning and IOT, University of Southern California*
[g]*Tennessee State Board of Education*
[h]*Strategic Initiatives Group, Office of Digital Learning, Massachusetts Institute of Technology*
[i]*Center for Science and the Imagination, Arizona State University*
[j]*World Wildlife Fund*
[k]*Institute for the Science of Teaching & Learning, Arizona State University*

Received 4 November 2016; received in revised form 19 March 2018; accepted 23 April 2018

Abstract

ICAP is a theory of active learning that differentiates students' engagement based on their behaviors. ICAP postulates that ***I****nteractive* engagement, demonstrated by co-generative collaborative behaviors, is superior for learning to ***C****onstructive* engagement, indicated by generative behaviors. Both kinds of engagement exceed the benefits of ***A****ctive* or ***P****assive* engagement, marked by manipulative and attentive behaviors, respectively. This paper discusses a 5-year project that attempted to translate ICAP into a theory of instruction using five successive measures: (a) teachers' understanding of ICAP after completing an online module, (b) their success at designing lesson plans using different ICAP modes, (c) fidelity of teachers' classroom implementation, (d) modes of students' enacted behaviors, and (e) students' learning outcomes. Although teachers had minimal success in designing *Constructive* and *Interactive* activities, students nevertheless learned significantly more in the context of *Constructive* than *Active* activities. We discuss reasons for teachers' overall difficulty in designing and eliciting *Interactive* engagement.

Correspondence should be addressed to Michelene Chi, Mary Lou Fulton Teachers College, Arizona State University, P.O. Box 872111, Tempe, AZ 85287-2111. E-mail: michelene.chi@asu.edu

Keywords: Active learning; Cognitive engagement; Constructive learning; Co-constructive learning; Collaborative learning

1. Introduction

This paper describes a 5-year project that attempted to teach K-12 teachers about a theory of cognitive engagement called ICAP, which defines different ways that students can engage with instructional materials to learn more deeply. ICAP stands for four cognitive engagement modes, labeled as *Interactive*, *Constructive*, *Active, and Passive* (Chi, 2009; Chi & Wylie, 2014). After a professional development (PD) training in ICAP, teachers were asked to translate ICAP into practice. Their success at translating ICAP was assessed by (a) pre- and post-tests of their understanding of ICAP; (b) the design of their own lesson plans; and (c) the implementation of their own lesson plans. To further gauge the success with which teachers translated their knowledge of ICAP into practice, (d) we checked the way students enacted the activities that the teachers had designed, and finally, (e) students' learning was measured by pre- and post-tests. Before describing our instruction about ICAP to teachers, we first review how cognitive engagement and related constructs such as "active learning," "deep versus shallow processing," "on-task versus off-task," and "hands-on versus minds-on" have been defined and conceptualized in the literature, as well as how these terminologies are related. We then introduce the ICAP theory and its predictions, followed by a description of this translation project. We close with a discussion of the unique features of this project, the challenges teachers faced, and insights we have gained.

2. What is engagement and how is it related to "active learning?" Terminologies

Engagement is a construct that is discussed primarily in the K-12 education literature, whereas active learning is discussed predominantly in the post-secondary literature either in the context of flipped classrooms or online learning, as well as in the educational technology and machine learning literature. These terminologies are used widely and broadly, without concrete definitions. Below, we briefly highlight what the terms mean in these three sets of literature, and we refer to other comprehensive reviews that are available for more details.

2.1. School engagement in the K-12 education literature

School engagement has been a prominent construct in the education literature because it is associated with academic achievement. Broadly, it refers to students' level of commitment and involvement with schooling, and comprehensive reviews have been provided (see Fredricks, Blumenfeld, & Paris, 2004; Sinatra, Heddy, & Lombardi, 2015).

Engagement is typically explored in multifaceted ways, most notably from the emotional, behavioral, and cognitive perspectives, with various measures of engagement used for each perspective. Measures of each perspective generally show a positive relationship with academic achievement. Below we briefly describe the three major perspectives of school engagement, its measures, and its relationship with student learning/achievement.

Emotional engagement refers to students' affective reactions, such as attitudes, happiness, boredom, interests, and values toward teachers, classmates, and subject domains taught in school. Emotional engagement is typically measured by self-reports using survey instruments, and occasionally by experience sampling (Csikszentmihalyi, 1988; Fredricks et al., 2004). Positive emotions promote greater engagement than negative emotions (Broughton, Sinatra, & Nussbaum, 2013); in turn, this greater engagement associated with positive emotions results in increased academic achievement (Heddy & Sinatra, 2013; Pekrun & Linnenbrink-Garcia, 2012).

Behavioral engagement refers to participation in schooling at several grain sizes. At a coarse grain size, behavioral engagement refers to students attending school, participating in extracurricular school activities and doing homework. Once in the classrooms, a finer grain size might be positive conduct such as following rules and adhering to classroom norms. Once students are in the context of instruction, an even finer grain size of behavior engagement may refer to students' effort, persistence, resilience, concentration, paying attention, and contributing to class (Skinner & Belmont, 1993). Measures of behavioral engagement vary depending on the grain size. School attendance is easy to measure objectively in terms of absences and tardiness, whereas conduct and persistence might be measured by teachers' ratings, and levels of participation can be measured by either teachers' ratings or self-reports based on survey questions. Observation techniques using scales such as identifying off-task or deeply involved behaviors have also been used. Again, behavioral engagement has been shown to be related to achievement (Marks, 2000), although Sinatra et al. (2015, p. 2) noted that behavioral engagement may be related only to shallow recall-type of assessment questions, and it may not be related to higher order achievement.

Cognitive engagement has been conceptualized broadly as a student's investment in learning (Wehlage & Smith, 1992). Investment appears to refer to metacognitive effort, such as trying to be strategic and self-regulatory by reflecting on how best to learn (Greene, 2015), such as preferring to solve harder or more challenging problems (Newmann, Wehlage, & Lamborn, 1992). Investment does not seem to refer to cognitive efforts, such as spending time resolving misunderstandings about a problem, and so on. Overall, defining cognitive engagement from a seemingly metacognitive perspective conflates it with motivational constructs (Sinatra et al., 2015), such as adopting learning rather than performance goals (Dweck, 1986), or persisting on challenging tasks (Zimmerman, 1990). Cognitive engagements have also been measured by surveys and self-report questionnaires designed to elicit individual preference for hard work, coping strategies for perceived failures, and self-regulatory strategy usage, as well as how students set goals, plan, organize, and monitor their study efforts. Although some observational techniques of behaviors have been tried, such as documenting evidence of learner persistence,

it is assumed that cognitive engagement cannot be inferred readily from behavioral assessment or from self-report measures (Fredricks et al., 2004, p. 68).

One of the most well-specified definitions of cognitive engagement is to conceive of it as the type and degree of cognitive strategy use. The degree of strategy use can be assessed by survey items. For example, Greene and Miller's (1996) scale of cognitive engagement specifies three subscales, one on self-regulation (with questions such as *I planned out how I would study the material for this exam*), one on the use of deep processing strategies (such as *When learning new materials*, *I summarize it in my own words*), and one on the use of shallow processing strategies (such as *I underlined main ideas as I read for course assignments*). Deep and shallow strategies are grounded in the depth of processing framework (Craik & Lockhart, 1972), in that deep strategies refer to those that involve the use of, linking and integrating with prior knowledge, and elaborating the to-be-learned materials. Shallow processing refers to using rote processing strategies such as rehearsing and verbatim memorization (Greene, 2015, p. 15). However, according to Dinsmore and Alexander (2012), there is a great deal of variation in how deep and shallow processing are conceptualized.

Aside from measuring cognitive engagement by survey items related to deep and shallow processing strategies, there are also numerous studies that tested the effectiveness of various strategies per se, such as the advantage of having students pose questions (King, 1992), or reason from evidence, and strategies of inquiry learning (Klahr & Nigam, 2004). There are also studies that pitted the advantage of one strategy over another. For example, Webb, Troper, and Fall (1995) have shown that giving explanations is clearly beneficial to students' learning, more so than receiving explanations, unless students further use the received explanations to try to solve problems. Numerous studies have also compared the benefits or lack of benefits of collaboration while learning. Overall, research on specific strategies sometimes obtain discrepant results, and there are also no clear and concrete definitions for how the variety of studying strategies are different from each other (other than to claim them as deep or shallow), nor whether one is more engaging than another and why.

In summary, several points can be gathered from this K-12 literature on the construct of school engagement, whether from an emotional, behavioral, or cognitive perspective. The first point is that students can engage to varying degrees, and the more engaged a student is, the higher the achievement. However, neither systematic definitions nor metrics are given for "degrees of engagement." A second point is that of these three perspectives of engagement, cognitive engagement seems to be the least well-defined (Greene, 2015; Sinatra et al., 2015) and conflates with related constructs such as motivation, self-regulation, metacognition, and strategy usage. Perhaps because of its broadness and vagueness, it is important to note that deep/meaningful cognitive engagement has not always been linked to achievement (Ravindran, Greene, & DeBacker, 2005; cited in Greene, 2015; Sinatra et al., 2015). Moreover, without concrete and operational definitions, it is difficult to inform teachers on how to increase cognitive engagement.

A third point is that cognitive engagement also suffers from being measured by survey instruments, and moreover, as Greene (2015) has pointed out, in surveys, the engagement

variable often becomes the outcome, rather than a predictor of achievement. Other methods to assess cognitive engagement are emerging, especially dynamic measures such as observation protocols (Greene, 2010), trace analyses (Winne, 2010), learning analytics (Gobert & Sao Pedro, 2017), and experience sampling. Curiously, the cognitive engagement literature does not talk about the cognitive processes involved in being engaged, nor what measures can be used to stand in for the underlying cognitive processes.

2.2. Active learning at the post-secondary level

The literature on active learning typically refers to how students engage in classrooms, during instruction. Research at the post-secondary level has been the most prominent in promoting and using active learning in college classrooms, perhaps because it (active learning defined mostly as group or collaborate work) has been shown to be effective for quite some time by Harvard physicists (Crouch & Mazur, 2001; Hake, 1998) and more recently promoted by Nobel Laureate in physics, Carl Wieman (Deslauriers, Schelew, & Wieman, 2011). In fact, Eric Mazur (Bajak, 2014; Crouch & Mazur, 2001) had called for a ban on lecturing and promoted group work instead, along with flipped classrooms (Grabinger & Dunlap, 1995). A recent comprehensive meta-analysis of 225 studies in science domains (Freeman et al., 2014) has shown unambiguously that active learning has the potential to enhance student learning, compared to passive learning. Thus, in contrast to inconsistent findings relating cognitive engagement to achievement (as pointed out above, Greene, 2015; Sinatra et al., 2015), implementations of active learning have had tremendous successes in enhancing student learning and retention at the college level. This may be due in part to the easier binary discrimination of active versus passive learning, whereas cognitive engagement is difficult to measure and seems to vary in degrees.

However, what is active learning exactly? What kind of student activities constitutes active learning and what kind not? As examples, we briefly indicate here how active learning has been defined or practiced in two sets of post-secondary literatures. In college-level instruction, active learning has often been defined in two ways: either from the students' perspective, dichotomously as a contrast to passive learning (i.e., students are either doing something or not doing anything extra, other than listening to lectures), or from the instructors' perspective, in terms of what the instructors do, that is, lecturing or not lecturing. Even though crossing these two dichotomous factors provides four cells, passive learning is most often defined merely as the context in which students are learning when instructors are lecturing, whereas active learning is defined as everything else that students could be doing while not being lectured, often doing collaborative/interactive activities in small groups or dyads. In short, college instructors have been encouraged to refrain from lecturing because it is assumed that lecturing leads to passive learning, but no operational definitions have actually been provided for what counts as active learning.

In the professional development literature, active learning refers to the general notion that a professional development program should embed active learning strategies throughout the program itself. Such strategies include a variety of teacher activities, such as requiring teachers as learners to practice under simulated conditions, to review student

work, to observe expert teachers, or requiring teachers to be observed by other teachers, followed by feedback and discussion (van Driel, Meirink, Van Veen, & Zwart, 2012).

Typically, active learning in various literatures focuses predominantly on encouraging group work or learning in dyads, either in the format of cooperative learning, collaborative learning, or peer teaching. Again, they are all mentioned as useful activities for active learning without specifying whether they have differential benefits for learning, nor how they should be operationally defined. For example, interactive engagement is sometimes defined in a circular way as methods "designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities" (Hake, 1998, p. 65). This definition suggests that all interactive methods are productive for active learning, which is clearly not true (see Chi & Menekse, 2015).

This extremely brief overview is intended only to suggest that active learning is a popular construct because learners who are active (typically meaning students who are doing something) do learn more than passive learners (typically meaning students who are doing nothing), whether they are college students or teachers. Although a variety of active learning activities have been exemplified across different literatures as described above, overall, no explicit parameters and concrete operational definitions exist, nor have metrics been provided with respect to what kind of activities are active learning activities, and how to determine whether one active learning activity is better than another. Nevertheless, active learning is a popular construct because of the well-documented improved learning outcomes, but teachers and instructors face the practical challenges of not knowing how to design active learning activities, other than to refrain from lecturing.

2.3. Educational technology and machine learning

In the educational technology literature, active learning often means requiring students to look at and pay attention to the instructional materials, and interactive learning often refers to engaging with systems that require students to make a response to a system's actions. For example, in the educational data-mining literature (Baker, 2016) as well as the intelligent tutoring system literature (D'Mello, Olney, Williams, & Hays, 2012), active leaning often means having the system be able to detect whether students are looking at the relevant materials presented on the screen or not. When students are detected to be looking at the learning materials at the right time, then students are considered to be actively engaged. Moreover, students are considered actively interacting with the systems regardless of what kind of responses students are making, such as whether students are required to select an option from a drop-down menu, or whether students are asked to make a free response. In short, students' interactive responses are not differentiated.

Besides engaging with instruction in either an active or passive way, one could also be disengaged with instruction, generally meaning that students could be off-task, such as goofing off or sleeping while the instructor lectures. In the context of online learning or other learning environments, disengagement can be detected when students are gaming the system, such as in "help abuse" (Aleven, McLaren, Roll, & Koedinger, 2006),

meaning that students exploit scaffolding help by clicking through the hint sequence to get the most explicit hint, which usually provides the answer, rather than attempt to solve the problem after seeing each hint. Other disengaged behavior might be doing a totally off-task behavior such as surfing the Web. One could define very precisely when students are disengaged using automated real-time detectors with log files. Systems have been built using features such as the average time between actions input by the students, the frequency and duration of pauses, and so forth (Gobert, Baker, & Wixon, 2015).

In the machine learning literature, active and passive learning is conceptualized in more or less the same dichotomous way. The goal of machine learning is to create computer systems that can improve itself (or learn) based on experiences. Experiences are derived from interactions with some data, such as categorizing thousands of labeled instances of a new variety of fruit—some are sweet and some are sour, let's say (Settles, 2012). After such experiences, the system can learn by inducing which kind of fruit is sweet and which kind is sour. This is considered a "passive" system. However, to label all instances of fruit initially as sweet/sour is often not possible or sometimes unavailable. Thus, a more desirable "active" system is one in which the system can query and select instances to test, based on some algorithmic decision. Thus, active and passive are used in the same dichotomous way in machine learning as in the general literature, in that the active system is one that can take actions, make queries, or do experiments with the data (Cohn, Ghahramani, & Jordan, 1996), while the passive system does not.

2.4. Summary

In summary, Fig. 1 is provided to clarify the use of terminologies, especially as it relates to how we use our terminologies of *Interactive*, *Constructive*, *Active*, and *Passive*, to be described in the next section. In the first row of Fig. 1, practitioners talk about being "on-task" or "off-task," often meaning whether students are at minimum, engaged with instruction or goofing off. Thus, "on-task" maps onto all four of the ICAP engagement modes. In the machine learning literature, "cognitively engaged" versus "cognitively disengaged" are terms similar in meaning to "on-task" and "off-task." In the post-secondary literature, "active learning" is defined as students doing something with the instructional materials, versus students doing nothing (or "passive learning," as shown in row 3 of Fig. 1). Thus "active learning" maps onto the *Interactive*, *Constructive*, and *Active* modes of ICAP, and "passive learning" maps onto the ICAP *Passive* mode. In the education literature, within cognitive engagement, there is a discrimination between using "deep processing strategies" (or processing that is generative and makes inferences, thus map onto the *Interactive* and *Constructive* mode) versus "shallow processing strategies" (which maps on to the *Active* and *Passive* modes). This same mapping applies to the practitioners' ideas of "minds-on" versus "hands-on" (as shown in row 4 of Fig. 1). In the next section, we will specify and define our terminologies within our theory of cognitive engagement and the predictions of our theory.

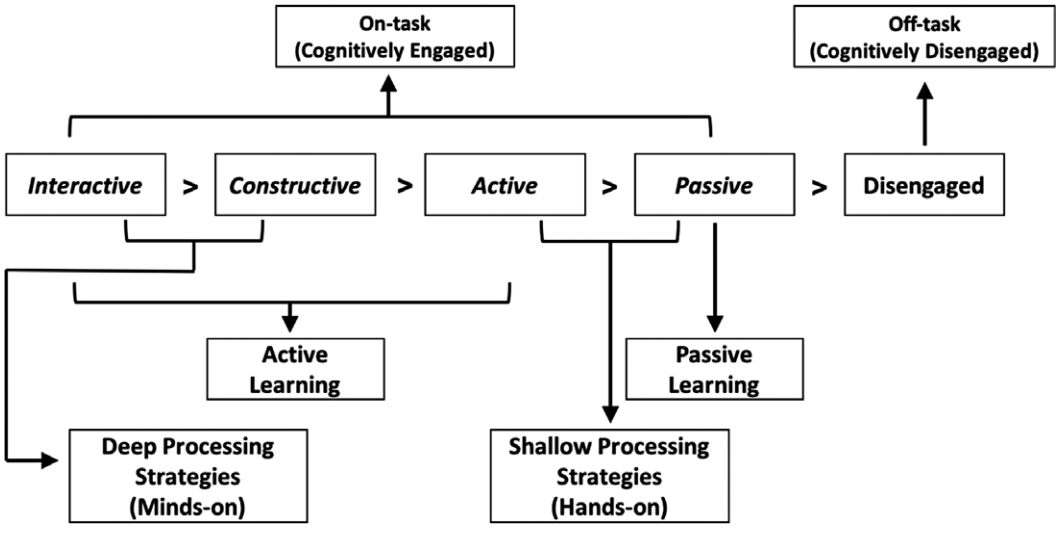

Fig. 1. Terminologies corresponding to ICAP.

3. The ICAP theory for active learning

As the brief review of the engagement and active learning literature shows, good operational definitions, theories, and metrics are lacking, and they are needed to know how to foster them in practice. Moreover, the literature suggests that cognitive engagement, as currently assessed and defined, seems to promote only shallower learning. To rectify these two problems, the ICAP theory was developed to define cognitive engagement or active/passive learning (the two terms are used interchangeably) in ways that can promote deeper learning.

The ICAP theory was first introduced in 2009 (Chi, 2009), in which the paper proposed three cognitive modes of engagement (*Active*, *Constructive*, *and Interactive*), along with evidence in the literature in support of ICAP's predictions that *Interactive* > *Constructive* > *Active*. ICAP was further extended in 2014 (Chi & Wylie, 2014) to include the *Passive* mode, because numerous laboratory and classroom studies the paper cited contrasted the *Passive* mode with one of these three alternative active modes. Thus, the term *Active* in ICAP is a label referring to one mode of engagement, whereas the term "active" in active learning is a broad term referring to all three modes of cognitive engagement, as indicated in Fig. 1. (ICAP was also referred to briefly as DOLA, for differentiated overt learning activities [see Menekse, Stump, Krause, & Chi, 2013].)

ICAP encompasses three components: a taxonomy of four engagement modes and the operational definition of each mode, a metric that can define the degree of engagement based on the cognitive processes corresponding to the four behavioral modes, and a hypothesis that can predict the hierarchical levels of student learning as a function of the mode of engagement. Support for the theory's predictions is culled from findings in

published studies in the literature. Because numerous studies in the literature whose findings have been shown to support the theory have been cited in prior three papers (Chi, 2009; Chi & Wylie, 2014; Fonseca & Chi, 2011), as well as our own laboratory study in the context of learning college engineering concepts (Menekse et al., 2013), we will only describe two studies here, carried out with a much younger population (since the prior papers reviewed mostly studies with adult or K-12 population), just to illustrate how findings are interpreted to suggest support for ICAP. Thus, in this paper, we will briefly describe primarily the taxonomy and the hypothesis here to understand our translation project, with a little more explicit elaborations about the cognitive thinking processes underlying the hypothesis. It is important to remember that cognitive engagement involves thinking processes, even though as we noted above, the literature on cognitive engagement rarely mentions the underlying thinking processes.

3.1. A taxonomy and operational definitions of four modes of cognitive engagement

Faced with a huge range of engaging activities that could be undertaken by college students, primary and secondary students, teachers, and even toddlers, ICAP imposed a taxonomy that can categorize activities into four broad types/modes based on definable differences with respect to the overt behaviors and products students produce or express. The four modes of ICAP are ***I****nteractive*, ***C****onstructive*, ***A****ctive*, and ***P****assive*. Because these terms, *Interactive*, *Constructive*, *Active*, *and Passive* are labels for the four modes of overt behaviors, we will capitalize and italicize them, to distinguish our definitions of them from other usage of these terms in the literature. These four modes are differentiated and operationally defined by the activities students are asked to do, which can often be observed overtly, along with the products students are asked to produce (to be elaborated below).

ICAP imposes three main assumptions. The first assumption of the ICAP theory is that students' overt behaviors and student products, together, can determine the mode of student's cognitive engagement. Although we totally agree that overt behaviors alone are not a good indicator of the underlying cognitive processes in general (Peterson, Swing, Stark, & Waas, 1984), we claim that behavioral engagement, along with student products, jointly, may be an adequate (but not perfect) measure to reflect the differentiated underlying cognitive processes that students are undertaking. In other words, our assumption is that overt behaviors (along with student products) can be differentiated, and these differentiations reflect differences in the underlying thinking processes. So we are using students' overt behaviors (and their products when necessary) as indicators to reflect cognitive engagement. Thus, we are not talking about behavioral engagement, as described in the preceding section.

The second assumption of the ICAP theory is that overt behaviors and the resulting products imply potentially distinguishable underlying cognitive *knowledge-change processes* that may occur. We introduce the term *knowledge-change processes* as domain- and task-general processes that will cause changes in one's knowledge, based on four elementary cognitive processes. Assuming that knowledge can be represented as node-link

structures, we assume that these four elementary processes are sufficient to illustrate how different behaviors can elicit various combinations of these elementary processes, thereby resulting in a metric of degree of engagement that corresponds to different levels of learning. The four elementary processes that are sufficient for our illustration of changes are: *storing*, *activating*, *linking*, and *inferring*. Of course, there may be many more other elementary processes, such as revising or changing, and so forth. But these four are sufficient to justify our hypothesis (to be described below).

The third assumption that ICAP makes is that the correspondence between overt behaviors and the underlying knowledge change processes is not perfect, but good enough. That is, we assume that, by and large, the correspondence holds in the majority of the times. For example, if we use an entire class as a unit of analysis, then it is likely that proportionately what the majority of the students are doing overtly reflects how they are thinking. If the unit of analysis is a single student, then over a period of time, the majority of the time s/he is devoted to a particular overt activity corresponds by and large to the thinking processes of that mode. We now provide definitions for each mode.

3.1.1. Paying attention, or the Passive *mode of engagement*

In our taxonomy, we define *paying attention* as the behaviors of being oriented toward and receiving information from the instructional materials without overtly doing anything else related to learning. Examples of paying attention include reading a text silently, watching a video, or listening to an online lecture without undertaking any other visible activities. We label paying attention as the attentive/*Passive* mode of engagement.

The corresponding underlying knowledge-change processes can be assumed to be *storing* the incoming information to which students are paying attention, perhaps storing it in an isolated fashion. Because the correspondence between overt behaviors and cognitive processes is not perfect, it is possible for students to be covertly processing the learning materials deeply, but overtly appearing only to be *passively* engaged. However, as stated above, we assume that by-and-large, when students are *only* paying attention, they are more likely to be merely taking in (or encoding) the information and *storing* it in isolation, without *activating* prior knowledge and *linking* new information with activated prior knowledge. Thus, their encoding of this new information may result in encapsulated or inert knowledge (Whitehead, 1929). The point is, even though students who are paying attention *could* be thinking more deeply about the materials, but on average, they are likely not. Thus, our attentive/*Passive* mode corresponds more or less to the general notion of passive learning in the literature, as indicating that students are not doing much overtly in terms of engaging with the learning materials, other than looking at the instructional materials. Note that this is the mode of engagement detectable by data-mining techniques, as mentioned earlier.

However, paying attention, resulting in *storing* new information, is obviously adequate for learning materials at a shallow level or in an isolated way, and such stored information can be retrieved and used, especially if the appropriate context is provided. Many simple procedures can be learned via passive learning. For example, learning how to operate an ATM machine may require paying attention only, to memorize the sequential

steps. Many test questions also only require the recall of passively stored information. Thus, *Passive* engagement is quite adequate for some contexts of learning, and it is obviously much better to be attentively engaged than being disengaged (Gobert et al., 2015).

3.1.2. Manipulating, or the Active *mode of engagement*

Learners' engagement with instructional materials can be operationalized as *manipulative* if some form of overt action or physical manipulation is undertaken, without providing any new information. Examples of manipulative activities can be: *pointing* to or *gesturing* at parts of what learners are reading or solving (Alibali & DiRusso, 1999), *pausing and rewinding* parts of a video tapes for review (Chi, Roy, & Hausmann, 2008), *rotating* objects (James et al., 2002), *underlining* (which is similar to *highlighting*, or *clipping*-and-*pasting*) certain text sentences (Katayama, Shambaugh, & Doctor, 2005), *copying* some problem solution steps (VanLehn et al., 2007), *mixing* certain chemical amounts in a hands-on laboratory (Yaron, Karabinos, Lange, Greeno, & Leinhardt, 2010), *choosing* a justification from a menu of options (Conati & VanLehn, 2000), or *repeating* what was already said (O'Reilly, Symons, & MacLatchy-Gaudet, 1998). In all these examples, the instructional materials are being manipulated, but the learners have not provided any new information beyond the instructional materials. For example, in the case of speaking, when students repeat or rehearse what was presented, that would be considered a physical or motoric *manipulative* activity since the content of the utterances does not contain information that goes beyond what was presented. We label manipulating instructional materials as the manipulative/*Active* mode of engagement.

Cognitively, the knowledge-change processes associated with the various manipulative actions of copying, underlining, choosing, and so forth, are that they cause attention to be focused on what is being manipulated. The outcome is that what is in focus may *activate* relevant prior knowledge, allowing the new information to be *linked* and *stored* with the activated prior knowledge, resulting in the new information being assimilated or embedded with this activated prior knowledge. The consequences of *activating* prior knowledge and *linking* is that the resulting *stored* knowledge is more complete and embedded with prior knowledge, thus also making the new knowledge more strengthened for easier retrieval. Thus, manipulative/*Active* engagement involves three elementary cognitive processes (*store*, *activate*, and *link*), and it can be quite adequate for learning in many situations, explaining why "hands-on" activities often facilitate learning.

3.1.3. Generating, or the Constructive *mode of engagement*

The ICAP taxonomy defines *generative* behaviors as those in which learners produce externalized ideas containing information that goes beyond what was provided in the learning materials or instruction. To meet the criteria for generative, the outputs of generative behaviors could be a product, such as a concept map, but the product must show evidence of new ideas that go beyond the information given, defined literally when matching the generated product with the instructional materials. For example, if a student were to compare two cases, the similarities and differences the student comes up with would be generated ideas because similarities and differences were not presented in the

instructional materials. In short, being generative means more than just generating an external physical product (such as a concept map), but the product must contain additional ideas not given during instruction or in the instructional materials. We label the behavior of being generative as the *Constructive* mode of engagement.

There are many examples of generative/*Constructive* behaviors in a learning context, such as *explaining* to others or to oneself (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & LaVancher, 1994), *taking* notes in one's own words (Trafton & Trickett, 2001), *posing* problems (Mestre, 2002), *asking* questions (Graesser & Person, 1994), *drawing* a concept map (Biswas, Leelawong, Schwartz, Vye, & The Teachable Agents Group at Vanderbilt, 2005), *predicting* (Schauble, Glaser, Duschl, Schulze, & John, 2009), *inventing* (Schwartz, Chase, Oppezzo, & Chin, 2011), *arguing*, *inducing* hypotheses, *self-evaluating* or *monitoring* one's understanding, *creating* a timeline (Dawson, 2004), and so on, assuming that all of these activities result in outcomes containing additional ideas that go beyond the original learning materials. Thus, our meaning of *Constructing* subsumes all processes of generating, whether it is to infer a small piece of knowledge, or whether it is to induce a complicated piece of knowledge, such as a hypothesis.

Two caveats to point out here. Although many *Constructive* behaviors show visibly that students are producing information beyond what was presented in the instructional materials, such as making a concept maps or drawing diagrams when none was obviously provided in instruction, or asking questions that were not asked in the materials or by the instructor, sometimes which ICAP mode of activity a student is engaged in may need to be further discriminated and confirmed by students' products. For example, if a student is taking notes, a student could be taking verbatim notes by copying (thus being *Active*) or taking notes in students' own words (being *Constructive*), so that the mode of the same overt behavior of taking notes can only be disambiguated by comparing the notes to instructional materials. Thus, as we stated above, although overt behaviors alone may not be adequate to indicate cognitive processes, overt behaviors coupled with the produced products can be fairly accurate in determining which ICAP mode students are engaging. The second caveat to note about the *Constructive* mode is that it is operationally defined simply and straightforwardly as students generating content that extends beyond the instructional content. Thus, being *Constructive* does not mean that a student will generate knowledge that is new to the content domain (such as a new formula) or even new to oneself. Occasionally a student may even be retrieving some already known stored knowledge related to the instructional content materials; we cannot easily tell from behaviors when retrieving versus generation occurs. However, because we are studying engagement in the context of learning new information, it might be safe for us to assume that most of the time, when students are generating some ideas that extend beyond the information provided in the instruction, they are not retrieving knowledge but are more likely generating knowledge.

Cognitively, the knowledge-change processes associated with being *Constructive* requires that the learners generate new knowledge by *inferring*, either from *activated* prior knowledge or from knowledge integrated and *linked* with new instructional content.

Thus, generative processes include all four elementary processes of *activating* prior knowledge, *linking* with activated prior knowledge, *inferring* from prior knowledge or the newly integrated knowledge, and *storing* the linked and inferred knowledge. Note that the inferences generated could be minute, much like self-explanations (Chi et al., 1989, 1994), and need not be complex reasoning processes such as inductive, deductive, or abductive inferring. Moreover, the inferred knowledge is "new" only in the sense that it was not presented in the content of instruction or the instructional materials. Thus, "new" does not mean "new" in the sense of being novel to the domain, such as a new discovery. We literally consider engagement to be generative/*Constructive* when the student produces some new knowledge extending beyond what was presented. These ideas are similar to traditional ideas of "deep" or elaborative processing (Anderson & Reder, 1979). Greene (2015), for example, considers deep engagement as involving deep elaborative processing, or the "intentional creation of more complex knowledge structures by integrating the new information with prior knowledge," and shallow engagement as involving "rote processing" such as "rote rehearsal and verbatim memorization strategies." We would consider the deep engagement strategies as *Constructive*, and shallow engagement strategies as *Active* (see Row 4 Fig. 1 again).

Our idea of generative/*Constructive* is compatible with several broad philosophical orientations of "constructivism" in the literature, such as an approach to instruction in which a learner's prior knowledge is considered (Ausubel, Novak, & Hanesian, 1986), or the perspective that students should "construct" their own unique systems of knowing rather than "being told" or "instructed" by a teacher, as proposed initially by Piaget (1930), then later by Bruner (1961) and Papert (1980). The ICAP definition of being generative/*Constructive* is compatible with these definitions, but it is more specific, operational, literal, and used as a verb, referring to the generative processes students undertake. Thus, our definition of generative/*Constructive* is determined by comparing what students do behaviorally and cognitively, to what is already presented in the instructional materials. The knowledge-change processes involved, *activating*, *linking*, *inferring* and *storing*, will result in a more elaborated knowledge structure.

Our definition ignores dimensions such as how detailed the outputs are or how correct and complete they are that others have considered (e.g., Webb et al., 2014; considered details; Webb et al. (2009), considered correctness in students' explanations). By our definition, details and completeness are important to the extent that they are generated outputs that provide additional information not already presented by instruction. Therefore, we consider all of the following types of student moves discussed in the literature as of the same generative/*Constructive* kind: explaining, posing questions, making comparisons, elaborating one's own thinking, inventing, and so forth, even though they may engage more or less complicated cognitive processes. That is, our theory cannot and is not meant to predict whether one generative activity (e.g., inventing) is better than another generative activity (e.g., posing questions) for learning. Our theory only predicts that generative activities are superior to manipulative activities, as assessed by students' learning outcomes.

3.1.4. Collaborating, or the Interactive *mode of engagement*

The terms collaborative/*Interactive* in the context of ICAP refers to interactions between two peers (or a small group), often through dialogs, that meet two criteria: (a) both partners' utterances must be primarily generative/*Constructive*, that is, adding ideas beyond what is already presented in the learning materials, and (b) each partner's contributions address or engage the other partner's contributions, thereby they are mutually and reciprocally generating or co-generating. For example, if speakers build-on, elaborate, justify, challenge, or question each other's ideas, then they are mutually and reciprocally co-generative because elaborations, justifications, challenges, and questions typically add information not originally provided in the instructional materials, and each partner is addressing and engaging with the other's questions, challenges, or explanations.

Based on our definition of collaborative, the knowledge-change processes are the same as ones involved in being generative, with the added variation that each speaker can not only *infer* from their own knowledge, but they can also *infer* from the knowledge articulated by the partner, as well as *infer* from partner's knowledge integrated with one's own knowledge, or *infer* from the partner's *inferred* knowledge. In short, collaborative interactions involve the knowledge-change processes of *store*, *activate*, *link*, *infer-from-own*, and *infer-from-other*. This suggests that interactive collaboration has the potential of creating innovative knowledge that neither partner could have generated alone, resulting in each partner having a more enriched knowledge structure.

These two criteria of collaborative in our definition of the *Interactive* mode (mutually-and-reciprocally generative) are consistent with many general definitions of collaboration as transactive dialogs in the literature, such as Damon (1984) and Hogan, Nastasi, and Pressley (1999), as well as the construct of dialogical reasoning, in which each individual of a dyad is listening to and considering the views of others, preferably in a way that adds to and elaborates upon what the other speaker is saying, and also the construct of collaborative knowledge-building (Brown & Campione, 1994; Scardamalia & Bereiter, 1996). However, our definition of co-generation in the *Interactive* mode is simpler and does not impose the additional criteria of requiring that the dyads arrive at shared knowledge or negotiate to converge on a shared understanding (such as an agreed upon consensus) and shared representation of the materials (Rochelle, 1992), nor the requirement that new constructs must emerge from ongoing interactions. Again, our definition of co-generation is based on comparing each speaker's contribution with both the instructional materials (i.e., whether it is generative or not in adding new knowledge), and comparing each speaker's contribution with the partner's preceding contribution, in terms of whether it adds further knowledge to what has already been contributed.

Under our simpler definition, we can consolidate all types of collaborative interactions requiring mutually and reciprocally generative dialog as more or less beneficial to the same degree. Therefore, we do not discriminate among the benefits of different patterns of co-generative dialogs, such as disputational talk, cumulative talk, or exploratory talk (Mercer, 1996). Moreover, our knowledge-change view of productive student-to-student interactions, as defined by the pattern of co-generating of knowledge by both partners, focuses on this pattern of interactions as a source of beneficial learning outcomes, rather

than on the conditions of interactions, such as the nature of the task or the competency levels between two peers.

3.1.5. Summary of the taxonomy and terminologies

We have defined cognitive engagement in an operational way based on students' behaviors and products. In our theory, engagement is the process of interacting with instruction or instructional materials. Therefore, engagement is cognitive irrespective of the assigned mode of a particular activity. That is, it is not the case that "hands-on" or manipulative activities are not cognitive. Cognitive processes do occur; it's a matter of which cognitive processes are occurring for "hands-on" activities compared to other "minds-on" types of activities.

Despite our speculations of what the underlying cognitive knowledge-change processes may entail while students interact with the learning materials, our definition of engagement is pragmatically based on the absence or presence of overt, observable behaviors, as well as whether the products contain information that went beyond the instructional materials (rather than based on identifying the underlying thinking processes per se). Defining engagement in this way is practical because (a) such overt behaviors and products can be more clearly *operationalized* by a straightforward comparison with instructional materials, as opposed to relying on judging the covert cognitive processes that may underlie the overt behaviors, (b) such definitions are more *concrete* for teachers to rely upon when they design lesson plans, and (c) such definitions, based on overt behaviors, allow teachers to more easily *detect* in situ whether students are appropriately engaged, and finally (d), behaviors and products are *malleable* and more easily elicited (consistent with the current view that engagement is malleable; Gobert et al., 2015; Reschly & Christenson, 2006).

The taxonomy should be conceived of only as a guideline for differentiating various active learning activities. Many engagement activities may not fall neatly into one or another mode, but perhaps in-between two modes, when the products/outcomes are considered, as these two examples show. If we suppose students are integrating the information presented in two different sources, are they being generative or just manipulative? Their engagement mode can be determined more accurately from the outcome of their integration. If the result of integration is some new ideas that can be inferred from the integrated understanding, then that would be generative. On the other hand, if the student merely concatenated two pieces of information, one from each source, then it is more likely just manipulative. A second example is the following. We have defined the manipulative/*Active* mode as requiring motoric manipulation, thereby causing attention to be focused on the manipulated aspects of instruction/instructional materials. Suppose a student reads a particular sentence out-loud from a long text passage. This is easily classified as *Active* since it involved motoric speaking of a selected sentence, thus entailing attention to be focused on it. However, if a student reads the entire passage mindlessly, in the sense of not adding greater emphasis or intonation to any portions of it, then this may fall closer to the *Passive* mode since such reading does not lead to focused attention on any specific part of the text. From a practical perspective, it does not matter how accurately

an instructor classifies an activity that s/he has designed; it is more important for an instructor to know how to design or upgrade an activity to a higher ICAP mode if possible.

Finally, our taxonomy introduces general engaging processes as knowledge-change processes, which refer to how knowledge is changed while learning. These general knowledge-change processes are applicable in all instructional contexts, and they differ from the encoding processes of specific learning tasks that are unique to the tasks. For example, both reading a text or listening to a lecture are learning tasks and they have unique encoding processes specific to them. Reading involves taking in the presented information through decoding the visual word and interpreting the meaning of a word through lexical access, then storing a representation of sentences that were read; while listening to a lecture requires the processes of segmenting sound waves and so forth. Thus, encoding processes underlying learning tasks such as reading a text or listening to a lecture are unique to the tasks, whereas engagement processes, by our definition, are general processes of changing knowledge once information is encoded. This suggests that engagement processes are applicable and generalizable to all types of learning contexts because they are relevant to how knowledge is changed while engaging, thereby allowing ICAP and its hypothesis to be applicable across domains, age groups, and student activities.

3.2. The hierarchical predictions of the ICAP hypothesis

The four ICAP modes have a hierarchical relationship behaviorally in that collaborative behavior requires that each student is individually generating with his/her partner's contributions; being generative often requires that students are physically manipulating, and being physically manipulative requires that students pay attention to the content being manipulated. In this way, one mode subsumes another mode in the following order: *Interactive* subsumes *Constructive*, which then subsumes *Active*, which then subsumes *Passive*, suggesting a hierarchical relationship in the order of learning outcomes, such that $I > C > A > P$.

However, the plausible knowledge-change processes that correspond to each of the four modes and the resulting knowledge structures can serve as a more informative metric for deriving the same hierarchically subsuming relationship. That is, being collaborative requires that partners generate inferences from his/her own knowledge, as well as from his/her partner's knowledge, so that all four elementary processes of *infer-from-own and infer-from-other*, *activate*, *link and store*, occur. Being generative alone means a student engages in the processes of *infer-from-own*, *active*, *link and store*, without the benefit of being able to *infer-from-other's* knowledge. Being manipulative engages the processes of *activate*, *link* and *store*, while being attentive engages the processes of *store* only, often without *linking* reuslting in more enriched knowledge structures. From this elementary knowledge-change processes perspective, each mode engages a different set of knowledge-change processes, resulting in either enriched, elaborated, embedded, or encapsulated knowledge structures, suggesting that the same $I > C > A > P$ hierarchical relationship holds. Thus, from both the behavioral perspective and the cognitive knowledge-change processes perspective, the operational definition provided in the taxonomy

generates the $I > C > A > P$ hypothesis in terms of students' levels of learning, with paying attention achieving the lowest level of learning, despite it being often regarded as an adequate way to engage (as shown above in our review of the literature), and collaborating as potentially achieving the highest level of learning, but only if students are collaborating in a mutually-and-reciprocally generative or co-generative way.

An important caveat noted earlier but worth repeating again is that ICAP makes predictions about the level of learning as a function of different modes of activities. It cannot make accurate predictions comparing activities within the same mode, such as comparing forming a hypothesis versus deriving similarities and differences (both in the *Constructive* mode), or comparing repeating a procedure versus rotating an object (both in the *Active* mode). To make differentiation for activities within the same mode requires more fine-grained analyses of the processes needed for each type of activities, not just the knowledge-change processes. Moreover, other factors may come into play. For example, when students rotate an object to observe it, other features of the object may become more salient, thereby providing extra information, whereas when students repeat a procedure, no new information may present itself. In short, the postulated knowledge-change processes are only adequate to differentiate the ICAP mode at which students are engaged with a specific task, not the relative advantage of one task over another task within the same mode.

3.3. Research-based support for ICAP

As described in Chi (2009), Chi and Wylie (2014), Fonseca and Chi (2011), and Menekse et al. (2013), ICAP's hierarchical predictions are supported by hundreds of studies in the literature when mapping the various conditions onto the ICAP modes, allowing for pair-wise comparisons of one ICAP mode with another mode. For example, using the same task of concept mapping, a study that compared concept-mapping alone (that can be interpreted as typically a generatively/*Constructively* engaging activity) and concept-mapping with a peer (a collaboratively/*Interactively* engaging activity), shows that doing so collaboratively enhances learning more than doing it individually, providing support for the $I > C$ pairwise comparison of the $I > C > A > P$ hypothesis. To avoid repetitions, here we only describe two studies focusing on very young children's learning that were not cited in our prior papers.

A classic finding in emergent literacy shows that toddlers whose parents not only read to them, but also engage them in "non-immediate talk," is related to children's later performance on measures of vocabulary, story comprehension, and definitions. "Non-immediate talk" was defined in a way that connotes generative/*Constructive* engagement: It is talk that "goes beyond the information contained in text or illustrations to make predictions; to make connections to the child's past experiences, other books, or the real world; to draw inferences, analyze information, or discuss the meaning of words" (De Temple & Snow, 2003, p. 19). Thus, the benefit of parents reading to a toddler and engaging in non-immediate talk can be easily understood from the ICAP's perspective when focusing on the toddler participation: Toddlers engaged in non-immediate talk were being generative/*Constructive*, whereas toddlers who were just read-to were being

attentive/*Passive*, supporting the $C > P$ part of the $I > C > A > P$ hypothesis. Note that the mapping of interventions in published results to ICAP modes often requires that we transfer the focus of an intervention in terms of what the *learners* did (the toddlers in this case), rather than what the instructors/experimenters did (the parents in this case). So in the toddler study, the mapping is based on what the toddlers did, and not based on what the mothers did.

A second easily interpreted study is by Legare and Lombrozo (2014). Very young children were shown five interlocking gears with a crank that turns a fan. Among the gears is an irrelevant "topper" piece. Five-year-olds participated in one of two conditions. In the "watch" condition, they merely watched as the experimenter turned the crank-fan for 40 seconds. In the "explain" condition, the children had to tell the experimenter "how this works" while the crank is being turned for 40 s. Neither group received any feedback. The results show that during assessment, the "explain" group was significantly better at giving an explanation of the causal-functional relationship between the crank and the gears to allow the machine to work, and it was also substantially better at excluding the irrelevant non-functional "topper" piece in their explanations. Other analyses in this study bear out the same pattern of superior results for the "explain" group. The results support ICAP's prediction in that the "explain" group was explaining while watching, thus being generative/*Constructive*, whereas the "watch" group merely watched, thus being attentive/*Passive*, supporting the $C > P$ comparison within $I > C > A > P$.

3.4. General summary and further clarification of ICAP

ICAP is a parsimonious and comprehensive theory that defines how students can engage with instructional materials cognitively, in a concrete and explicit way that is generalizable across learners' age, content domain, and context (e.g., teachers learning in the context of professional development, college students learning in the context of a lecture hall, middle school students learning in a science class, or in afterschool and other informal contexts, and toddlers learning from their parents). Moreover, because ICAP focuses on what the learners do to engage, we can un-confound and separate out the role of the instructor from the role of the learners. That is, what learners do can be defined independently from how instructors teach. For example, it is not necessarily the case that an instructor lecturing implies students must learn *passively*, as dictated by the common definition in the literature, with two empty cells in the previously mentioned 2×2 perspective, crossing lecturing or not with active or passive learning. It is possible to couple lecturing with prompts or questions to encourage and elicit higher ICAP modes of engagement. Finally, ICAP introduces the idea that different modes of interacting with the learning materials lead to different levels of learning outcomes, with the collaborative/*Interactive* mode having the potential of producing the deepest learning with possibility of discovering innovative knowledge, followed by the generative/*Constructive* mode. The ICAP hypothesis was generated on the basis of the subsuming hierarchical relationships of both the behaviors and the hypothetical knowledge-change processes, resulting in knowledge structures varying from enriched to elaborated to embedded to encapsulated.

Confirmation of the ICAP hypothesis by a huge number of studies in the literature suggests that our definitions of the behavior modes are valid.

ICAP can also serve as a tool for various other purposes, as discussed in Chi and Wylie (2014). For example, the ICAP modes can serve as a rubric to code student products, as well as dialog patterns (Chi & Menekse, 2015). However, one important implication of ICAP, not previously discussed, is that ICAP provides an important tentative definition of deep versus shallow engaging activities based on the presence or absence of the knowledge-change process of *infer*. Because both the *Constructive* and *Interactive* modes include the process of *infer*, whereas neither of the *Passive* nor *Active* modes include the process of *infer*, we anticipate that there may be a bigger gulf between *Constructive*-and-*Interactive*, versus the *Passive*-and-*Active* modes. From this knowledge-change perspective, the presence/absence of *infer* can be used to characterize the gulf between shallower (the *Passive* and *Active* modes) and deeper thinking (the *Constructive* and *Interactive* modes).

This gulf also seems to correspond to the "hands-on" versus "minds-on" distinction. Although it is tempting to say that there is a significant divide between manipulative and generative learning outcomes because the former is *physically* active whereas the latter is *mentally* active, this distinction in the popular view is incorrect, as stated earlier, because in ICAP, both manipulative and generative activities involve cognitive knowledge-change processes. See Fig. 1 again.

Most importantly, we need to emphasize that the different levels of learning as a function of different modes of engagement cannot be captured in measures of learning outcomes unless the assessment instrument is deep or sensitive to enriched knowledge derived from inferences. That is, because the manipulative/*Active* mode can cause new information to be encoded with activated prior knowledge, the resulting knowledge can be more complete, easily retrieved, but it only reflects what was taught, so that assessment measures that basically request recall of information are sufficient to assess how well students have encoded the presented information. However, to know whether students have understood the instructional materials more deeply, inferences are required to generate new ideas, and such deeper understanding can only be assessed by deeper questions, questions that assess knowledge that go beyond the information given. Unfortunately, standardized tests typically do not measure deep learning.

In this paper, we take ICAP in a new direction. As Greene (2015, p. 25) has noted, "Although it is difficult to discern implications for practice from this work" on engagement, now that we have a concrete definition of cognitive engagement as provided by ICAP, we agree with her that "the time is ripe for more applied research that would focus on the potential instructonal implications" of engagement. Because ICAP has been well supported by evidence from laboratory and classroom studies, the goal of this translation project is to see if we can teach teachers to understand ICAP well enough so that they can translate their understanding to the successful design and implementation of higher modes of activities, to be verified by students' learning. Thus, students' learning outcomes will be viewed as a reflection of teachers' success at designing and implementing higher ICAP modes, and not as our assessment of the ICAP hierarchy per se.

4. Developing an online module as professional development to introduce ICAP for teachers

ICAP is a theory about how students engage to learn, not how instructors teach. However, as a theory of learning, ICAP can be translated into a theory of instruction in terms of how teachers can facilitate and elicit students' engagement. An ICAP theory of instruction based on eliciting cognitive engagement means that teachers need (a) to learn and understand what cognitive engagement is in terms of ICAP, (b) to know how to design lesson plans that incorporate higher modes of engaging activities, (c) to recognize and know how to design deeper questions to assess student learning, and (d) to know how to implement ICAP-designed classroom activities properly in real time. Proper implementation includes all aspects of the lesson, such as in the instruction they give students, the amount of time they allocate to the activities, the questions they ask to facilitate cognitive engagement, and so forth. Therefore, the goal of our project was to teach K-12 teachers to learn and understand cognitive engagement per ICAP and then see if they can translate their understanding into practice.

Teachers' learning and translation into practice were measured sequentially in unfolding stages in terms of (a) changes in teachers' knowledge, (b) teachers' design of contrasting lessons using that knowledge, (c) teachers' behavioral practices in implementing their lesson plans, and whether teachers' knowledge and behavior are validated by (d) students' enactment, and (e) students' learning. In the next section, we first describe the online module to teach ICAP, then describe the five stages of assessment.

4.1. The ICAP online module

The approach we took to help teachers learn about ICAP was to develop an online module about ICAP to provide teachers as professional development. After several iterations, our resulting online module consisted of explanations that specified student behavioral characteristics for each mode of engagement in a concrete and operationalized way. The module also provided examples of common classroom activities and explained the activities in terms of the ICAP mode they elicited. Exercises embedded in the module asked teachers to improve sample classroom and homework activities according to the ICAP framework and receive some form of feedback.

The ICAP module was developed iteratively over a 3-year period. The first iteration was presented in person with two community college instructors knowledgeable about cognitive psychology. The second iteration was an online version accompanied by a 1-week face-to-face workshop with 11 6th–12th grade teachers recruited via email through local teacher networks. The third iteration was a completely online version, implemented on the WISE platform (University of California, Berkeley, 2016), and given to 40 pre-service teachers. Those teachers were then assessed for their understanding of ICAP. Based on their learning outcomes and informal feedback, revisions were made, resulting in the fourth iteration. This paper describes the findings from our study using the fourth iteration of the ICAP module.

This fourth iteration of the online ICAP module consisted of four sections with comprehension questions embedded throughout the first three. The first section detailed the ICAP taxonomy of student engagement behaviors, outlined the hypothesized knowledge-change processes underlying each mode, and asked teachers to classify various student activities by ICAP mode. The second section presented three common activities (note-taking, concept-mapping, and worked examples) and explained how each could be implemented at different ICAP modes, based on our review of the literature of these common activities (Chi & Wylie, 2014). This section also contained exercises for which teachers were asked to list common learning activities they had used in the past, to classify each of them by ICAP mode, and then to describe how they would "bump up" three of the activities to higher ICAP modes. For example, if a teacher originally had students listen to a lecture on natural selection (attentive/*Passive*), she could suggest bumping it up by having students take guided notes (manipulative/*Active*), write a summary including any questions they had at the end of a lecture section (generative/*Constructive*), or answer a series of discussion questions about the lecture content with a partner after the lecture (collaborative/*Interactive*). Our module was not "intelligent" so teachers did not receive direct feedback on their description of how they would "bump up" their three chosen activities. Instead, teachers were provided criteria that a correct answer would have included and they were asked to self-check their answers regarding how accurately they had "bumped up" an activity to a specific mode.

The third section of the ICAP module contained information on the creation of deeper assessment items since only deeper questions are sensitive to the deeper understanding gained from the *Constructive* and *Interactive* modes. As we explained above, if we only required students to recall some information, then we only need shallow questions that can assess whether students have stored that information in a strong retrievable form, based on manipulative/*Active* engagement. On the other hand, if we need students to have a deeper understanding of the materials, then we need deeper questions that can assess whether students have engaged generatively with the learning materials or co-generatively with a partner, so that their knowledge structure is richer (with added inferences).

During our face-to-face professional development, we discovered that teachers did not know how to create, nor understand what we meant by shallow and deep questions. We had to come up with a new way to convey the idea of shallow versus deep questions, by differentiating easy versus hard questions and recall versus inference questions. An easy question was defined as one for which teachers expected a greater number of students could answer correctly after the lesson, whereas a hard question was defined as one for which fewer students were expected to answer correctly. Recall questions were defined as those that require the students to remember facts or concepts presented during learning, whereas inference questions were said to be those that require students to make inferences or new connections between facts and concepts not presented in the lesson materials. Examples of each question type and writing tips for assessment question design were provided. Finally, teachers were told about the importance of assessing student learning with inference questions, not only for our research purposes, but also for detecting the type of deep learning associated with higher ICAP modes (*Constructive* and *Interactive*).

Therefore, they were instructed that all quizzes and exams should contain some proportion of both easy and hard questions and recall and inference questions. In the module, teachers were asked to create four types of questions: easy recall, hard recall, easy inference, and hard inference; they were provided a list of criteria for recall and inference questions after submitting their responses in order to self-check their work.

The fourth section of the module presented tips for successful implementation gleaned after a previous group of teachers completed a prior version of the module and taught lessons designed to target specific modes of engagement. For example, a list of 12 potential inference question stems, taken from King (1992), such as "What is a new example of. . ., What would happen if. . ., Explain why. . .Why is it important?" was provided to help teachers design *Constructive* questions (i.e., questions that are likely to elicit *generative* responses). It was also suggested that teachers take a less dominant role in the classroom during *Interactive* activities.

Finally, lesson plans, class materials, and student products were provided from two successfully implemented lessons, taught at different ICAP levels, so that teachers could see explicitly how a successful lesson might look. For example, a lesson on tone and mood was designed for an *Active* class, a *Constructive* class, and an *Interactive* class. The lesson plans for each were provided, followed by sample PowerPoint slides, student notes, and worksheets.

In addition to these four sections, pre- and post-tests were developed to assess how well teachers understood the content from all the sections of the ICAP module.

4.2. Using the ICAP module as professional development

The effectiveness of the ICAP module was tested in a charter school in which the module was a required professional development unit for all teachers in grades K-12. There were approximately 135 full-time teachers in the school system and 102 teachers completed at least some of the ICAP module. Forty-three teachers completed all three parts of the ICAP module (the pre-test, the module, and the post-test). Of those reporting demographic information (n = 21), the male to female ratio was 1:6, and predominately white or Caucasian. The average age was 30 years old (SD = 9.2), with an average of 4.6 years of teaching experience (SD = 5.5).

Teachers spent approximately 3 h to complete the module, with a median time of about 3 h and 20 min. This included time spent on the pre- and post-tests, which typically took between 18 and 20 min to complete. A partial mastery learning design was used in four ways. First, the first three sections included embedded knowledge checks and activities that required perfect completion to proceed through the module. A typical knowledge-check activity asked teachers to drag and drop examples of activities into *Passive*, *Active*, *Constructive*, or *Interactive* categories. Second, at the end of first section, teachers also completed an eight-question quiz and received immediate feedback embedded in the online module once they submitted their answers. There was no limit to the number of attempts that could be made; however, learners could proceed to the next question after submitting just one attempt, irrespective to the answers' correctness.

Finally, the third section included 8 quiz questions, scored as before, and 16 short-answer questions that they self-checked via feedback given online once the answers were submitted. All feedback statements were embedded in the module and delivered automatically when an answer was submitted. There were no learning activities associated with the fourth section, as this section was simply intended to provide exemplary cases.

From the 43 teachers who completed all three parts of the ICAP module, 13 volunteered to continue participation in our study. Continued participation meant that they went on to design activities within their lesson plans according to the ICAP framework (as taught in the ICAP module) and taught those lessons as planned. These 13 teachers taught a variety of subject matters for grades 7–11.

Following their completion of the online ICAP module, 7 of the 13 teachers created lesson plans on two topics, and the other 6 teachers created lesson plans on three topics, for a total of 32 topics. These 32 topics ranged from science (10), math (8), language arts (8) and foreign language (6). Teachers then created two variations or "paired-lesson plans" for each topic to correspond to two ICAP modes of the teachers' choosing. One teacher created a triple lesson plan for one topic. Thus, a total of 65 lesson plans were created for the 32 topics. These 65 lesson plans were distributed across the ICAP modes in the following ways: 2 in the *Passive* mode, 19 in the *Active* mode, 21 in the *Constructive*, and 23 in the *Interactive* mode, in the sense that they were designed for classes that the teachers intended to be at their pre-specified mode. Appendix A tabulates all the data we are analyzing for this paper, and Row 1 shows the distribution of the lesson plans across modes.

The two variations of each paired-lesson plan differed mostly in the activities within the lesson plans. Some activities within each paired-lesson plans were similar/common, while others were different/unique, when compared with each other. The unique activities were those intentionally manipulated by the teachers to correspond to their chosen ICAP mode, whereas the common activities were shared in both of the paired-classes. For example, paired-classes might have received the same lecture (common activity), but it may have been followed by an *Active* activity in one class and a *Constructive* activity in the paired-class, and these two activities would be unique. Thus, the unique activities should correspond to the ICAP mode the teachers had intended for each of the paired-classes. The teachers also designed one set of pre–post questions for each paired-lesson plans since the paired lesson plans are on the same topic.

The teachers sent their lesson plans to two members of the research team, who then reviewed the plans and gave teachers feedback within three business days. There were no instructions or scripts that the researchers followed with respect to what feedback to give or how often; feedback was based on how well the researchers judged that teachers' design choices adhered to the desired ICAP mode. The researchers gave both positive and negative feedback about the lesson designs and assessment questions. Some teachers received feedback only once, whereas others received it twice. The feedback was brief, consisting of email responses, such as asking teachers to keep students in non-interactive conditions from doing group work, asking teachers to have stronger directions in *Active* conditions to prevent *Constructive* engagement, or requesting teachers to include more

deep assessment items on their pre- and post-tests. As a result of the feedback, the intended ICAP mode for 6 out of 65 classes were changed.

After receiving this feedback, the paired-lesson plans were taught to two different classes, to be referred to as paired-classes. For instance, the activities within a lesson on trigonometry ratios were manipulated so that one class would be taught at a manipulative/*Active* mode, and the other class at a generative/*Constructive* mode. The 13 teachers implemented 31 paired-lesson plans in 31 paired-classes (or 62 individual classes), with one lesson plan manipulated at three different variations of ICAP modes and implemented in three classes. In total, teachers implemented 65 lesson plans in 65 classes, involving 719 students, with some students participating in more than one class.

To summarize, teachers took the ICAP module as professional development and then designed pairs of lesson plans with each variation within a pair corresponding to a different ICAP mode (with the exception that one teacher designed three variations). They then implemented each of the paired-lesson plans in two classrooms (creating paired-classes), corresponding to the ICAP mode by which each of the paired-classes was designed. Each of the paired-classes was assessed by the same pre- and post-test questions because they were on the same topic, but differed only in the ICAP mode. In this way, teachers are transferring their understanding of ICAP into the design of their lesson plans according to specific ICAP modes of their choosing and the implementation of their lesson plans.

5. Effectiveness of the ICAP module and teachers' implementation

Five sets of findings will be reported here that assess (a) how well teachers understood and learned the ICAP module content, (b) how well teachers were able to transfer their learning into the design of lesson plans, (c) the fidelity with which teachers implemented ICAP in their classes, (d) how the students enacted the activities compared to teachers' intentions regarding engagement modes as revealed in students' products, and (e) students' learning outcomes as a function of the intended ICAP modes. For these analyses, "intended mode" was defined as the ICAP mode teachers indicated on their lesson plans at which each class would be taught. For clarity, henceforth, classes are indicated as *Active*, *Constructive*, or *Interactive* based on teachers' intention, and students' behaviors are indicated as manipulative, generative, or collaborative.

5.1. Teachers' learning outcomes from the ICAP module

Teachers' understanding of ICAP was assessed from (a) the pre- and post-test data of the 43 teachers who completed all three components of the module, and the (b) misunderstandings revealed by the 13 teachers who actually implemented ICAP in their classrooms.

5.1.1. Learning as assessed by pre- and post-tests

Learning for the 43 teachers was assessed by 46 questions on the pre- and post-test, and their average proportion correct is shown in Fig. 2 (left-most set of bars). Overall

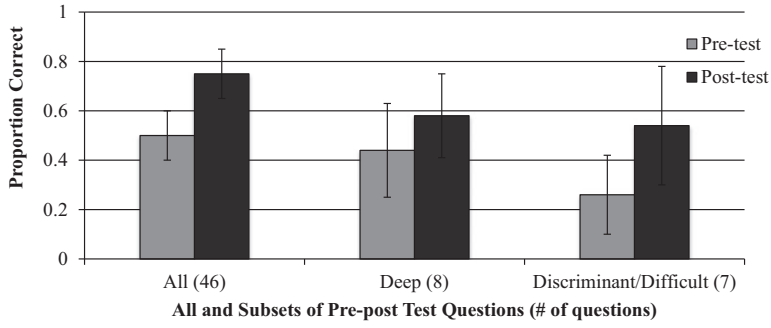

Fig. 2. Teachers' learning of the ICAP module assessed by all, deep, and discriminant pre- and post-test questions.

performance was 0.50 (SD = 0.10) at pre-test and 0.75 (SD = 0.10) at post-test. A repeated measures ANOVA model showed significant overall learning from pre-test to post-test, $F(1, 42) = 2{,}375.6$, $p < .001$, $\eta_p^2 = 0.983$.

Since only 43 teachers completed all three parts of the ICAP module (the pre-test, the module, and the post-test) and only 13 participated in our study, we needed to explore whether teacher performance on the pre-test assessment influenced their decision to (a) fully complete the module or (b) to participate in the teaching aspect of our study. We analyzed teachers' pre-test scores to determine if there were significant differences between different teacher "types," where types represented teachers' level of participation, for example, completed pre-test only, completed pre-test and module only, completed module and taught, for a total of five subgroups. There were no differences in pre-test score as a function of participation level ($F(4, 92) = 0.878$, $p = .480$). This result tells us that teachers' performance at pre-test did not influence their decision to participate in further aspects of our study.

Although overall improvement from 50% correct at pre-test to 75% correct at post-test seems impressive, for a more complete understanding of pre/post-test performance we conducted two post hoc analyses. Eight items from the post-test that tapped deep (vs. shallow) knowledge of the ICAP framework were selected for this first post hoc analysis; items were identified subjectively based on the extent to which they tested the participants' understandings of the nuance of the ICAP framework. For instance, one question asked the participants what level of engagement would "students learning definitions from flashcards" be classified, according to ICAP. From an ICAP perspective, studying flashcards is *Active* because students can pick and sort the cards into different piles; that is, students could manipulate the cards and focus attention on each one as it is being picked up and sorted. For these eight deep items, the pre-test proportion correct was 0.44 (SD = 0.19), and the post-test proportion correct was significantly higher at 0.58 (SD = 0.17), $F(1, 42) = 13.831$, $p = .001$, $\eta_p^2 = 0.248$ (see Fig. 2, middle columns).

Because it is possible that the items in our pre- and post-test were not well written, an additional post hoc analysis was conducted, in which seven items were identified that

proved to be both fairly difficult (item difficulty $p < .66$) and good discriminators of teachers' ability (discrimination index > 0.30). Pre-test performance on these seven items averaged a proportion correct of 0.26 ($SD = 0.16$), and at post-test 0.54 ($SD = 0.24$); the increase in teachers' scores was significant $F(1, 42) = 37.984$, $p < .001$, $\eta_p^2 = 0.475$ (see Fig. 2, columns on the right).

In sum, although teachers appeared to have understood ICAP when assessed by all the pre–post questions well (with 75% correct at post-test), the harder questions showed that they were only correct 54% to 58% of the times, because many of the questions were definitional only, assessing their understanding of the ICAP labels.

5.1.2. Teachers' misunderstandings

Because teachers' learning was more modest, we needed to know what misunderstandings they have about ICAP. We identified the majority of teachers' misunderstanding in the following subjective way. Six laboratory members studied the 65 videos of the 13 teachers' implementation. They then discussed, based on their subjective impressions, what they thought they saw was teachers' misunderstanding, and these impressions were consolidated into a list of nine misunderstandings, categorized according to (a) misunderstanding about the ICAP modes or (b) misunderstanding about implementation. For example, teachers sometimes think that physically producing a product (such as a concept map) is generative/*Constructive*, without realizing that they need to examine the content of the concept map to see if new knowledge was actually produced. Five other misunderstandings were culled from other sources of data (such as the feedback researchers gave to teachers' design of their lesson plans) for a total of 14 distinct misunderstandings. These misunderstandings confirm the interpretation of more modest learning as assessed by the harder pre- and post-test questions.

In summary, based on all the items in our post-test, the results suggest that teachers have learned quite a bit about ICAP from our module (with 75% correct at post-test). However, using either subjectively selected harder questions or psychometrically selected discriminant questions, both analyses show that teachers' learning was much more modest, around 54% to 58%. This suggests that they did not have a nuanced understanding of the ICAP theory, confirmed by our subjective collection of 14 distinct teachers' misunderstandings from the videos of their implementation and other sources.

5.2. Teachers' design of the lesson plans

We next examined how well teachers designed their lesson plans. Although two researchers did give teachers feedback on their design, as described above, that feedback was cursory and given in real time with short turn-around time, whereas here, we present detailed in-depth analyses of their lesson plans. Most lesson plans included at least some of the following: learning objectives, class procedures, general descriptions of activities, activity worksheets, time allotments for activities, scoring guides, scans of in-class readings, PowerPoint slides, oral scripts, URLs to online activities, and expected student behavior. Because of this wide range and variety of instructional materials, it was

impossible to compare lesson plans on all these dimensions. However, because all lesson plans contain in-class activities and assessment questions, we therefore focused on them as the sources of our data analyses.

5.2.1. Teacher-designed lesson activities

Each lesson plan contained on average three activities. As stated above, for each paired-lesson plan, there were some activities that were common across the paired-classes and some activities that were unique between them. The activities and materials targeted at the two paired-classes are easily distinguished within each paired-lesson plan.

Because only 2 of the 65 teacher-designed lesson plans were intended to be *Passive*, we excluded them from our analyses. In addition, 11 of the lesson plans were excluded due to incomplete lesson plan materials, leaving a total of 52 lesson plans that were analyzed. Of these 52 classes, 14 were intended to be *Active*, 18 intended to be *Constructive*, 20 intended to be *Interactive* (see Appendix A, Row 2).

Within this set of 52 lesson plans, a total 105 unique activities were designed (2.02 per class, *SD* = 1.18), 29 intended for the *Active* classes, 29 intended for the *Constructive* classes, and 47 intended for the *Interactive* classes, so the greatest number of unique activities were designed to be intended for the *Interactive* mode (see Appendix A, Row 3). A total of 26 common activities were designed (1.00 per class, *SD* = 1.09) with each common activity used twice (i.e., in both paired-classes, resulting in implementing 52 common activities, distributed across the three modes of ICAP classes, as shown in Appendix A, Row 4). Three analyses will be reported with respect to teachers' ability to design ICAP mode-appropriate lesson plans.

5.2.1.1. Written instruction for common and unique activities within lesson plans: Activities embedded in lesson plans vary substantially from one activity to another activity, so how can we compare and assess them? The only part of activities that is available for all activities is the written instructions (or "directives") provided on how to carry out the activity worksheets. Consequently, the written directives for an activity became our source of data. Again, how do we code such activity directives? We arrived at a novel way to assess teachers' worksheet directives, which is to analyze the verbs or verb phrases in the written instruction. That is, each directive for an activity could be segmented based more-or-less on a verb, such as "move," "label," "measure," "sketch," "compare," or "determine," along with a noun phrase. For example, directive segments for activities might say "Circle the best choice" or "Use the rule to help you find the answer to these problems." "Use the rule to help you find the answer to these two problems" would be coded as a manipulative/*Active* directive because students were asked to apply a rule that was already given. Coding included the noun phrase because the noun phrase, which provided the context, can sometimes map the verb more accurately. For example, "connect" two ideas is generative/*Constructive*, whereas "connect" two nodes is more likely manipulative/*Active*.

Out of the 105 unique activities that teachers had, 88 of them had written directives, and these 88 directives were segmented into 232 segments (see Appendix A, Row 5 for

the distribution of these segments across the intended class modes). Out of the 26 common activities that teachers had designed that were identical across the paired-classes, 17 had written directives; and these 17 directives were segmented into 46 segments (see Appendix A, Row 6). There was a total sample of 278 directive segments (46 + 232) across the sample of 105 (88 + 17) unique and common activities with written instructions.

The first result shows our approximate classification of the variety of distinctive verbs that was culled from the sample of 278 directive segments, determined by the verbs alone, without the noun phrase. Appendix B displays five alphabetized lists of verbs that were used by the teachers in this sample of 278 directive segments, categorized into ICAP modes. Surprisingly, there were more than twice as many distinct manipulative/*Active* verbs (58) as there were generative/*Constructive* verbs (28) used, with *Interactive* verbs being the fewest (9). This suggests that teachers have the largest repertoire of manipulative verbs. Notice that even though *Interactive* activities were the most prevalent in that teachers had designed 47 of them (vs. 29 for *Active* and 29 for *Constructive*, see Appendix A, Row 3 again), and collaborative directives were the most prevalent with 92 segments of them for *Interactive* classes (vs. 66 segments for *Active* and 74 segments for *Constructive* classes, see Appendix A Row 5), there was a very small number of distinct collaborative verbs used in teachers' instruction. Because none of these classes were intended to be *Passive*, it is comforting to see that few attentive verbs (6) were used; and because only two classes were designed intentionally as *Passive*, we cannot judge the relative frequency (6) of the *Passive* verbs as compared to the verbs in the other modes. There were also 10 vague words that could not be categorized.

Having a large repertoire of manipulative verbs available does not necessarily imply that teachers were more likely to use them. Our second result examined whether teachers were more likely to use manipulative verbs for their activities by counting the frequency with which each type of verb mode was used in the 278 written directive segments for both the unique and common activities. Using the verbs and the associated noun phrase, each segment of instruction was coded as mapping on to a specific ICAP mode. Three coders separately coded 20% of the written directive segments, and the Krippendorff's alpha reliability was 0.61, 95% CI [0.47, 0.74]. Reliability is lower than Krippendorff's (2004) suggested 0.667 because the noun phrases provided more variability in interpretations. The three coders then met and resolved disagreement through discussion. After that, each of the three coders individually coded one-third of the remaining written instruction segments.

Fig. 3 shows the pattern of distribution of the ICAP modes of verbs teachers used in written directives. The finding shows that teachers had a tendency to provide manipulative directives most frequently for both unique and common activities, and collaborative directives least frequently.

We have established that teachers prefer to give manipulative directives in general. The question now is, how "accurate" were the teachers in designing their unique activities. By accurate, we mean whether the ICAP mode of their written directives matched or corresponded to the mode they had intended for each of their classes. For example, for

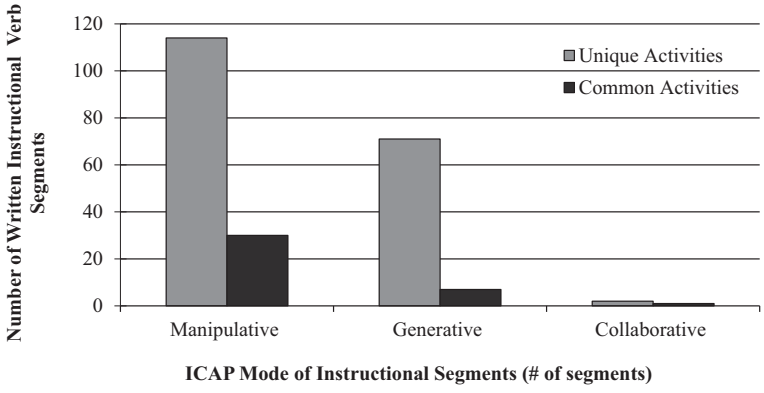

Fig. 3. Usage frequency of verbs mapped into ICAP modes for unique and common written directives.

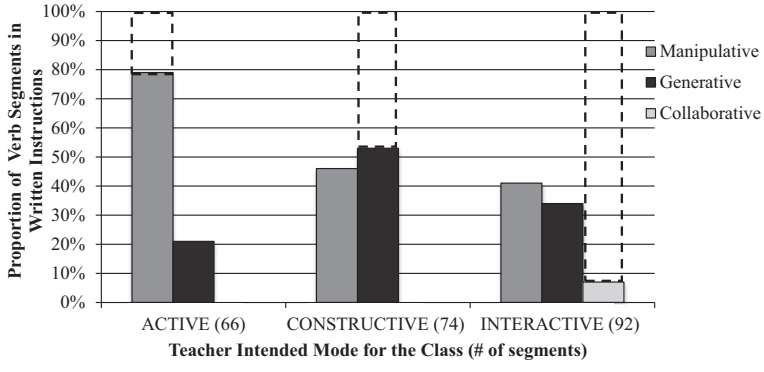

Fig. 4. Teacher design: Verb segments in written directives of unique activities.

an intended *Constructive* mode of a class (e.g., a class designed to be a *Constructive* class), we would expect that the majority (ideally close to 100%) of the directive segments would correspond to that generative mode. For this analysis, we analyzed accuracy for the unique activities only because they were intentionally manipulated by the teachers to vary across paired-classes. As shown in Appendix A, Row 5, there were a total of 232 directive segments for the unique activities, and each verb segment was coded as either *manipulative*, *generative*, or *collaborative*.

Fig. 4 shows the distribution of the verb segments in written directives of unique activities for each mode of class. For classes intended to be *Active*, the written instruction appropriately used manipulative verbs 79% of the time, and generative verbs 21% of the time (the difference is highly significant, $\chi^2(1, N = 66) = 21.879$, $p < .001$), with no interactive verbs used. This suggests that *Active* classes were designed more or less accurately because their written directives were accompanied predominantly by mode-appropriate manipulative verbs (i.e., 79% of the verbs used were manipulative, albeit ideal would be 100%, as shown by the dotted line in Fig. 4). For *Constructive* classes,

however, only 53% of the written directives used the appropriate *generative* verbs (as opposed to the ideal of 100%, as shown by the dotted line), with 46% being *manipulative* verbs (with no significant difference, $\chi^2(1, N = 74) = 0.216$, $p = .642$). Similarly, for *Interactive* classes, only 7% of the written directives used *collaborative* verbs, which was significantly fewer than manipulative verbs ($\chi^2(1, N = 44) = 23.273$, $p < .001$) and significantly fewer than generative verbs ($\chi^2(1, N = 38) = 17.789$, $p < .001$). In short, only the written directives for activities in the *Active* classes were designed somewhat accurately, in the sense that the largest proportion (79%) of the instructions were written in a mode that corresponded to the intended mode of the classes, significantly greater than the alternative directive modes. The *Interactive* classes were the least appropriately designed, in terms of the proportion of collaborative directives given, which were significantly fewer than the other two modes.

5.2.1.2. Actual mode elicited by questions within worksheets: The preceding analyses coded the verbs used in the written directives given for an activity. A more sensitive measure of the accuracy of a designed activity in terms of its correspondence to the intended ICAP mode of a class is to code each question embedded in an activity (for activities that can be coded that way). There were 14 worksheets from the unique activities that had multiple questions that could be coded this way, and Appendix A Row 7 shows their distribution across the classes. Because collaborative instruction was typically given only in the overall directive for an activity rather than in the context of individual questions, so no question was coded as collaborative. Therefore, each question within these 14 worksheets was coded as eliciting either manipulative or generative responses. Manipulative questions were ones in which the expected answer can be found in the learning materials; this included combining multiple parts of the learning materials. For example, if the question asks, "What does the Shaman do to identify witches?" and the answer was found in the reading, then this was coded as a manipulative question. Generative questions were ones for which the expected answer cannot be found in the learning materials; to answer correctly, students were required to generate inferences, connections, or new knowledge.

Fig. 5 shows the proportion of manipulative and generative worksheet questions for the intended ICAP mode of the classes. Again, a similar pattern of results is obtained when we coded at a more fine-grained level by the mode of each question on a worksheet, as when we coded the verb phrases in the written directives for each worksheet (see Fig. 4). That is, the *Active* classes were designed the most accurately in that 66% of their worksheet questions were manipulative, requesting information that was already provided in instruction, compared to 33% of the questions were generative, and the difference is significant, $t(8) = 2.404$, $p = .043$, $d = 1.416$. For the *Constructive* classes, 54% of the questions in the worksheets were manipulative, and 46% were generative and the difference was not significant, $t(4) = 0.827$, $p = .827$, $d = 0.190$. For the interactive classes, 37% of the questions were manipulative and 63% were generative, and the difference was not significant, $t(10) = 2.049$, $p = .068$, $d = 1.183$.

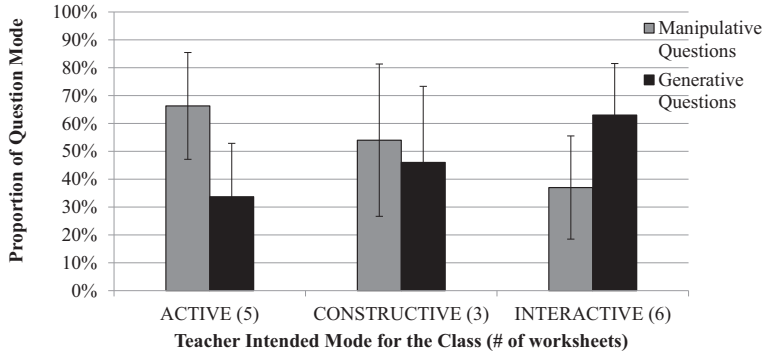

Fig. 5. Teacher design: Questions embedded in worksheets.

Overall, Figs. 4 and 5 show that worksheets were not designed uniformly to reflect the ICAP mode of the class, and teachers were most accurate with *Active* classes. For both the *Constructive* and the *Interactive* classes, there were comparable number of questions eliciting manipulative and generative responses. Thus, the similarity in the pattern of the results for both a coarse-grained coding of the written directives for activities and at a finer-grained coding of individual questions within an activity suggests that teachers are biased to elicit manipulative responses from students.

5.2.2 Teacher-designed assessment questions

As part of designing their lesson plans, the teachers were also responsible for developing the assessment items used to measure student learning for each topic. The assessment questions were identical for each set of paired-lesson plans on the same topic, allowing us to compare students' performance between the paired-classes.

Section 3 of the ICAP module both explained to teachers the need to assess students with harder inference questions and demonstrated how to write them. It is paramount that teachers design some proportion of the assessment questions to be inference questions, since the prediction of the ICAP hypothesis is based on deep understanding. We stated above that because teachers had difficulty understanding the difference between shallow and deep questions, we changed our instruction in the module to distinguish easy from hard and recall from inference questions. Easy/hard was defined in terms of how many students they think can answer the question correctly. We anticipated that easy/hard was an easier definition for teachers to understand, since they had experience with questions in terms of how many students got them right, whereas they may have more difficulty discriminating recall/inference.

The 13 teachers designed a total of 263 post-test questions for the 32 lesson plans, averaging 8.2 questions per lesson plan. For the majority of these post-test questions, they were supposed to indicate for each question whether it was easy/hard and recall/inference. But not every teacher specified the category for each question. Of the ones labeled, 178 of them were labeled as easy/hard, and 166 of them were labeled as recall/inference. The

two sets of questions overlapped but are not identical. The accuracy of their question type was determined in the following ways: For the easy/hard ones, the analysis was based on students' actual performance on the post-test questions, using a median split as the criterion for whether a question was in fact answered correctly or not (thus easy or hard). For the recall/inference questions, two researchers determined whether the questions were recall or inference, based on the instructional materials teachers provided.

Table 1 shows that there were great agreements between the teachers' judgment of whether questions were easy or hard with actual student performance. For example, 82% of the easy questions were in fact answered correctly by students, and 74% of the hard questions were not. Table 2 shows that researchers also agree by and large with the teachers' designation of a recall question 87% of the time, and an inference question 72% of the time.

Tables 1 and 2 show that the percentage of agreement between the teacher-designed questions and the researchers' judgment was similar to the percentage agreement for easy/hard and actual performance. This suggests that teachers understood the distinction between recall and inference questions. Although we had expected teachers to be familiar with the concept of easy versus hard questions based on their classroom experiences, we were surprised that they understood the distinction between recall and inference questions, since our earlier face-to-face PD with a different group of teachers showed that they had difficulty understanding the distinction between shallow and deep questions. We wondered whether this discrepancy between what we expected them to be able to understand and how they did perform on the recall and inference questions might be due to our provision in the fourth section of the module 12 question stems taken from King (1992) (e.g., "What do you think causes...?" and "Why is ____ important?"), which might have helped them design inference questions accurately.

In summary, our coding and analyses of how accurately teachers designed their lesson plans after completing the ICAP module, in terms of (a) their written directives for the activities, (b) the specific questions within the activity worksheets, and (c) their assessment questions, show two general findings. First, on the whole, teachers seemed to be most accurate in designing manipulative activities for *Active* classes, and least accurate in designing collaborative activities for *Interactive* classes, based on their usage of verbs in their written directives. Second, teachers were relatively accurate in designing assessment questions, using either the easy/hard or recall/inference descriptions, although accuracy in

Table 1
Percentage of agreement between teacher-designed easy and hard questions and students' actual performance

Teacher-Designed Question Types (178 total)	Students' Actual Performance	
	Easy	Hard
Easy (n = 108)	82%	18%
Hard (n = 70)	26%	74%

Table 2
Percentage of agreement between teacher-designed recall and inference questions and researchers' coding based on design

Teacher-Designed Question Types (166 total)	Researchers' Re-coding	
	Recall	Inference
Recall (97)	87%	13%
Inference (69)	28%	72%

the design of recall/inference questions could have been influenced by the availability of question stems that we had provided.

5.3. Fidelity of teachers' implementation

The above analyses assessed the accuracy of teachers' design of the content of their lesson plans. Design is an activity or task that teachers can carry out prior to a class. In this section, we analyzed the success of teachers' implementation in real time as captured in class videos. The videos captured the entire class time, which was generally 50 min long for junior high students and 90 min long for high school students. We distinguished between instructional tasks and miscellaneous, non-instructional tasks. Non-instructional tasks included taking the pre- and post-tests, transitioning between two portions of the class, assigning students to groups, collecting worksheets, and so forth. These miscellaneous tasks occupied up to 20% of the class time and were excluded from the subsequent three analyses.

The fidelity of teachers' implementation was analyzed for the 63 non-*Passive* classes in three ways. First, we examined the amount of time allocated to common and unique activities, to verify that more class time was in fact devoted to the ICAP-mapped unique activities. This is essentially a dosage analysis (see Stump et al., 2018, for description of the coding process). There would be no opportunity for us to see the predicted ICAP contrast in students' learning unless more time was spent on the unique activities. Second, we coded the ICAP mode of the oral directives for the unique activities. Third, we examined whether the assessment questions designed for each lesson were implemented in a way that is consistent with the intended ICAP mode.

5.3.1. Dosage: Amount of class time spent on common and unique activities, from the videos

The lesson plan analyses show that a total of 157 activities were implemented (105 unique and each of the common activities was implemented twice for 52 common activities). Because there was an average delay of 10.8 days between design and implementation, we must re-identify again what were the actual implemented common and unique activities, to know the proportion of instructional time devoted to them, as well as the

engagement mode elicited during that time, before we can accurately assess whether teachers' implementation of various ICAP modes had any impact. Although teachers were not told to do this explicitly, we had hoped that more time would be devoted to the unique activities since they were the manipulations that constituted the intervention, compared to the common activities. If teachers had devoted a larger proportion of class time to common activities that were not intended as an ICAP manipulation, then there may be very little effect of the intended ICAP mode mapping on the class as a whole.

For each video corresponding to a class, we identified the existence of each activity based on the presence of oral directives. The oral directives were transcribed and time coded, then used to demarcate the beginning and end of an activity (i.e., its boundaries). A total of 168 activities were identified across the 63 videos (excluding the 2 *Passive* classes), very close to the 157 identified in the lesson plans.

We then determined if an activity was unique or common based on the communicative intention of the oral directives from the paired-classes. If the oral directives elicited the same type of student engagement between two activities in paired-classes, the activities were considered common. Since teachers never gave the same instructions word for word in both paired-classes, we allowed slight deviations in the wording of instructions between common activities. Out of the 168 implemented activities identified from the videos, 127 of them were unique and 41 of them were common; again, similar in distribution to the 105–126 split of designed unique and common activities. Finally, we used the time codes from the activity boundaries to determine the average duration of common and unique activities across classes.

Overall, across all 63 classes, teachers devoted an average of 7.4 min per class to common activities and 42.4 min per class to unique activities. Thus, teachers did spend almost six times more class time on the manipulated unique activities than the common ones, so that any significant differences in student achievement can be reasonably attributed to their ICAP manipulation.

5.3.2. Accuracy of the oral directives

As shown in Fig. 4, we had analyzed the ICAP mode of the written directives that teachers had included in the design of their activities and compared our coding with the teachers' intended ICAP mode for their classes. To assess their implementation fidelity, we transcribed teachers' oral directives for the unique activities in the videos, and segmented and coded them in the same way as the written directives, based on directive verbs plus some expected action. There was a total of 181 segments, distributed across the class modes as shown in Appendix A, Row 8. Three coders separately coded 20% of the oral instructional segments, and the Krippendorff's alpha reliability was 0.58, 95% CI [0.46, 0.69]. The three coders then met and resolved disagreement through discussion. After that, each of the three coders individually coded one-third of the remaining oral directive segments.

The pattern of results for the oral directives, shown in Fig. 6, is again similar to the written directives (as shown in Fig. 4). As Fig. 6 shows, teachers' oral directives were most accurate for *Active* classes, in that 64% of their oral directives described

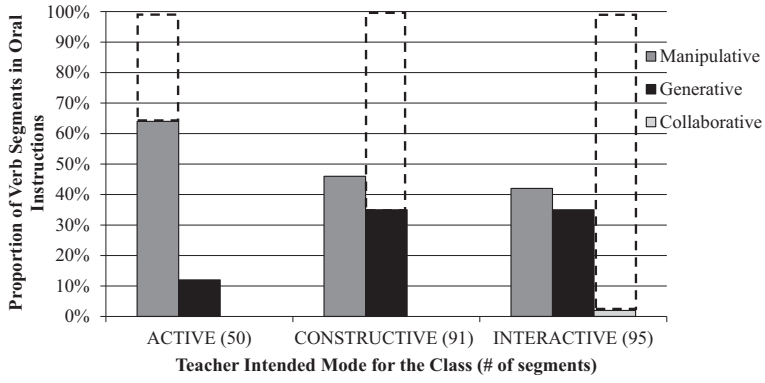

Fig. 6. Teacher implementation: Verb segments in oral directives for unique activities.

manipulative activities for their *Active* classes, significantly greater than 12% of the oral directives that were generative ($\chi^2(1, N = 38) = 17.789$, $p < .001$). For their *Constructive* and *Interactive* classes, the directives were about equally divided between manipulative (46%) and generative (35%), with no significant difference between them for either the *Constructive*, $\chi^2(1, N = 75) = 0.486$, $p = .486$ or the *Interactive* classes $\chi^2(1, N = 75) = 0.486$, $p = .486$). The dotted lines again show the ideal proportions that we had expected.

With only 2% of oral directives being collaborative for *Interactive* classes, it was significantly less than both the manipulative ($\chi^2(1, N = 43) = 31.837$, $p < .001$) and the generative directives ($\chi^2(1, N = 43) = 31.837$, $p < .001$). This analysis shows again that overall, teachers' oral directives were most accurate for activities intended for *Active* classes, and they were least accurate for the activities intended for *Interactive* classes, replicating the results for their verbal directives.

5.3.3. Distorting the assessment question type

In Tables 1 and 2, we showed that teachers' design of assessment questions for each lesson revealed that they could accurately design easy/hard questions, based on students' answer correctness. They were also able to accurately design recall/inference questions, according to the researchers' subjective judgment. In this section, we recoded teachers' recall/inference assessment questions based on the oral directives teachers gave in the videos. This recoding could only be done for the teacher-designated recall/inference questions, and not the easy/hard questions because the easy/hard questions were confirmed in terms of students' learning outcomes, which we have already shown in Table 1, and not from classroom implementation, as captured in the videos.

More specifically, as before, a previously designed recall question in lesson plans remained a recall question if we confirmed that the content information needed to answer the question was given during implementation in the classroom video either orally or as writings on a whiteboard, or provided by the lesson materials presented during class. If

the answer was not provided, the question was recoded as an inference. Similarly, for each inference assessment question originally designed in lesson plans, we analyzed the corresponding video data to verify whether the teachers provided facts, concepts, or connections that answered the question. If so, the question was recoded as a recall question. That is, an inference question that required students to generate new knowledge for an answer can be converted to a recall question if the teacher provided the answer to that question during the class instruction.

Table 3 shows the results of this recoding based on implementation compared to teachers' original labels from lesson plans. After this video analysis, 98% of the teacher designed recall questions remained a recall question in that the answer to such a question was provided in the lesson materials. However, 75% of the questions designed to be inference questions were recoded as recall questions because the answers were presented during classroom instruction. Thus, teachers' original labels of recall and inference for post-test questions as designed differed significantly from the actual mode of the questions as implemented, based on analyses of the content of the classroom videos, $\chi^2(2, N = 166) = 31.354$, $p < .001$.

The conversion of many inference questions to recall questions during implementation of classroom instruction negated the accuracy advantage of their originally designed inference questions (as shown in Table 2). This finding suggests that our prior interpretation, that their designed accuracy could have benefitted from having the question stems that we had provided, seems even more plausible. That is, teachers' conversion of an inference question to a recall question during implementation shows that they may not have understood the essence of an inference question after all. Thus, the discrepancy between teachers' lack of understanding of the distinction between shallow and deep questions in an earlier PD, and their apparent understanding of recall versus inference questions in this current study (as shown in Table 2) can be explained by teachers' usage of the provided question stems, which could be used without understanding what an inference question is.

The 32 post-tests included a total of 69 inference types questions, or 2.2 inference questions per lesson plan. However, after recoding, only 25% (or 17 of the 69 questions) remained an inference question. This means that over half of the post-tests did not

Table 3
Percentage agreement between teacher-designed recall and inference questions and researcher's re-coding based on implementation

Teacher-Designed Question Types (166 total)	Researchers' Re-coding	
	Recall	Inference
Recall (97)	98%	2%
Inference (69)	75%	25%

include a single inference question (17 of the 32 lessons), and almost half of the other post-tests included only a single inference question (7 of the remaining 15). Because deep learning benefits associated with generative and collaborative engagement are only measureable by inference questions, the teacher-designed assessment questions for their lessons were not adequate for discriminating between the learning benefits of *Constructive* and *Interactive* classes from *Passive* and *Active* classes.

In summary, while there was not perfect fidelity between intention and implementation, teachers did spend the majority of the class time devoted to the unique ICAP activity mode that they had designed as intended. However, their written directives and oral directives for activity worksheets and the actual worksheet questions all provided the same pattern of distortion, in that only about half of the directives for *Constructive* classes were generative, with almost non-existent collaborative directives for *Interactive* classes. Only the *Active* classes were implemented somewhat accurately, in that the majority of the directives for *Active* classes was manipulative. Overall, the *Interactive* classes were implemented the least accurately. The distortion also occurred in the way teachers converted their deeper inference assessment questions to shallower recall questions during implementation.

5.4. Fidelity of student enactment to activities based on product coding

Even well-designed and fidelitously implemented lessons cannot guarantee that students enact the activities in the ICAP mode that teachers intended. In this section, we investigated whether products that students generated while enacting class activities confirmed engagement at teachers' intended ICAP mode.

5.4.1. Coding of student products

Students' written products from both the unique and the common activities were analyzed and taken from paired-classes in which student written products were available, and the length of time engaged in both unique and common activities did not differ between the paired-classes by more than 25% of the total minutes for the lesson. The selected sample of written student products were either: (a) verbal notes or answers to questions, (b) visual concept maps, diagrams, posters, or drawings, or (c) numeric responses to mathematical problems, taken from 65 activities embedded in 31 classes, developed by 11 teachers, for a total of 1,171 written products (distribution across classes is shown in Appendix A, Row 9, lower set of numbers).

Because student products or responses varied so radically from activity to activity, the segmentation and coding process was specific to the format of the activities. Segmentation for verbal data was done at a sentence or phrase level, an idea or argument chain level, or at an explanation level, whichever was most appropriate for the assignment. Segmentation for visual data was dependent on whether the visualization was a drawing (segmented as a single unit), a diagram (segmented at structural level), or markings of text (segmented by non-content features like paragraphs). For example, in a lesson on "ions," students were asked to draw representations of ions, such as sodium chloride (Na^+Cl^-).

The scoring guide developed for this class indicated that each structure (e.g., electrons, orbitals, nucleus), behavior (e.g., "only 2 electrons in this orbital"), and function (e.g., labeling orbitals as energy levels, or valence electrons as ones that will react) would be segmented as one idea unit and then coded. Lastly, segmentation for numeric data was done at the problem level or the equation level, which generally elicited only one ICAP mode. Generally, segmentation was done to create a unit-of-analysis that was most appropriate for determining student engagement modes throughout an activity. The 1,171 student products were segmented into 18,131 segments (see Appendix A, Row 10).

After segmentation was completed, each segment was coded as manipulative or generative by determining if the students were required to recall or infer information. It was not possible to evaluate collaborative generation of information for *Interactive* activities from their products because we were not able to audio-record nor videotape the students while working. Without such information, we were not able to determine if co-generation of information occurred. Therefore, this analysis of student enactment is limited to whether students engaged manipulatively or generatively. Consequently, in 8 of the 11 classes intended as *Interactive*, our product coding could only verify whether written student work was manipulative or generative. This is still important information because, according to the ICAP framework, *Interactive* tasks should involve knowledge construction.

For the same reasons given for the analysis of recall and inference questions, when coding the student product segments, it was necessary to simultaneously review the source of class content—teachers' PowerPoint presentations, readings, or lectures—to determine how students were engaging. If students were generating new information by making inferences beyond what was given, the segment was coded as generative. If they were not generating new information, such as copying notes or applying algorithms, the segment was coded as manipulative. In some cases it was difficult to determine student engagement mode without knowing whether students wrote their responses before or after the teacher reviewed the correct answer, such as when teachers asked inference questions in class, instructed students to record their responses, and then reviewed the correct answer. In these instances, if a student's response was identical to teacher's answer, then it was coded as manipulative, and if it was different, it was coded as generative. For example, during a lesson on "enzymes," the teacher instructed students to take notes. Suppose the teacher used the terms "puzzle piece" and "lock and key" when referring to the appearance or function of an enzyme in the lecture, and if a student's notes contained those terms, then it was coded as manipulative. However, if the notes contained a qualitatively different description, it was coded as generative. For classes in which the teacher demonstrated a problem solution, and then asked students to solve problems that used the same steps or formula, the segments were coded as manipulative.

Because of the massive amount of student product segments to code (18,131 segments), seven members of the research team divided up the coding task. After coding was completed, inter-rater reliability was determined for 28% of the data by having a second member of the research team code the data independently. The data chosen for the reliability analysis represented the variety of activity types across language arts, science, and math classes. The inter-rater agreement for each class ranged from 74% to 100%.

The results from the two coders were compared, and any discrepancies were resolved by discussion.

Fig. 7 shows the average proportion of each type of student enacted segments for the activities. For teacher-intended *Active* classes, 98% of the student enacted segments (30.0 out of 30.6 segments per class) were manipulative, which is significantly greater than the proportion of generative segments (2% or 0.6 segments) for those same activities, $t(14) = 2.627$, $p = .034$. This means student enacted accurately, as expected because these were intended to be *Active* classes.

However, students produced on average only 17% (3.7 segments) and 14% (3.2 segments) of generative segments for classes intended to be *Constructive* and *Interactive*, respectively. Ideally, students should be engaging predominately (preferably 100%, as indicated by the dotted line in Fig. 7) generatively in those two class modes. Note that even though we could not code for co-generation in *Interactive* classes, co-construction does require that each partner must also be generative; thus, we would expect a large proportion of generative activity in *Interactive* class. Instead, students responded manipulatively 82% of the times for the *Constructive* classes and 85% of the times for *Interactive* classes. Therefore, the majority of students' responses were enacted manipulatively, regardless of the intended mode of the class.

Despite the low frequency of generative segments enacted by students in the *Constructive* classes, the proportion of generative segments was significantly different for the three intended modes of classes, $F(2, 28) = 2.959$, $p = .031$, with significantly greater proportion for *Constructive* classes when compared to *Active* classes ($p = .032$), but not between *Constructive* and *Interactive* classes, nor between *Interactive* and *Active* classes, based on a post hoc comparisons using Bonferroni error rate adjustment. This shows that even though students distorted their enactments, the teachers' design of the class activities, especially for the *Constructive* classes, did have some effect in the proportion of generative segments they produced.

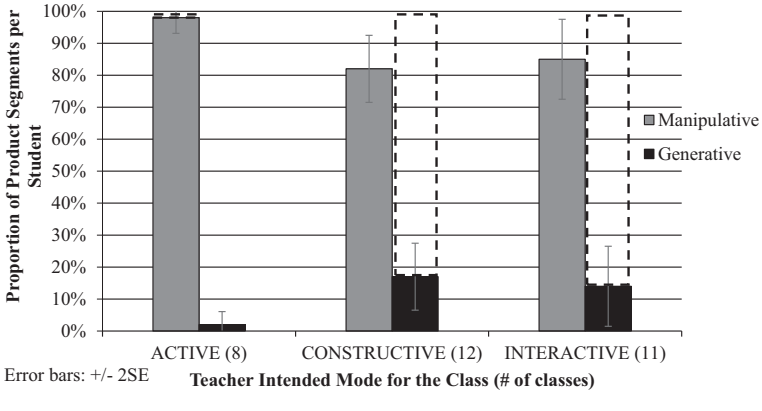

Fig. 7. Student-enacted written product segments for worksheet activities.

5.4.2. Students' enactments to specific worksheet questions

Because oral directive for an activity is typically delivered prior to students' execution of the activity, it is easy to imagine that teachers' oral instructions did not stick in students' minds. So a more direct way to look at the impact of the mode of teachers' design of activity is to look at the explicit requests made by specific questions embedded within activity worksheets. Therefore, this analysis focused on the 117 questions within the 14 worksheets that were selected previously (see Fig. 5 again & Appendix A Row 7) based on availability of student enactment data.

Each of the 117 worksheet questions was simply coded as to whether it requested a manipulative or a generative response, regardless of the intended mode of the class the worksheets were embedded in. A worksheet question was coded as manipulative if it could be answered verbatim or retrieved directly from the written materials or PowerPoint slides teachers provided, often using verbs similar to the ones they used for the directives (as shown in Appendix B). There were 65 *Active* questions, and two examples are as follows:

Name three enzymes and the reactions they are involved with.

Sketch a line that connects the data points. (The data points are provided.)

A worksheet question was coded as generative if the answer required the students to generate new content or inferences not present in these materials. There were 52 generative questions (see Appendix A Row 11), and two examples are:

What is the main idea of these two paragraphs?

Suggest how this experiment could be changed to investigate the effect of temperature on the activity of catalase.

After the mode of questions was coded, we re-plotted the previously coded student responses as a function of the question mode. Fig. 8 shows the mode of students' responses to each mode of questions: For *Active* questions, 95% of response segments were manipulative, which is appropriate since an *Active* question requires only a manipulative response; however, for *Constructive* questions, only 42% of the responses were generative. This shows that even when a question explicitly requests a generative response, only 42% of students' response segments were generative. This may be explained by the fact that generating a *constructive* response takes more effort, requiring the process of inferring.

However, what is more important to compare in Fig. 8 is that when a question explicitly requested a generative response (i.e., a *Constructive* question), students were more likely to respond generatively (42% of the times) than when a question requested a manipulative response (i.e., an *Active* questions), then students rarely responded generatively

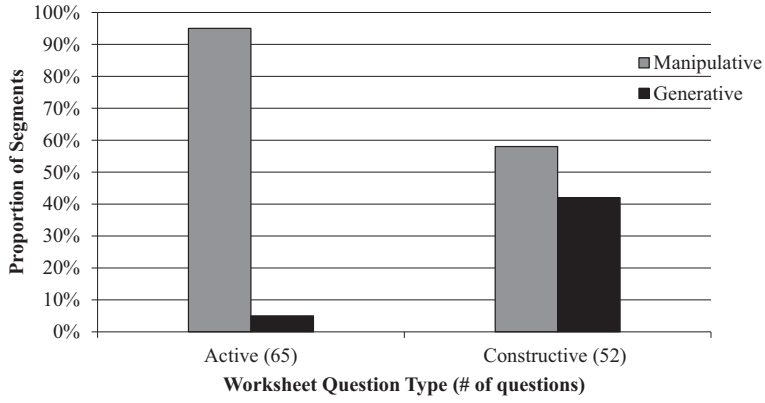

Fig. 8. Student enacted segments for specific worksheet questions.

(5%). This difference (42% vs. 5%) is highly significant $\chi^2(1, N = 2373) = 490.84$, $p < .001$.

This pattern of results shows that, overall, students did have a tendency to respond in manipulative ways (95% and 58%) regardless of whether a question elicited a manipulative or generative response, respectively. Nevertheless, when a question explicitly requested a generative response, students were more likely to respond generatively compared to when a question requested a manipulative response (42% vs. 5%). This result is extremely promising, suggesting that there is clearly an advantage to train teachers to create questions that elicit generative responses. That is, students are less likely to respond manipulatively to *Constructive* questions (58%) than *Active* questions (95%). Thus, coding individual questions based on what each question explicitly requested is a more sensitive measure of the benefit of responding generatively to a *Constructive* question.

5.5. Student learning outcomes from lesson implementation

When testing the validity of ICAP from evidence in the literature, we mapped a researcher's manipulation of the two contrasting conditions in their laboratory study into ICAP modes, and then compared students' performance outcomes for the two contrasting modes. We showed that the performance outcomes generally supported ICAP's predicted I > C > A > P direction (Chi, 2009; Chi & Wylie, 2014). In this study, teachers implemented each of their designed lessons in a pair of classes, with one class designated to be at one ICAP mode and the other paired-class designated to be at another ICAP mode. Each paired-class was assessed using the same teacher-constructed post-test, thereby allowing us in principle to contrast the classes as we had done with the studies from the literature.

For our current classroom data, although we originally planned to compare and contrast whether there were pairwise class differences in student learning in the ICAP predicted direction, based upon the two ICAP modes teachers had designed for each lesson,

such pair-by-pair comparisons unfortunately cannot be expected to support ICAP's prediction for all of the following reasons:

1. Teachers had an imperfect understanding of ICAP from the ICAP module, as indicated by their modest pre- and post-test performance on the harder more nuanced questions (see Fig. 1) and the misunderstandings they revealed.
2. Consequently, teachers did not design their lesson plans accurately with respect to the ICAP mode, as indicated by the incongruence in the proportion of the written directives for the unique lesson activities and the intended class mode (Fig. 3), as well as actual worksheet questions (Fig. 4). That is, if a class was intended by teachers to be a *Constructive* class, only 53% of the written directives pertains to using generative verbs, and only 7% of the written directives pertains to collaborative co-construction for intended *Interactive* classes. And since we did not carry out careful analyses until after all the data were collected, we did not provide adequate feedback for teachers to modify their lesson activities.
3. Although teachers' design of the assessment questions was initially more accurate with respect to the ICAP mode, they converted their inference questions into recall questions during implementation, leaving no inference question at all for about half of the lessons. This means that the post-test questions could not serve as an accurate assessment of learning in the contrasting paired-classes because inference questions were required to detect the differentiated learning gains associated with higher ICAP modes.
4. The fidelity of teacher implementation was inconsistent, as determined by the oral directives captured in the videos, even though they did allocate more class time to unique activities that were meant to differentiate their two paired-classes. That is, their oral directives for their unique activities mirrored the results of their written (designed) directives, in that there were very little generative (35%) directives for the *Constructive* classes and almost non-existent collaborative (2%) directives for the *Interactive* classes (see Fig. 6 again).
5. Finally, students did not enact activities in the ICAP mode as intended by class. Students predominately enacted manipulative responses across activities in all three modes of classes (see Fig. 7 again).

These design, implementation, and enactment limitations present too much noise for us to expect that each pairwise comparison of ICAP modes to result in the ICAP predicted direction. In general, research "in the wild" or authentic classrooms has to anticipate unexpected outcomes; that is the main purpose of testing an intervention in authentic classrooms. Hence, we carried out instead two other analyses on students' learning outcome data from their performance on the post-tests compared to the pre-tests. For both analyses, we aggregated the data over all the classes of each ICAP mode, and compared it with classes of another ICAP mode. In other words, we compared all the lessons of a given mode as a collective (instead of making pairwise comparisons for each paired-classes) and looked at the effect of engagement mode on learning gain.

5.5.1. Analyses of learning gains by pooling classes with the same ICAP mode

When comparing student learning outcomes across teacher-intended ICAP modes, we continued to exclude the two *Passive* classes from analyses. Students' individual improvement from pre- to post-test was calculated using a normalized gain score, g, (the proportion of possible gain a student made with respect to their pre-test score) described in Hake (1998). To examine the effect of engagement mode on student learning gains, the normalized learning scores (g) were first collapsed by teacher intended mode—all *Active* classes formed one group, all *Constructive* classes formed another group, and all *Interactive classes* formed a third group. A one-way ANOVA showed a significant main effect of ICAP mode, $F(2, 1354) = 8.497$, $p < .001$, $\eta_p^2 = 0.012$. Pairwise comparisons between modes of pooled classes further showed that *Constructive* classes collectively resulted in significantly greater student learning than *Active* classes ($g = 0.392$ and $g = 0.305$, respectively), $p = .001$ and *Interactive* classes ($g = 0.315$), $p = .002$. There was no significant difference in student learning between *Active* and *Interactive* classes, $p = .913$.

After consideration of the different content areas teachers taught, we realized that some classes were foreign language classes, and we had conjectured in Chi and Wylie (2014) that learning foreign language is less receptive to the benefit of *generative* activities because no meaningful elaborations or justifications can be provided for syntax or vocabulary. That is, because a large component of learning a new language in middle and high schools is rote memorization of new vocabulary and syntax, it is adequate for students to engage in *Active* activities to learn a new language; *Constructive* engagement such as generating meaningful justifications for grammatical rules are not beneficial since no meaningful reasons can be provided. Our study contained 12 foreign language classes assigned ICAP labels exceeding the *Active* mode. Allowing foreign language classes, which do not require *Constructive* engagement to be merged, with classes that benefit from generative behaviors may have distorted the effect of the ICAP level seen in the previous analyses. Thus, the present aggregate analysis excludes the 2 *Passive* classes and the 12 foreign language classes, leaving 51 classes (15 *Active*, 18 *Constructive*, 18 *Interactive*). Exactly the same pattern of results was found based on a one-way ANOVA: That is, there was a significant main effect of ICAP mode, $F(2, 1207) = 7.737$, $p < .001$, $\eta_p^2 = 0.013$; and pairwise comparisons between modes (not paired-classes) again showed that *Constructive* classes collectively resulted in significantly greater student learning than *Active* classes ($g = 0.394$ and $g = 0.301$, respectively), $p = .001$, and *Interactive* classes ($g = 0.318$), $p = .005$. There was no significant difference in student learning between *Active* and *Interactive* classes, $p = .808$. Fig. 9 shows the mean normalized learning gain scores for the 51 classes.

5.5.2. Analysis of learning gains with respect to ICAP mode using a multilevel model

In addition to the one-way ANOVA, we tested the effect of ICAP on student learning gains using a multilevel model on the 51 classes. As teacher level variability was not significant, a two-level model was used nesting students within classes. A nested model test indicated that, overall, ICAP mode had no effect on student learning gains after

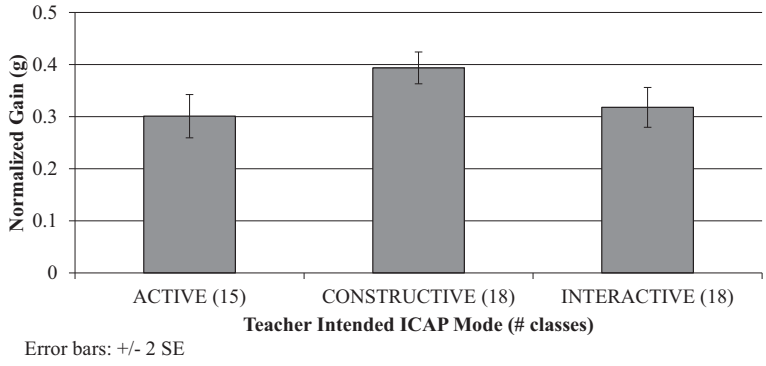

Fig. 9. Student mean normalized learning gain scores, aggregated across class modes.

controlling for pre-test, $\chi^2(2) = 0.149$. The failure of the nested model test to emerge as significant overall may reflect our limited sample size; only 51 classes, divided between three experimental conditions. However, the comparison between the *Constructive* condition and the *Active* condition did emerge as significant ($z = 2.055$, $p = .040$), again with the *Constructive* condition outperforming the *Active* condition.

In sum, we can interpret the significant finding in our model as highly suggestive that —when students' learning in *Constructive* classes outperformed *Active* classes, this may be accounted for by the modest increase in elicitations of generative responses (from 21% in *Active* classes to 53% in *Constructive* classes, see Fig. 4 based on written directives, and from 12% in *Active* classes to 35% in *Constructive* classes, see Fig. 6 based on oral directives). All three analyses (the ANOVA with either 63 or 51 classes, and the nested model with 51 classes) provided consistent results: that overall, students' learning was significantly better in *Constructive* classes than in *Active* classes, averaged across different teachers, different samples of students, different content topics, and so forth. Our theory suggested that due to the presence and absence of the inferring process, there should be a larger gulf between learning from generative versus manipulative engagement, respectively. The fact that students learned significantly more in *Constructive* than *Active* classes, despite the substantial distortions in teachers' implementation and students' enactment, confirms our assumption that the inferring process may be producing this gulf.

The superior learning outcomes of *Constructive* classes might also be accounted for by the increased number of student-enacted generative responses, as measured by the segments in students' products. As shown in Fig. 7, students produced a significantly greater number of generative segments in *Constructive* classes than within *Active* classes, and produced a greater number of generative segments for *Constructive* questions than *Active* questions (Fig. 8). These results suggest that even a modest amount of *Constructive* engagement (as indicated by 17% of enacted generative segments, shown in Fig. 7) was sufficient to improve the amount of student learning in *Constructive* classes overall, as compared to *Active* classes.

Although the comparison between *Constructive* and *Interactive* classes in terms of student normalized gain scores in the one-way ANOVA was opposite the predicted direction

according to ICAP, it is most likely due to the frequent implementation errors for *Interactive* classes (see Figs. 4 and 6) such that students were not co-generating, resulting in poorer student learning outcomes.

5.6. Summary of findings from this classroom implementation study

After having learned about ICAP through an online professional development module, teachers were asked to design lesson plans for pairs of classes, in which one class of the pair was designed to be at one ICAP mode and another class of the pair was designed to be at another ICAP mode. The difference between the paired-classes translated into differences in the number of activities that were either shared between the paired-classes (referred to as common activities) or not (unique activities). At a gross level, teachers succeeded in the following ways. First, although ideally they should have allocated all activities of one ICAP mode to one class and all activities of another ICAP mode to another class to maximize differences in ICAP modes of the paired-classes, this was not realistic in authentic classrooms nor was this 100% allocation explicitly stated in the ICAP module. Therefore, it was quite acceptable that at least they designed about three times as many unique activities as common (or shared between the paired classes) activities, and this translated into devoting six times more class time to unique activities (42.4 min) than common activities (7.4 min). Second, teachers were also accurate in the design of their assessment questions for each lesson plan, when the questions were defined as easy/hard and when question stems were provided for their design of recall/inference questions. The accuracy of their design for easy/hard questions was verified by the fact that students did have a harder time answering the hard questions correctly, compared to the easy ones. In short, their design process was somewhat accurate when assessed at a gross level, in terms of the number of activities relevant to the intended class mode and assessment question types.

However, teachers displayed shortcomings when we examined their designed activities in greater detail, such as their written directives for worksheet activities. Overall, the mode of their worksheet directives reflected predominantly the *Active* mode requiring manipulative student responses, even though the class was intended to be in either the *Constructive* or the *Interactive* modes (see Fig. 5 again).

During implementation, similar to their design of verbal directives for worksheets and actual questions in their worksheets, teachers tended to distort the intention of their designed activities and questions. That is, they were inaccurate in their oral directives for *Constructive* or *Interactive* classes by providing mostly *Active* directives. For their assessment questions that were originally designed to be inference questions, they often inadvertently converted them to recall questions by providing the answers in class, either orally or within written materials. This affected the validity of the post-test assessments in that they could not adequately measure students' learning gains associated with a deep understanding of the material.

Regardless of how teachers implemented their activities and what kind of responses they elicited, students uniformly tended to give manipulative responses (between 82%

and 98% of the time) during all three ICAP modes of activities. For example, when teachers intended their classes to be a *Constructive* class, students responded overall with generative comments only around 17% of the times; instead, they tended to respond manipulatively 82% of the times (see Fig. 7).

A more fine-grained analysis of student enactment—by explicit individual questions embedded within worksheets showed that students, on average, did produce more generative responses when a question explicitly elicited a generative response. This suggests that having teachers explicitly direct students to be generative did increase students' proportion of knowledge generation attempts, from 17% (as determined by the overall worksheet instructions) to 42% (when elicited explicitly in the context of a question, compare Figs. 7 and 8 again). This suggests that students were somewhat responsive when teachers explicitly elicit generative responses, although nowhere near 100%.

Because teachers did not implement their lesson pairs perfectly, nor did students enact them as intended, there was no point in looking at students' learning outcomes per each paired-classes in a contrastive way, especially since the assessment questions were also not sufficiently deep to detect ICAP's prediction of learning outcomes. Therefore, our analyses combined all classes within each ICAP mode and examined the data as a collective. The distortions in teachers' implementation of the *Interactive* classes (in terms of almost non-existent collaboration directives for *Interactive* classes) and the limited design of only two *Passive* classes forced us to assess the validity of the ICAP hierarchy with only the *Active* and *Constructive* modes. However, it is important to note that we were testing the prediction of the hierarchy based on implementation in classrooms that were carried out by teachers who had acquired some understanding of ICAP; we were not evaluating ICAP's predictions in terms of the researchers' implementation of ICAP in the classrooms. Thus, we are evaluating the success/limitations of a translation project carried out by teachers. Excitingly, despite the various limitations in teachers' implementation and students' enactment, and the various shortcomings of the ICAP module unearthed by this ambitious project, our results provide promising indication that for teachers to add even small amounts of *Constructive* activities, for which students actually engaged in generative behaviors by producing ideas that go beyond what was given to them, was beneficial for learning in classrooms. That is, students' normalized learning gain scores did show a significant improvement for *Constructive* classes compared to *Active* classes, using both an analysis of variance and a nested model analysis. This suggests that even with weak implementation accuracy, students did learn substantially more from the *Constructive* than the *Active* classes averaged across a variety of topics, students, grade levels, and so forth.

6. Discussion

In this closing section, we discuss the unique aspects of this project, the challenges we faced, and the lessons learned. We then consider how these challenges can be addressed in moving forward, and we reflect on our top–down translation approach.

6.1. Uniqueness of this project

This project is unique in several ways. First, it has been an ambitious attempt to translate a broad evidence-based theory about cognitive engagement/active learning to practice in many authentic classroom settings. Other projects in the literature that have attempted to translate laboratory findings into classroom practice operate at a much smaller scale, such as focusing on translating the implementation of a specific learning strategy, such as self-explaining (Chi et al., 1989) into practice; and there are many examples of how such translation of a specific strategy could proceed. Typically, the translation might consider how one would prompt for self-explanations in the classrooms (Renkl, 1997), sometimes relying on technology to do so. Often such approaches scaffold students directly to use a strategy, without asking teachers to intervene. To translate a theory which is applicable across all content domains and age groups, we had to figure out what approach to take. Our approach of teaching teachers first (in the form of an online module PD about ICAP) and then asking teachers to transfer what they have understood into design and implementation is clearly non-typical and operates at a larger scale.

Second, the professional development (PD) module itself is also unique in that it aims to teach teachers about how students engage to learn. In contrast, PD typically focuses on instructing teachers on (a) how to teach, such as using various strategies of teaching (e.g., pausing frequently, or knowing what kind of questions to ask, such as revoicing, Michaels, O'Connor, & Resnick, 2008), as well as (b) teaching teachers how to manage their classrooms, how to orchestrate, how to assess students, or (c) teaching teachers strategies for teaching a particular subject domain (or pedagogical content knowledge), and that may include teaching them what students' misconceptions are and how to teach in a way that addresses their misconceptions (e.g., how they think mathematically; Carpenter, Fennema, Peterson, Chiang, & Loef, 1989). Few, if any, PD teaches teachers general understanding of how students learn.

Third, this project is unique in that it involves two tiers of translation: The researchers had to decide how to translate a theory into knowledge for teachers, and the teachers then had to decide how to translate what they have understood into lesson design and implementation, essentially a form of transfer. Asking teachers to translate what they have understood about our theory into practice reveals deficits in the theory and indicates what needs to be improved. For example, we now realize that ICAP needs to specify more clearly what collaborative interactions involve, and that students may need scaffolding to respond more generatively.

Fourth, teachers' successes at translation were evaluated in five extensive ways, as described in this paper. Assessments included teachers' knowledge of ICAP, teachers' design of lesson plans, teachers' implementation of lesson plans, students' enactment, and students' learning outcome. Such comprehensive assessment of PD is unique, as Desimone (2009) noted, because very few studies have provided links between the PD, teacher knowledge, teacher practice and student achievement (Carpenter et al., 1989; Franke, Carpenter, Levi, & Fennema, 2001).

Fifth, we made every attempt to document and standardize our PD module so that we know what is being taught and how effective it is, so that it has the potential for scale up. Currently, most PDs in practice are delivered without an assessment of its impact, or how it can be replicated in exactly the same way for a different group of teachers.

6.2. Challenges and lessons learned

The challenges that we have uncovered as we translate ICAP into practice are important lessons learned about both classroom practice and our PD module. We describe three major challenges.

The most prominent revelation from our study was finding that teachers had the greatest difficulty implementing instruction for *Interactive* activities, based on analyses of their written and oral verb directives. We propose two possible reasons. First, it may be that teachers were more focused on the conditions of collaboration (e.g., whether students were sitting next to each other, who to pair with whom, whether the tasks were conducive to discussion) rather than on the pattern of the dialogs. Thus, teachers are not aware that simply telling students to "work with your partner" will not suffice to elicit co-generative collaborative behavior, that is, working together in a mutually reciprocal way. Of the nine collaborative verbs they used (in Appendix B), seven of them can be conceived of as terminologies that mostly refer to working with a partner. Perhaps only two verbs ("agree upon" and "debate") might be conceived of as referring to dialog patterns. As we have established elsewhere (Chi & Menekse, 2015), using ICAP as a coding lens to code each partner's contribution shows that there is a variety of dialog patterns possible when two peers work together, and not all of these patterns can be considered co-generative. In short, teachers are conceiving of collaboration as a physical task, requiring certain conditions, whereas researchers are conceiving of collaboration as requiring a certain dialog pattern. This discrepancy is an important lesson learned about teachers' conception of collaborative learning.

The second reason is that we now recognize that our own PD was faulty in not specifying how students should collaborate verbally. We re-examined our ICAP module for instruction given on collaboration and found that we had 20 instances of presenting collaboration verbs, as shown in Fig. 10, with 12 unique descriptors/verbs used. Among these 20 instances, only two (i.e., "debate" and "challenge or confirm each other's ideas") could be conceived of as implying the co-generative type of dialog pattern. Therefore, it makes sense that teachers did not use the proper verbs or verb phrases in their directives to students to elicit mutual-and-reciprocal type of co-constructive collaboration. Thus, the lesson learned is that we need to improve our instruction in the PD module about how to convey to teachers about how to facilitate students' co-generative collaboration.

The second challenge is teachers' tendency to design *Active* activities that elicited *manipulative* responses when they intended to design *Constructive* activities that elicited *generative* responses. Because they were quite accurate in designing *Active* activities in that those activities correctly elicited a majority of *manipulative* responses, we can only surmise that teachers were much more familiar with designing *Active* activities than

Constructive activities (see Fig. 3). The lesson learned here is that we should provide better feedback on their lesson design (recall that two researchers gave them cursory feedback only), and perhaps also provide more practice opportunities in designing *Constructive* activities.

The third challenge is that teachers had great difficulty implementing the changes or the behavior in the desired intended ICAP mode in real time. For example, we showed that they essentially converted the inference assessment questions they had designed into recall questions during implementation by giving away the answers to those questions. Upon reflection, we should not be surprised at teachers' distortions in implementation since implementation is a process of transferring understanding to behavioral change. From the cognitive science literature, transfer, even within cognitive knowledge per se, such as from transferring knowledge of how to solve a simple problem almost identical to a presented example to a more complex problem, is almost impossible to achieve (Chi & VanLehn, 2012), and a change in behavior such as smoking, is also impossible to achieve even when people are motivated to stop smoking. Here, we are considering transfer from knowledge to behavioral practice and moreover, the practice is a very complicated dynamic skill of teaching in real time in a complex setting. Nevertheless, it is promising to see that a modest amount of professional development about ICAP (3 h on average) was adequate in helping our teachers increase their *generative* learning activities, for example, from about 20% for *Active* classrooms to over 50% for *Constructive* classrooms (see Fig. 4). Such improvement may have been sufficient to elicit more generative engagement from students, which led to improved learning in the *Constructive* mode overall compared to the *Active* mode, consistent with the commonsense impression that "minds-on" activities enhances learning more so than "hands-on" activities. The lesson

20 instances of these VERBS about collaborating	
• **Work** together (2)	• **Act** as a tutor or tutee (2)
• **Discuss** with a partner (2)	• **Talk** with a tutor
• **Debate** with other students	• **Interact** with a computer tutor
• **Talk** with a teacher, partners, or participate in dialogues (3)	• **Do** some task with a partner (2)
• **Collaborate** with teammates	• **Explain** to another (2)
• **Review** with a friend (2)	• **Challenge** or **confirm** each other's conclusion

Fig. 10. Verbs used in the ICAP module.

learned here is that we cannot expect transfer from newly learned knowledge (about ICAP) into behavior easily (i.e., implementing lesson plans in class).

6.3. Reflections and current directions

In moving forward, we face many challenges that may be difficult to overcome without further research. The most difficult challenge is to know how to help teachers understand and convey to students what effective collaborative dialog patterns are and how to carry out such collaborative interactions. In point of fact, facilitating students on how to collaborate in a co-generative way is a challenge that has not been resolved in the literature. Many approaches are currently being explored and devised, such as providing a script to scaffold collaborators. The majority of these techniques may be effective in the context of the task or domain students are being scaffolded. But the skill of collaboration may not have been learned adequately to transfer to other tasks or contexts. Even though we have provided precise concrete definitions for what kind of dialog patterns is most effective for collaborative learning (Chi & Menekse, 2015), nevertheless, the challenge remains in how to translate our specification into ways that teachers can easily understand so that they can convey this specification to students.

Moreover, we are aware of other obstacles to teachers implementing collaborative activities. In our current work, we find that teachers are reluctant to implement collaborative activities because they hold several false pedagogical beliefs. For example, they think that they would lose control of classroom management if students are grouped and free to interact, or they cannot detect which pairs or groups of students are working effectively and which groups are not, or once they approach a group that has difficulty, they may not know how to diagnose their difficulty easily and quickly. Future work would need to approach the development of teacher training from the perspective of conceptual change, such that teachers' negative beliefs about the utility of collaborative practice are directly challenged.

Our translation project was very top–down, in that a theory was first developed and instruction to explain that theory was then provided to teachers, expecting them to modify their teaching accordingly. Although our top–down approach incorporates many aspects of bottom-up design-based research approach (such as the use of mixed methods, involving multiple iterations, situated in real classrooms, testing an intervention), nevertheless, we should incorporate a teacher partnership earlier in the research cycle. For example, our current effort seeks a teacher's inputs in how a lesson should be designed to meet ICAP criteria. The challenge of a research-practitioner partnership however, is to be able to accommodate the revisions required by the teacher, yet maintain the integrity of the initially designed intervention, so that we can accurately assess the impact of the intervention without compromise, as predicted by the theory.

Although conducting such multi-year and multi-facetted research projects is effortful, including hours of laborious and herculean coding efforts, the potential payoff of verifying the utility of a theory that has broad applicability for student engagement that functions with equal effectiveness in all classroom contexts, is exciting.

Acknowledgments

The authors are grateful for funding provided by the U.S. Department of Education, Institute of Education Sciences, grant number: R305A110090, for the project titled "Developing Guidelines for Optimizing Levels of Students' Overt Engagement Activities," and grant number R305A150432 for the project titled "Developing and Revising Instructional Activities to Optimize Cognitive Engagement." We are particularly grateful to both Dr. Josephine Marsh for her help in facilitating our work at ASU Preparatory Academy and to ASU Preparatory Academy for participating in this study. Comments from Joshua Morris and Lu Ding are greatly appreciated.

References

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education*, *16*, 101–128.

Alibali, M. W., & DiRusso, A. A. (1999). The function of gesture in learning to count: More than keeping track. *Cognitive Development*, *14*, 37–56.

Anderson, J. R., & Reder, L. M. (1979). An elaborative processing explanation of depth of processing. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 385–403). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ausubel, D. P., Novak, J. D., & Hanesian, H. (1986). *Educational psychology: A cognitive view* (2nd ed.). New York: Warbel and Peck. (original work published in 1978)

Bajak, A. (2014). Lectures aren't just boring, they're ineffective too, study finds. *Science*. Retrieved October 17, 2016, from: http://www.sciencemag.org/news/2014/05/lectures-arent-just-boring-theyre-ineffective-too-study-finds

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, *26*, 600–614.

Biswas, G., Leelawong, K., Schwartz, D., & Vye, N., & The Teachable Agents Group at Vanderbilt (2005). Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence*, *19*, 363–392.

Broughton, S. H., Sinatra, G. M., & Nussbaum, E. M. (2013). "Pluto has been a planet my whole life!" Emotions, attitudes, and conceptual change in elementary students' learning about Pluto's reclassification. *Research in Science Education*, *43*(2), 529–550.

Brown, A. L., & Campione, J. C. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229–270). Cambridge, MA: The MIT Press.

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review*, *31*, 21–32.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal*, *26*, 499–531.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1*, 73–105.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*, 145–182.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477.

Chi, M. T. H., & Menekse, M. (2015). Dialogue patterns that promote learning. In L. B. Resnick, C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (Ch. 21, pp. 263–274). New York: Routledge.

Chi, M. T. H., Roy, M., & Hausmann, R. G. M. (2008). Observing tutoring collaboratively: Insights about tutoring effectiveness from vicarious learning. *Cognitive Science*, *32*, 301–341.

Chi, M. T. H., & VanLehn, K. (2012). Seeing deep structure from the interactions of surface features. *Educational Psychologist*, *47*, 177–188.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*, 219–243.

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, *4*, 129–145.

Conati, C., & VanLehn, K. (2000). Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education*, *11*, 389–415.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, *69*, 970–977.

Csikszentmihalyi, M. (1988). Motivation and creativity: Toward a synthesis of structural and energistic approaches to cognition. *New Ideas in Psychology*, *6*(2), 159–176.

Damon, W. (1984). Peer education: The untapped potential. *Journal of Applied Developmental Psychology*, *5*, 331–343.

Dawson, I. (2004). Time for chronology? Ideas for developing chronological understanding. *Teaching History*, *117*, 14–24.

De Temple, J., & Snow, C. E. (2003). Learning words from books. In A. van Kleeck, S. A. Stahl, & E. B. Bauer (Eds.), *On reading books to children: Parents and teachers* (pp. 16–36). Mahwah, NJ: Lawrence Erlbaum Associates.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*, 181–199.

Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science*, *332*, 862–864.

Dinsmore, D. L., & Alexander, P. A. (2012). A critical discussion of deep and surface processing: What it means, how it is measured, the role of context, and model specification. *Educational Psychology Review*, *24*, 499–567.

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, *70*, 377–398.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, *41*, 1040.

Fonseca, B., & Chi, M. T. H. (2011). The self-explanation effect: A constructive learning activity. In R. E. Mayer & P. A. Alexander (Eds.), *The handbook of research on learning and instruction* (pp. 296–321). New York: Routledge.

Franke, M. L., Carpenter, T. P., Levi, L., & Fennema, E. (2001). Capturing teachers' generative change: A follow-up study of professional development in mathematics. *American Educational Research Journal*, *38*, 653–689.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*, 59–109.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*, 8410–8415.

Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist*, *50*, 43–57.

Gobert, J. D., & Sao Pedro, M. A. (2017). Digital assessment environments for scientific inquiry practices. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment frameworks, methodologies, and applications* (pp. 508–534). Hoboken, NJ: John Wiley & Sons Inc.

Grabinger, R. S., & Dunlap, J. C. (1995). Rich environments for active learning: A definition. *ALT-J*, *3*(2), 5–34.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, *31*, 104–137.

Greene, B. A. (2010). Teacher quality and student success: Testing the K20 Science Professional Development Model (K20 Science) for Rural Science Teachers. NSF Final Report, Award No. 0634070.

Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. *Educational Psychologist*, *50*, 14–30.

Greene, B. A., & Miller, R. B. (1996). Influences on achievement: Goals, perceived ability, and cognitive engagement. *Contemporary Educational Psychology*, *21*, 181–192.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*, 64–74.

Heddy, B. C., & Sinatra, G. M. (2013). Transforming misconceptions: Using transformative experience to promote positive affect and conceptual change in students learning about biological evolution. *Science Education*, *97*, 723–744.

Hogan, K., Nastasi, B. K., & Pressley, M. (1999). Discourse patterns and collaborative scientific reasoning in peer and teacher-guided discussions. *Cognition and Instruction*, *17*, 379–432.

James, K. H., Humphrey, G. K., Vilis, T., Corrie, B., Baddour, R., & Goodale, M. A. (2002). "Active and "passive" learning of three-dimensional object structure within an immersive virtual reality environment. *Behavior Research Methods, Instruments, & Computers*, *34*, 383–390.

Katayama, A. D., Shambaugh, R. N., & Doctor, T. (2005). Promoting knowledge transfer with electronic note taking. *Teaching of Psychology*, *32*, 129–131.

King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal*, *29*, 303–323.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effect of direct instruction and discovery learning. *Psychological Science*, *15*, 661–667.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*, 411–433.

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, *126*, 198–212.

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal*, *37*(1), 153–184.

Menekse, M., Stump, G. S., Krause, S., & Chi, M. T. H. (2013). Differentiated overt learning activities for effective instruction in engineering classrooms. *Journal of Engineering Education*, *102*, 346–374.

Mercer, N. (1996). The quality of talk in children's collaborative activity in the classroom. *Learning and Instruction*, *6*, 359–377.

Mestre, J. P. (2002). Probing adults' conceptual understanding and transfer of learning via problem posing. *Journal of Applied Developmental Psychology*, *23*, 9–50.

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliverative discourse idealized and realized: Accountable Talk in the classroom and in civic live. *Studies in Philosophy and Education*, *27*, 283–297.

Newmann, F. M., Wehlage, G. G., & Lamborn, S. D. (1992). The significance and sources of student engagement. In F. M. Newmann (Ed.), *Student engagement and achievement in American secondary schools* (pp. 11–39). New York: Teachers College Press.

O'Reilly, T., Symons, S., & MacLatchy-Gaudet, H. (1998). Brief research report: A comparison of self-explanation and elaborative interrogation. *Contemporary Educational Psychology*, *23*, 434–445.

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

Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 259–282). New York: Springer.

Peterson, P. L., Swing, S. R., Stark, K. D., & Waas, G. A. (1984). Students' cognitions and time on task during mathematics instruction. *American Educational Research Journal*, *21*, 487–515.

Piaget, J. (1930). *The child's conception of physical causality*. New York: Harcourt Brace & Company.

Ravindran, B., Greene, B. A., & DeBacker, T. K. (2005). Predicting preservice teachers' cognitive engagement with goals and epistemological beliefs. *The Journal of Educational Research*, *98*, 222–233.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, *21*(1), 1–29.

Reschly, A., & Christenson, S. L. (2006). Promoting school completion. In G. Bear & K. Minke (Eds.), *Children's needs III: Understanding and addressing the developmental needs of children* (pp. 103–113). Bethesda, MD: National Association of School Psychologists.

Rochelle, J. (1992). Learning by collaborating: Convergent conceptual change. *Journal of the Learning Sciences*, *2*, 235–276.

Scardamalia, M., & Bereiter, C. (1996). Adaptation and understanding: A case for new cultures of schooling. In S. Vosniadou, E. De Corte, R. Glaser, & H. Mandl (Eds.), *International perspectives on the design of technology-supported learning environments* (pp. 149–163). New York: Routledge.

Schauble, L., Glaser, R., Duschl, R. A., Schulze, S., & John, J. (2009). Students' understanding of the objectives and procedures of experimentation in the science classroom. *Journal of the Learning Sciences*, *4*, 131–166.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, *103*, 759–775.

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(1), 1–114.

Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educational Psychologist*, *50*, 1–13.

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, *85*, 571.

Stump, G. S., Li, N., Kang, S., Yaghmourian, D., Xu, D., Adams, J., McEldoon, K., Lancaster, M., & Chi, M. T. H. (2018). Coding dosage of teachers' implementation of activities using ICAP: A video analysis. In E. Manalo, Y. Uesaka, & C. A. Chinn (Eds.), *Promoting spontaneous use of learning and reasoning strategies: Theory, research, and practice for effective transfer* (pp. 211–225). New York: Routledge.

Trafton, J. G., & Trickett, S. B. (2001). Note-taking for self-explanation and problem solving. *Human-Computer Interaction*, *16*, 1–38.

University of California, Berkeley. (2016). WISE-Web-based Inquiry Science Environment [Online Computer Software]. Retrieved October 17, 2016, from: https://wise.berkeley.edu/pages/wise-advantage.html

van Driel, J. H., Meirink, J. A., Van Veen, K., & Zwart, R. C. (2012). Current trends and missing links in studies on teacher professional development in science education: A review of design features and quality of research. *Studies in Science Education*, *48*, 129–160.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, *31*, 3–62.

Webb, N. M., Franke, M. L., De, T., Chan, A. G., Freund, D., Shein, P., & Melkonian, D. K. (2009). "Explain to your partner": Teachers' instructional practices and students' dialogue in small groups. *Cambridge Journal of Education*, *39*, 49–70.

Webb, N. M., Franke, M. L., Ing, M., Wong, J., Fernandez, C. H., Shin, N., & Turrou, A. C. (2014). Engaging with others' mathematical ideas: Interrelationships among student participation, teachers' instructional practices, and learning. *International Journal of Educational Research*, *63*, 79–93.

Webb, N. M., Troper, J. D., & Fall, R. (1995). Constructive activity and learning in collaborative small groups. *Journal of Educational Psychology*, *87*, 406.

Wehlage, G. G., & Smith, G. A. (1992). Building new programs for students at risk. In F. M. Newmann (Ed.), *Student engagement and achievement in American secondary schools* (pp. 92–118). New York: Teachers College Press.

Whitehead, A. N. (1929). *The aims of education*. New York: Macmillan.

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist*, *45*, 267–276.

Yaron, D., Karabinos, M., Lange, D., Greeno, J. G., & Leinhardt, G. (2010). The ChemCollective – Virtual labs for introductory chemistry courses. *Science*, *328*(5978), 584–585.

Zimmerman, B. J. (1990). Self-regulating academic learning and achievement: The emergence of a social cognitive perspective. *Educational Psychology Review*, *2*, 173–201.

Appendix A: Data sources and frequencies

		Teachers' Intended Class Mode				
		P	A	C	I	Total
1	**Lesson Plans Designed**	2	19	21	23	65
2	**Lesson Plans Analyzed**		14	18	20	52
3	**Unique Activities** in lesson plan		29	29	47	105 (88 with written directives)
4	**Common Activities (each implemented twice)**		17	16	19	26 × 2 = 52 (17 with written directives)
5	**Segments of written directives** within 88 Unique Activities		66	74	92	232
6	**Segments of written directives** within the 17 Common Activities	2	20	10	14	46
7	**Worksheets with Student Products**		5	3	6	14
8	**Segments of oral instruction** within Unique Activities		49	75	57	181
9	**Student Products** taken (from 31 classes)		291 (8)	435 (12)	445 (11)	1,171 (31)
10	**Segments of student Products**		6,832	5,778	5,521	18,131
11	**Student Enactment** to actual questions		65	52		117

Appendix B: The variety of verbs identified from instructional directives for worksheet activities

Attentive Passive 6	Manipulative Active 58		Generative Constructive 28	Collaborative Interactive 9	Vague 10
Engage	Add	Keep track	Ask questions	Agree upon	Answer
Go through	Annotate	Label	Brainstorm	Answer with partner	Complete
Listen	Attack	List	Build	Debate	Email
Look	Avoid	Match	Come up	Discuss	Make
Observe	Bend	Measure	Comment	Exchange	Prepare

(continued)

Appendix. (continued)

Attentive Passive 6	Manipulative Active 58		Generative Constructive 28	Collaborative Interactive 9	Vague 10
Read	Break down	Move	Compare	Help	Respond
	Calculate	Name	Connect	Participate	Speak
	Categorize	Number	Construct	Share	Think
	Check	Order	Create	Work with group/	Work
	Choose	Paraphrase	Decide	partner	Write
	Circle	Pick	Defend		
	Click	Place	Determine		
	Complete	Plot	Draw		
	Confirm	Practice	Explain		
	Consider	Re-organize	Generate		
	Copy	Recall	Graph		
	Cover	Record	Justify		
	Cross out	Refer to	Plot		
	Delete	Review	Predict		
	Describe	Rewrite	Put/explain/write		
	Email	Round to	in own words		
	Expand	Show	Represent		
	Factor	Stimulate	Set goal		
	Fill in/out	Take down	Sketch		
	Find	Tape	Solve		
	Fold	Type	State		
	Follow	Use	Suggest		
	Guess		Summarize		
	Identify		Support		
	Include				
	Keep notes				

Data source: worksheets of 73 activities (for which worksheets were used).