

INQUIRY SCIENCE INSTRUCTION OR DIRECT? EXPERIMENT-BASED ANSWERS AS TO WHAT PRACTICES BEST PROMOTE CONCEPTUAL DEVELOPMENT OF SIGNIFICANT SCIENCE CONTENT

Abstract

The science education has overwhelmingly adopted a guided-inquiry perspective on science teaching. Nevertheless there remains debate about instructional approaches across a spectrum from direct through guided inquiry to open discovery. Questions about the efficacy of inquiry-based instruction linger. Proponents point to supportive studies, but critics counter that little of this research is sufficiently controlled. In various states there is political pressure for a return to direct instruction. The lack of convincing research for inquiry is thus of concern, hence with NSF/IERI funding we executed a 4-year experimental study with randomized subjects to test the efficacy question with regard to science conceptual development. In this symposium we describe the project, present and interpret the findings from both cognitive and practical perspectives.

Paper presented at the 2009 Annual Meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

William W. Cobern, David Schuster, Betty Adams, Brooks Applegate and Brandy Skjold,
Western Michigan University
Adriana Undreiu, University of Virginia's College at Wise
Janice D. Gobert, Worcester Polytechnic Institute
Cathleen C. Loving, Texas A&M University

We are posting this document to the web on June 26, 2009; however, our report remains in progress. A further developed version will be posted by fall 2009.
<http://www.wmich.edu/way2go/>.

**INQUIRY SCIENCE INSTRUCTION OR DIRECT?
EXPERIMENT-BASED ANSWERS AS TO WHAT PRACTICES BEST PROMOTE
CONCEPTUAL DEVELOPMENT OF SIGNIFICANT SCIENCE CONTENT**

William W. Cobern, David Schuster, Betty Adams, Brooks Applegate and Brandy Skjold,
Western Michigan University
Adriana Undreiu, University of Virginia's College at Wise
Janice D. Gobert, Worcester Polytechnic Institute
Cathleen C. Loving, Texas A&M University

This research studied the efficacy of carefully designed inquiry instruction in science against equally carefully designed direct instruction, at the middle school grades. *Inquiry teaching of science* refers to teaching that reflects the investigative attitudes and empirical techniques that scientists use to discover and construct new knowledge. The origins of the modern day concept of science teaching as inquiry lie with the 1960s NSF-funded curriculum projects. In recent years, under National Research Council, National Science Foundation and American Association for the Advancement of Science leadership, the United States has developed a national commitment to the teaching of science *as inquiry* across the K-12 grades, and almost all state frameworks for K-12 science education have an inquiry focus. The science education research community has overwhelmingly adopted an inquiry pedagogy perspective for science education.

With the advent of these curriculum projects and standards documents, various stakeholders such as teachers, researchers, curriculum developers, and policymakers have been interested in the effectiveness of inquiry-based curricula and teaching with respect to science concept achievement. Research and evaluation projects of wide variety have been carried out and proponents of inquiry teaching claim that this body of work generally supports the effectiveness of inquiry instruction. Critics however point out that little of this research has been unconfounded and thus the research support for inquiry instruction is weak. Given the widespread adoption of inquiry methods for science teaching, the lack of unconfounded experimental research data in support of inquiry is cause for concern. Hence, the purpose of our research was to test the efficacy of Inquiry Instruction vis-à-vis Direct Instruction in a controlled comparative study. Instructional topics chosen for the research were conceptually demanding yet were appropriate at the 8th grade level and consistent with national frameworks. The research utilized an experimental design with students randomly assigned to inquiry and direct instructional modes groups. Both treatment and control situations were specified in detail and hence are in principle replicable. The design of instructional units for each mode was based on accepted models that have support in the literature. Teacher fidelity to intended method was monitored by independent observers blinded to teacher assignments. This paper reports on the design and development of the instructional units and assessment, the research procedures and methodology, and the project findings.

Literature Review

The Rise to Prominence of Inquiry Instruction in Science

The origins of the modern day concept of science teaching as inquiry lie with the 1960s NSF-funded curriculum projects (Anderson, 2003; DeBoer, 1991; Krajcik et al, 2001; Rudolph, 2002; Schwab, 1962). The roots of inquiry run even deeper. The movement to bring science into the school curriculum began in the late 1800s and from the very beginning, advocates envisioned science instruction based on experience with the physical world, the gathering of data, rational argument and the drawing of inferences from evidence and data. Thomas Huxley spoke of scientific training as “*practicing the intellect in the completest form of induction...in drawing conclusions from particular facts made known by immediate observation of Nature*” (DeBoer, 1991). Laboratory activities became ubiquitous in science instruction on the basis that effective pedagogy must reflect the true nature of a discipline. Science as a discipline is *both content and inquiry*, “*the warp and woof of a single fabric*” (Rutherford, 1964); thus it is reasoned that science instruction must be more than the clear explication of information. It must include the investigative processes that lead to concept development. In short, science instruction should be largely experience-based and inductive.

Since the late 1800s we have had a century of shifting curriculum tides and recurring controversies, with few certainties about what science should be taught at K–12 levels and how it should be taught. However in recent years there has developed a clear commitment by most science educators to the pedagogy of inquiry. Indeed under National Research Council and AAAS leadership, the United States has developed a national commitment to the teaching of science *as inquiry* across the K–12 grades (AAAS, 1990; NRC, 1996, 2000, 2001). Almost all state frameworks for K–12, science education have an inquiry focus. The science education community has overwhelmingly adopted an inquiry pedagogy perspective for science education, including the National Science Teachers Association (NSTA), the National Association for Research in Science Teaching (NARST), and the Association of Science Teacher Educators (ASTE). An ERIC keyword search using ‘inquiry instruction’ (or inquiry teaching) and ‘science’ yields over 600 published articles between 1980 and the present. Similar searches in the General Science Abstracts and Web of Science also yield many hits. Inquiry is omni-present in the language of the science education community, and for most in the science education community, an inquiry approach has become the *sine qua non* for science teaching. As of 2002, Anderson asserted that research regarding the teaching of science had matured and thus “tended to move away from the question of whether or not inquiry teaching is effective, and has become focused more on understanding the dynamics of such teaching and how it can be brought about” (2002, p. 6), implicit is the assumption that the case in favor of inquiry is established. In 2008, Brady explicitly stated that “We know how to teach science” meaning of course that we know to teach science by inquiry. The only question left to ask appears to be, “when will we do it?” (Brady, 2008, p 607).

So just what is the Pedagogy of Inquiry?

Inquiry teaching of science refers to teaching that reflects the investigative attitudes, empirical techniques and reasoning modes that scientists use to discover and construct new knowledge. According to the *National Science Education Standards*,

Learning science is something that students do, not something that is done to them. “Hands-

on” activities, while essential, are not enough. Students must have “minds-on” experiences as well. The *Standards* call for more than “science as process,” in which students learn such skills as observing, inferring, and experimenting. Inquiry is central to science learning. When engaging in inquiry, students describe objects and events, ask questions, construct explanations, test those explanations against current scientific knowledge, and communicate their ideas to others. They identify their assumptions, use critical and logical thinking, and consider alternative explanations. In this way, students actively develop their understanding of science by combining scientific knowledge with reasoning and thinking skills. (NRC, 1996, p. 2)

AAAS espouses a similar view:

Both [AAAS and the NRC] want students working in teams, both want them raising questions and exploring ideas for themselves, both want students to learn to evaluate ideas using evidence. The pedagogy for science teaching, then, is one that actively engages students in reasoning about scientific phenomena. (Kennedy, 1997, p. 9)

The hard question is: how best to reflect this in instruction? Table 1 below from the National Science Education Standards (NRC, 2000) summarizes a spectrum of possible approaches to inquiry depending on the degree of learner or teacher direction. One can easily see that inquiry instruction may take on a variety of forms. Different forms of inquiry may be appropriate for different situations or stages of instruction.

Table 1

Teacher instructs so that the:	Variations			
1. Learner engages in scientifically oriented questions	Learner poses a question	Learner selects among questions, poses new questions	Learner sharpens or clarifies question provided by teacher, materials, or other source	Learner engages in question provided by teacher, materials, or other source
2. Learner gives priority to evidence in responding to questions	Learner determines what constitutes evidence and collects it	Learner directed to collect certain data	Learner given data and asked to analyze	Learner given data and told how to analyze
3. Learner formulate explanations from evidence	Learner formulates explanations after summarizing evidence	Learner guided in process of formulating explanations from evidence	Learner given possible ways to use evidence to formulate explanation	Learner provided with evidence
4. Learner connects explanations to scientific knowledge	Learner independently examines other resources and forms the links to explanations	Learner directed toward areas and sources of scientific knowledge	Learner given possible connections	
5. Learner communicates and justifies explanations	Learner forms reasonable and logical argument to communicate explanations	Learner coached in development of communication	Learner provided broad guidelines to use sharpen communication	Learner given steps and procedures for communication
<p style="text-align: center;"><i>More</i> ← <i>Amount of Learner Self-Direction</i> → <i>Less</i></p> <p style="text-align: center;"><i>Less</i> ← <i>Amount of Direction from Teacher or Material</i> → <i>More</i></p>				

Inquiry teaching is also used to achieve a variety of purposes. It is considered motivational for students. It is considered the ideal way to teach students about the nature of science and

scientific inquiry. It is, of course, also used for conceptual development, which is our area of research interest. Regarding inquiry teaching for conceptual development, perhaps the best known format for structuring a science lesson on scientific inquiry lines is the Learning Cycle, first described in 1962 by Atkin and Karplus (1962) and in various forms by Karplus and others many times since (see Lawson, 2004). The Learning Cycle divides a teaching sequence into three parts: exploration, concept introduction, and concept application. According to Karplus:

During *exploration*, the students gain experience with the environment, they learn through their own actions and reactions in a new situation. In this phase, they explore new materials and new ideas with minimal guidance or expectation of specific accomplishments. The new experience should raise questions or complexities that they cannot resolve with their accustomed patterns of reasoning. . . . As a result, mental disequilibrium will occur and the students will be ready for self-regulation.... The second phase, *concept introduction*, provides social transmission. It starts with the definition of a new concept or principle that helps the students apply a new pattern of reasoning to their experiences.... The concept may be introduced by the teacher, a textbook, a film, or another medium. This step, which aids self-regulation, should always follow exploration and relate to the exploration activities.... In the last phase of the learning cycle, *concept application*, familiarization takes place as students apply the new concept and/or reasoning pattern to additional situations. (Karplus, 1977, p. 173-174)

There are also variations of the Karplus cycle, such as the ‘5-E’ learning cycle, but these usually have at their heart the basic Karplus stages. The Learning Cycle informed our work (see Schuster, 2007), as will be discussed in a later section on the construction of teaching units for research purposes. Before that, we turn our attention to the equivocal research record in support of inquiry teaching for science conceptual development.

Problematic Evidentiary Support for a Pedagogy of Inquiry

Inquiry pedagogy is part of what the 2004 AAAS Forum on Science Teaching called “scientific teaching”, which “involves active learning strategies to engage students in the process of science and teaching methods that have been systematically tested and shown to reach diverse students” (Handelsman et al., 2004). Referring to college teaching of science, Handelsman et al add that

Many scientists are still unaware of the data and analyses that demonstrate the effectiveness of active learning techniques.” Active learning techniques can mean many things, but to the extent that the authors include a pedagogy of inquiry there is a clear claim here that such a pedagogy has been soundly tested and found effective. (p. 521)

However, for all the money that went into the 1960s NSF-funded curriculum projects, little was allotted to the scientific study of instructional effectiveness (see Welch & Walberg, 1972; Welch, 1976). Moreover, research evidence at the time and over the next decade regarding the effectiveness of inquiry as a strategy of science instruction was mixed and sometimes negative (Adams et al., 2006; Anastasiow et al., 1970; Ausubel, 1961 & 1962; Craig, 1956; Kersh, 1962; Shulman & Keislar; 1966; Tai & Sadler, 2009; Wittrock, 1964). Still later, the Ivins (1985) research findings failed to support the superiority of inquiry instruction for conceptual development.

Although the research findings of Shymansky et al (1983) and Shymansky et al (1990) on the effectiveness of the NSF-funded curricula using meta-analysis techniques were supportive of inquiry, their studies were also open to criticism. Many of the assumptions necessary for their meta-analysis studies were arguable. Moreover, the studies were not specifically focused on

inquiry instruction but were general curriculum program evaluations, thus leaving the central inquiry question open. More recent meta-analyses of inquiry science teaching have similar shortcomings (Springer, Stanne & Donovan, 1999). Recent research by Secker & Lissitz (1999), Secker (2002), Tretter & Jones (2003), Udovic et al., (2002), White & Fredrickson (1998) among others, are more focused on inquiry and again the results are supportive of inquiry, but even these studies do not yield inferences sufficiently unconfounded that one can feel that the central issue about inquiry instruction has been adequately addressed, similarly noted by EDC (2007 & 2008).

A few years ago, a dispute broke out amongst physics educators when Lamoreaux (2001) criticized aspects of work in Physics Education Research (PER). Some of his remarks were overstated but getting at something important about research methodology when he noted that:

Leaving the assessment to those involved in a particular project is probably not wise. We need an independent group to thoroughly study both the efficacy and student impressions; if the idea of PER is serious, the funding agencies supporting PER research should make funds for an independent assessment freely available... (Lamoreaux, 2001, p. 633)

Unconfounded empirical evidence – as sought by the Institute of Education Sciences at the Department of Education and the NSF/IERI Program – that the pedagogy of inquiry is efficacious with respect to commonly held instructional goals is in short supply (see e.g., Mervis, 2004); indeed, one recent study of the research literature notes a significant decline in research rigor from 1984 to 2002 (EDC, 2008). Much of the research on inquiry since the 1960s has been confounded by the following threats to validity.

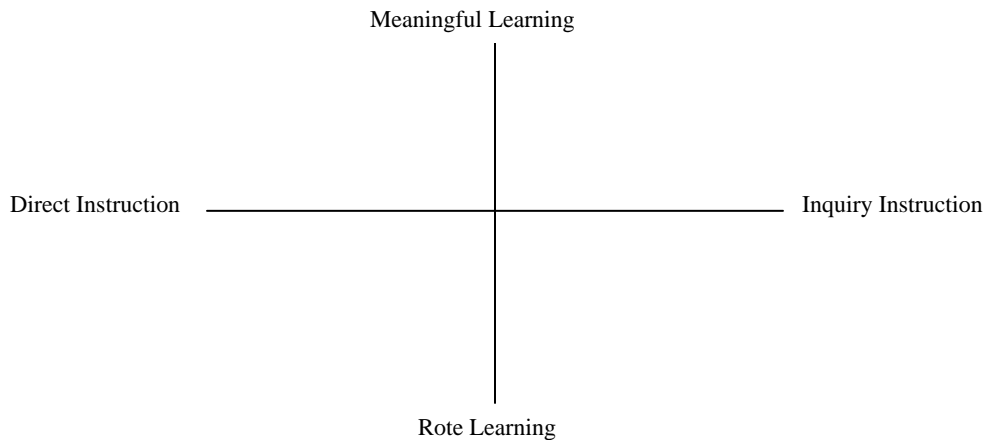
- 1) Comparative controlled studies which pit inquiry against worthy alternative instruction are few. There is often absence of a significant control. Too often the treatment is merely compared with poor or nebulous ‘traditional’ teaching so there is no fair and valid comparison.
- 2) As observed above, it is problematic to have evaluations that are not independent of the developers and researchers.
- 3) Few studies use randomized assignment of subjects to treatment groups or even quasi-experimental efforts to control for differences between subjects in treatment and control groups.
- 4) Too often there is an insufficiently detailed specification of treatment and control, in order that each may be exactly known and in principle be repeatable by other researchers.
- 5) In many reported studies, the treatment involves extended periods of teaching such as a semester. The difficulty here, and another threat, is that one cannot be sure what the ‘active agent’ is in a long series of lessons. The lessons, for example, may well be inquiry in nature but still involve other activities over the course of several weeks. In such a situation it is virtually impossible to know exactly what causes any improvement.

We elaborate on these five threats in our section on research design but the matter of a significant control requires immediate attention. In this research, our significant control, or fair test opponent, is Direct Instruction.

Direct Instruction

Direct instruction is the historic nemesis of inquiry teaching. When science education specialists work with teachers, they invariably talk about moving teachers away from behaviors considered

forms of direct instruction. The commitment to inquiry instruction is both understandable and laudable given that inquiry instruction has a natural appeal to anyone who loves both science and the teaching of science. Less commendable reasons for advocating inquiry are distorted caricatures of alternative modes of instruction, in particular expository or direct instruction. For example, it has been asserted that there was controversy over the *Benchmarks for Science Literacy* and the *National Science Education Standards* when they were first introduced due to their “focus on inquiry-based science - rather than memorization” (Brady, 2008, p 607). Whether anyone (especially science teachers) ever preferred instruction for memorization is dubious. The quote suggests that there are only two instructional options: instruction for memorization and instruction by inquiry, an opinion echoed by the *Committee on Prospering in the Global Economy of the 21st Century* (2009). The subtext here is that direct instruction as the opposite of inquiry instruction is thus instruction for memorization (e.g., Hein, 2004). This misrepresentation of teaching approaches was addressed relentlessly by David Ausubel (1961, 1962) more than forty years ago. He argued that the true issue was rote learning vis-à-vis meaningful learning, and that neither inquiry instruction nor direct instruction automatically leads to meaningful learning. Inquiry instruction can lead to rote learning; direct instruction can lead to meaningful learning, depending on how each is done. Novak (1976 & 1977) later adopted and elaborated Ausubelian views. We note that the Ausubelian literature uses the terms ‘discovery’ and ‘expository’ teaching where we use the more current language of ‘inquiry’ and ‘direct’ teaching.



The question is – is direct instruction potentially as effective as inquiry instruction for conceptual development and meaningful learning or even more so? As noted above, research evidence in favor of inquiry is inconclusive, not to mention some research evidence that appears to favor direct instruction (see for example, Anastasiow et al., 1970; Ivins, 1985; Schwartz & Bransford, 1998; Walberg, 1991; Wright & Nuthall, 1970). Moreover, direct instruction finds support in other areas of education such as developments from Distar, a 1960s project that was part of President Johnson's War on Poverty (Adams & Engelmann, 1996; American Federation of Teachers, 2003; Finn & Ravitch, 1996). According to these educators, effective instruction must be more teacher-led than student-directed. These views have an audience; in 2002, the recipient of the Council of Scientific Society Presidents’ annual award for Education Research was Siegfried Engelmann, who was cited for his research and development on direct instruction.

Perhaps the most prominent current dissenter from inquiry instruction is David Klahr (Chen & Klahr, 1999; Klahr, 2000 & 2002), who cautions that the “widespread belief that discovery learning is superior to Direct Instruction in early science education warrants careful empirical

assessment” (2002). Based on his research on the acquisition of science skills (such as controlling variables), he concludes that his

results challenge predictions derived from the presumed superiority of discovery approaches for deeper, longer lasting, and “more authentic” understanding of scientific reasoning processes, and suggest instead, a more *nuanced* examination of the most effective mixes and the most suitable matches between topic and pedagogy. (2002, p. 1, emphasis added)

Klahr’s assertions are not without support; see for example, Kirschner, Sweller & Clark (2006), Mayer (2004), and Sweller, Kirschner & Clark (2007). However, Klahr’s research involves open “discovery” approaches that are the most unstructured form of inquiry, and not the more *nuanced* and guided approach to inquiry as advocated by the NRC and AAAS, and his ‘direct’ mode arguably has some aspects of directed inquiry. Moreover, Klahr’s work is about acquiring certain science process skills not about science concept development. Thus on both counts, Klahr’s findings do not speak to the core question regarding inquiry instruction. Similarly, Sweller (2009) summarizes research findings regarding problem-solving and the use of worked-examples concluding that the superior effectiveness of direct instruction is supported by both empirical findings and cognitive theory; but he also seems to think that his conclusions legitimately extend to other learning outcomes such as concept development. Our view is that whatever are the findings of Klahr and Sweller in their respective areas of interest, we do not think that any conclusions can be drawn about conceptual learning unless there are research studies specifically addressing conceptual learning. Nevertheless, Klahr and Sweller draw attention to the precarious evidentiary support of inquiry instruction and that science Inquiry Instruction for concept development vis-à-vis Direct Instruction has not been subjected to experimental controlled studies comparable to Klahr’s work or the work referred to by Sweller.

We believe that the soundest conclusion from the research literature (from the early days of the NSF curricula to the present) is that experientially based teaching of science is more effective for conceptual development than non- experientially based teaching of science (Donovan & Bransford, 2005; EDC, 2008); not that specifically inquiry forms of experientially based teaching of science are more effective than direct instruction forms of experientially based teaching of science. The really interesting and significant question is not whether inquiry forms of experientially based teaching of science are more effective than non- experientially based teaching of science. That question is unequivocally answered in favor of inquiry. The really interesting and significant question is whether inquiry forms of experientially based teaching of science are more effective than direct instruction forms of experientially based teaching of science for concept development when both approaches are expertly designed and executed.

The research that we are now reporting cast Inquiry Instruction as the “treatment” approach versus Direct Instruction as the “control” approach. The specifics of instruction are discussed below, but in brief we attempted to test a model of Inquiry Instruction for conceptual development as advocated in the science education literature against a *worthy* opponent, that is, an expert model of Direct Instruction. We addressed this research question via field studies (see Anastasiou et al., 1970). The study of instructional effectiveness does not involve alternative approaches about which the science education community is neutral, or instructional approaches regarded as potentially *equal* rivals. The commitment to inquiry is deeply ingrained in the philosophical commitments of the science education community, thus any significant weaknesses inherent to Inquiry Instruction will have to be demonstrated in the field of actual practice. Moreover, an acceptable field study will have to ensure that the participating teachers use widely acceptable inquiry technique for concept instruction. Here we are fortunate to have

the work of the National Research Council (1996, 2000) upon which to draw. Even then it should not be expected that negative results from a competently conceived and executed field research would lead to a comprehensive rejection of inquiry. Such an event would be tantamount to a Kuhnian paradigm shift on the order of Ptolemaic cosmology giving way to Copernican—it happened, but not easily.

It is critical to bear in mind, however, that “received wisdom” should never be a barrier against good research. It is also important to recognize that teachers in the classroom are quintessential pragmatists, regardless of the philosophical commitments of scientists and science education professors, curriculum specialists, etcetera. Evidence of “what works” —especially if “what works” is shown to be feasible— is persuasive with teachers. Therefore, the results of a competently conceived and executed experimental field study that compares the outcome of good Inquiry Instruction with that of a worthy adversary, good Direct Instruction, will be persuasive with teachers. Moreover, tertiary level professionals will gain a better understanding of the strengths and weaknesses of Inquiry Instruction and of the long-term prognosis that some form of Inquiry Instruction is (or is not) the best method for achieving science education instructional goals with respect to the development of science conceptual knowledge, including knowledge of the nature of science and scientific inquiry. Of course, the entire science education community would justifiably feel vindicated by an unequivocal positive outcome in support of Inquiry Instruction. Anything less than this or (even a negative outcome) will lead first to more informed efforts at improving inquiry. More importantly, sound data and evidence will compel greater attention to the need for adopting data-supported instructional practices.

THE RESEARCH METHODOLOGICAL FRAMEWORK

Our research responded to the National Science Foundation’s Interagency Education Research Initiative (IERI/NSF 04-553) call for employing research and measurement designs that are demonstrably valid and reliable with a premium on experimental designs with random assignment of subjects (see American Educational Research Association, 2009). There is no need to rehearse the relative merits of quantitative and qualitative research, let alone the controversy, as this discussion is thoroughly documented elsewhere (e.g., Erickson & Gutierrez, 2002; Feuer et al., 2002a & b; Pellegrino & Goldman, 2002; St. Pierre, 2002). We think that quantitative research encourages precision of research (which we describe below) and that its findings can provide an important threshold, a floor, a common ground from which to pursue further research, including qualitative research, without becoming reductionistic. One does not need to give in to the extremes of reductionism to value the notion of “active agent”—so for example, what makes inquiry pedagogy “inquiry”? We know that there is a range of instructional activities that we call inquiry; what is that it that they all have in common that makes them “inquiry”?

Another way to look at precision is to consider the numerous threats to research validity. Precision is used to minimize such threats. We took it as a challenge in this research to protect against the numerous threats to validity that so often plague educational research. To meet this challenge we built into our research four critical features: A. specificity, B. fidelity, C. objectivity, and D. transparency.

A. Specificity

We address *specificity* of research in three areas on which the study depends critically, namely 1) The nature of inquiry and direct instruction, 2) The nature and quality of the instructional units, and 3) The nature and quality of the assessment, including its alignment with

unit objectives and content. In the sections below we specify the differences and commonalities between inquiry and direct approaches, discuss the design of parallel instructional units in the two modes, and describe the assessment requirements and features.

A1. Models for Inquiry and Direct Instruction

Single-word descriptors for instructional type, such as ‘traditional’, ‘inquiry-based’, ‘direct’, ‘lecture’, ‘hands-on’, ‘conventional’ etc, are inherently vague and ambiguous and thus open to various interpretations. To be clear and specific about what we actually mean by inquiry and direct science instruction we need to provide explicit *models* of instruction in each mode. These would include structure, components, sequencing and approach, thus making clear what is different between modes and what is common.

Below we describe the nature of our inquiry and direct instruction in terms of clear models with specified features.

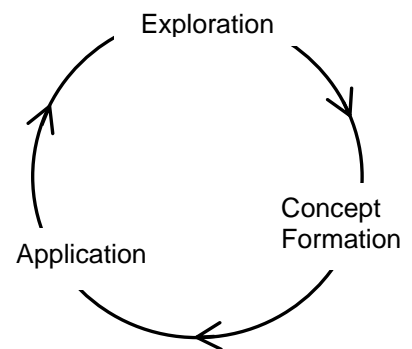
Model of Guided Inquiry Instruction

The Karplus learning cycle

Our model for inquiry instruction is the Karplus learning cycle, first proposed by Karplus and Atkin in 1962 and used extensively in various guises since then. The cycle has three major phases, as shown in the table and figure. Note that variations may occur in the names given to phases.

1. *Exploration*
2. *Concept formation*
3. *Application.*

The exploration and concept formation phases, in this order, together reflect a scientific inquiry approach. These phases are inquiry-based and inductive to a considerable extent (though not purely so). The application phase, where students develop the ability to apply the science and solve problems, is mostly deductive.



Model of Direct Instruction

In direct instruction, the theoretical concepts are presented and explained upfront by the teacher, and the main role of subsequent student activities is ‘illustration and verification of theory’. We call this instructional mode ‘direct active’; active because following instructor exposition there are student practical activities. Components of this direct model are as follows.

1. *Presentation and explanation*
2. *Illustration and confirmation*
3. *Application*

This sequence may also be viewed as a ‘learning cycle’ analogous to that for guided inquiry, but the term is not usually applied in the direct case.

Notes

The epistemological distinction between these inquiry and direct models is evident, residing in the character and sequencing of the first two stages of each model.

Questions and discussion occur in either model, though in inquiry it is likely to be guiding student thinking toward the desired conceptual understanding, while in direct it is likely to be clarifying or explaining concepts already provided to them.

Lab activities occur in each mode, although the approach would be mostly investigative toward theory for the inquiry mode, and confirmatory of theory for the direct mode.

Note that certain general practices would occur in any good instruction, whether inquiry or direct, and these aspects can be common to both modes, but the way the topic is approached differs and hence the respective roles, activities and discourse of teacher and students will also be different in the two modes.

Note also that our direct model is not simply transmission and passive reception of knowledge, which we might call 'direct passive'. Using the latter type of instruction as control would constitute a straw-man comparison.

Composite nature of all lessons

In reality no lesson or series of lessons is instructionally 'pure', i.e. homogeneous in all aspects with regard to intended method. All real lessons are a composite of elements, made up of many components over a teaching session or topic unit, whether the lesson is inquiry or direct or anything else, and each element or component of the composite lesson potentially contributes to the ultimate overall effects of the lesson.

Lessons viewed as composites can be dissected into identified constituent parts. Conversely, a lesson is constructed deliberately from known constituent parts. In very general terms these at least include: introduction, scene setting, what is spoken and how, explanation, demonstrations, activities and their nature, organization, participation, tasks, problems, assignments, assessment (formative and summative), discussion, questioning, dialogue, elicitation, discourse, and time on task, etc. Once we understand lessons to be composites, it becomes clear that even the most inquiry-oriented of lessons will not be homogeneously inquiry throughout in all aspects, and similarly for direct.

Thus no realistic unit will be 100% 'pure', in both essence and detail, whether designated as inquiry or direct. Nor would it be desirable or practical to attempt this; the result would likely be recognized as poor instruction generally. Nonetheless, even if there are no strictly 'pure' lesson modes, there is no doubt that there is a clear distinction, both epistemologically and practically, between inquiry and direct instruction. What then is the essential distinction to make between inquiry and direct modes, with respect to science conceptual development? We suggest that this resides in *how students come to the concept*. Do they come to it through exploration and concept invention, i.e. by a process of scientific inquiry, (science-in-the-making), or do they receive it as a finished product (ready-made-science), presented and explained by the instructor? Note that here we use the term *concept* rather broadly, to represent the scientific concepts, principles, laws, relations, or theories to be acquired in instruction.

Thus it is never a simple question to ask what is an inquiry lesson vis-à-vis a direct lesson; instead, we have to ask about the essence of inquiry and the essence of direct. In each case, given the composite nature of lessons, we must ask what the 'active agent' is regarding how a learner comes to the concept. Any research that hopes to minimize threats to validity must then

deliberately and specifically construct lessons using constituent parts in such a way that only the suspect ‘active agent’ is changed across the lesson types; stock lessons will not work.

In designing instruction then, whatever the mode, one first determines the central concepts to be learned, and whose understanding will be assessed, setting these as the main learning objectives. One then produces parallel frameworks outlining the inquiry and direct approaches to be taken, ensuring that the differences in instructional paths leading to the concepts are clear. This distinguishes ‘inquiry’ and ‘direct’ lessons, even if each of these is a composite overall, and is the ‘active agent’ which differs between the treatment and control groups in the research. Other elements of the composite lesson can be dealt with as appropriate for good instruction generally; one may perhaps view these elements as ‘carrier ingredients’, which need not necessarily be different for inquiry and direct instruction, as long as they are designed appropriately. Thus for example in an inquiry unit where the central objective is to investigate the relation between force and the motion, it is quite appropriate for the instructor to specify the object to be used in the activity (e.g. student sitting on a skateboard) and how to apply the force, rather than have students free to come up with any objects and methods they wish and try them out, ‘by inquiry’. They don’t have the basis for sensible choices at this stage, it is not the current goal, and the ensuring confusion will divert time and attention from the central objective, and thus take far longer than the corresponding direct lesson. Thus in both inquiry and direct modes the instructor can specify such ‘practical’ matters, and only the active agent needs to differ. Thus in the dynamics example above, the essential difference resides in whether the students *infer* the force-motion relation from their investigations, or are *told* the relation and then confirm it. The aim then for both instructional design and research is to focus on those essential elements of a composite lesson that distinguish inquiry and direct modes. This is what would make the learning outcomes potentially different, rather than a summative idea of ‘inquiry’ or ‘direct’ applied to the whole lesson generally. The assessment too, would focus specifically on those central concepts that were approached via the alternative modes.

‘Mixed’ instructional modes?

While a mixture of instructional strategies occurs on various *details* of instruction during any composite lesson, the idea of developing the *central concept* of a given lesson by a ‘mixture’ of modes makes little sense – either the students explore first then infer, or they are told first then confirm. A mixture (combining inquiry and direct instruction for the central science concept) would be neither fish nor fowl with respect to the ‘active ingredient’. It would be epistemologically (and educationally) unclear, probably situational and episodic, without theoretical warrant or guiding model, and would not help to answer the research question. We emphasize that since the ‘carrier’ ingredients of a lesson may be done in the most appropriate manner, a lesson might look something of a mixture overall to the casual observer. But this is not what we mean by mixed methods *with regard to the central concept*, and is simply a practical reality of all lessons. The nub of the question is: of the various lesson elements, what is the active ingredient we are after, and how shall that be *approached*?

A2. Development of the Instructional Units

The research study clearly depends on the nature and quality of the instructional units, else any research conclusions about efficacy would be questionable. Thus it is important that units be designed according to explicit criteria based on stated models of good instruction in each mode.

Chosen Topics

For our instructional units we chose two science topics that have substantial conceptual demand and are known to give students difficulties, namely force and motion and the seasons. Both units were conceptualized from the ground up as far as content development, approach, narrative and pedagogy were concerned, to constitute two-week instructional units. The criteria and principles for unit construction are described in a later section.

The two units are:

1. *Its Dynamic! – The relation between force and motion.* Deals with the concept of force and how motion relates to net force and mass – Newton’s first and second laws of motion.
2. *Its Illuminating! – Sunlight and temperature variations on earth.* Deals with some basic science (how light energy received depends on angle, distance and time), and its application to temperature variations on Earth, i.e. variation by location (latitude) and by time of year (the seasons).

Each unit was developed in parallel Inquiry and Direct versions.

Objectives

For each unit we produced both broad domain objectives and detailed topic-specific learning objectives. These were carefully devised to specify the level and range of conceptual understanding of the important science concepts which were to be the central focus of instruction. The objectives guided unit development, teaching and assessment. Objectives, instruction and assessment were all closely aligned and consistent with one another. This type of ‘overall curriculum coherence’ was important to achieve the goals of the research unambiguously. Sets of objectives for each unit are available online.

General requirements for good science instructional units – in either mode

There are general design considerations for developing good science instructional units, which would hold irrespective of mode. These include the following.

- *Substance.* The science topic and the target concepts should be substantial and important.
- *Conceptual.* The treatment should be mainly conceptual (but conceptually demanding) rather than being fact- and formula-based.
- *Learning challenge.* The topic should pose known learning challenges, e.g. where alternative ‘conceptions’ are prevalent.
- *Standards.* The topics should be in the US national standards and benchmarks.
- *Objectives.* Learning and assessment objectives should be clearly stated, in both broad terms and specific to the unit.
- *Coherent topic development.* The topic and concepts should be coherently developed in a logical sequence over several lessons, rather than occur as fragmented activities for example.
- *Focus.* It is important in any instruction, but particularly instruction with hands-on activities, to focus attention on the underlying message, rather than just on practical details or fun aspects, in order to make meaning of the activities.
- *Application.* A measure of students’ meaningful grasp of a concept is the ability to *apply* their knowledge, to answer questions, explain, and solve problems, in new but related situations. Application is thus a component of instruction (as in the models) and of course of assessment.

- *Length.* A unit should be of sufficient length to allow building up a substantial conceptual structure and promote students' ability to apply it. This requires several lessons at least. Note that individual lessons or parts of lessons are usually too short to meet the substance, challenge, coherence and application requirements. On the other hand whole programs, such as a semester course, contain many different topics, and can only address the global question of whether or not 'Curriculum A' is more somehow effective than 'Curriculum B; this is of course valuable but cannot easily target specifically what it is that leads to differential outcomes. Hence, for research purposes on instruction and learning in science, the specificity goal suggests that sets of lessons are preferable to individual lessons and to courses.
- *Engagement.* Student engagement is important so the units should preferably have an interesting storyline with hands-on and minds-on activities, supportive of the conceptual development of course.
- *Reflection and consolidation.* Reflection or 'debriefing' should occur at the end of each significant activity or set of activities, to identify the essence of what has been learned, pull the threads together, and consolidate.

Equipment for the activities and printed materials for the units

Once the science units were designed, the necessary equipment was ordered or constructed. Practical experience trying out the activities often led to revision of lesson details, or new ideas to incorporate.

Writing the lessons for the units, in equivalent parallel modes, along with teacher notes and narrative, was an unexpectedly demanding undertaking, which took much care, time and multiple revisions. This was essential as the units were the basis on which the research depended. Materials were produced for each unit, in both modes, and in both student and teacher versions. Student versions were essentially in worksheet form, while teacher versions had in addition alternating pages containing teaching notes and suggested teaching dialog, in the appropriate mode. Note that this teacher 'narrative' is intended for use in teaching preparation rather than to be used verbatim in the 'live' situation, since teaching dialog should flow naturally and flexibly in the classroom. Overhead projection sheets were also produced for teachers to use for selected aspects of lessons. (The sets of student and teacher materials for each unit are available online: <http://www.wmich.edu/way2go/>).

Comparing times for inquiry and direct modes

The conventional wisdom is that inquiry science teaching takes much more time than direct. Time-on-task is potentially a problematic issue when one wishes to test the comparative efficacy of two models with respect to conceptual development. One might not be able to convincingly claim for example that inquiry is superior to direct if the outcome might be due simply to the prosaic point of having spent more time on task rather than on some essence of inquiry. There are two issues here: does inquiry really take (significantly) more time than direct, for the specific instructional models we use, and to the extent this may be so, how can it be addressed?

Once one views and designs a lesson as a sensible composite with identified active agent, we find that the time issue, while still present, is not nearly as significant as is often claimed. To take a specific example, consider the target concept of the relation between force and motion (Newton's second law). Student practical activities, which take a fair time, would in fact be very similar in the two modes – in inquiry they would explore the phenomenon (person being pushed on a skateboard) and propose the relation from their observations, while in direct they would be told the relation and asked to verify it via the same basic activity. Of course the mindset is

different between modes, and thus the classroom discourse, but the actual hands-on activities take similar times either way. Thus some of what is believed about time comparisons stems from inadequate (or no) specification of what the instruction actually looks like (or should). It is quite clear that relatively unguided inquiry, open discovery, or attempts to do every subsidiary detail by inquiry, would lead to substantially longer times – but this is not what we would advocate and not what we test in this study. A properly constructed guided inquiry lesson should not take valuable time away from the ‘focus’ learning objective by trying to have students do absolutely everything, peripheral or not, by ‘inquiry’.

Nevertheless, we do find that our specified guided-inquiry instruction takes about a tenth longer than our active-direct instruction, a difference of about 5 minutes in 50. Of course we believe that in inquiry mode students have in addition gained invaluable experience with the nature of scientific inquiry. Nevertheless, leaving that aside for the moment, for research comparison purposes one might think of how to use the extra time in direct instruction; but how is that to be done without creating further differences? For example, if direct instruction is extended by adding drill and practice, time-on-task has been equalized, but a new variable has been introduced between approaches. To us it seems that the time difference is comparable to normal variations that occur between lessons depending on how things go, and that the ‘added value’ benefits obtained from a guided inquiry approach, beyond science content acquisition, are well worth it.

A3. The Assessment

Clearly the nature and quality of the assessment for ascertaining student attainment of science understanding is of crucial importance for the validity of the study, just as was true for the instructional units. Besides the quality requirement, the assessment needs to be closely aligned with the learning objectives and unit content. The nature of the assessment questions embodies our criterion for concept understanding: demonstrating the ability to *apply* the concept in relatively new situations, to explain, predict or solve conceptual problems. This is not factual knowledge or formulas; we aim at Bloom taxonomy levels 2 and 3, comprehension and application. Of course for comparing efficacy, the assessment must be the same for both instructional modes. The same tests are administered pre- and post-instruction by the project evaluators, in order to determine performance gain, both overall and on individual aspects of the course.

Note that it is important that the teachers implementing the lessons be blind to the assessment. Otherwise they might teach to the test, either inadvertently or deliberately, thereby thwarting the research goal. This requirement means that although the assessment is closely aligned to instruction, and students may have encountered questions of a similar (generic) type, the particular test questions must not be seen in advance by either teachers or students

Assessment design considerations and characteristics

- For the purposes of the research we decided to use an objective format, consisting of high quality multiple choice (MCQ) items, for ease and consistency of administration, scoring and analysis and comparison, for large numbers of students
- The assessment is closely aligned to the objectives and instruction; however the particular questions in the tests are ‘unseen’ to the students.

- The questions involve primarily science content understanding, tested via the ability to apply the concepts to new but related situations.
- Questions focus almost entirely on the important targeted concepts in the objectives, i.e. on the ‘active essence’ which is approached differently in the two instructional modes
- Some ‘generic’ aspects of science may also be assessed, e.g. control of variables, where this occurs in the topic experiments.
- Questions are conceptual rather than about factual knowledge, recall or calculation (except where simple mental arithmetic may be a way to test conceptual understanding).
- Questions are carefully formulated, tested and refined. The question stem is carefully worded to describe the situation clearly and unambiguously, preferably illustrated with a diagram. MCQ options provide plausible choices, with some distracters representing common ‘alternative conceptions’.
- Students are also asked to indicate their level of confidence in their answers. It is useful to know for example that a student is very confident in a wrong answer, rather than just giving ‘my best guess.’
- The compiled tests for each unit consist of about 24 MCQs. Since unit content proper spans about 6 lessons, there are roughly 4 MCQs on each lesson, each with a confidence level indicator. We would like more comprehensive assessment, but time and attention considerations also come into play, and the essential concepts seem to be adequately assessed.
- Note that the project does not involve explicit instruction or assessment with regard to scientific inquiry, although clearly the students in the inquiry classes experience it implicitly during their lessons. For them this may be regarded as an ‘added value’ benefit which is not assessed explicitly in the content tests. (The tests for each unit are available online: <http://www.wmich.edu/way2go/>).

B. Fidelity

Teacher *fidelity* to treatment and control assignments is the second critical feature of our design. In this research we followed the rule of “prepare and verify.” To begin with experienced teachers were recruited to the research by first asking them about what forms of instruction they were comfortable with. We wanted experienced teachers so that our teachers could concentrate on their assignments more than non-instructional classroom and student matters. We also sought cooperative teachers who were interested in the research and willing to take on the instructional roles assigned to them. Working with the teachers we assigned the teachers as inquiry or direct (treatment and control, respectively) according to their levels of comfort. We rejected the idea that teachers be asked to teach both instructional methods out of the reasonable concern that teaching both methods within a short time framework would increase the challenge of maintaining fidelity to method.

Teacher preparation was facilitated by the fact that the teachers all had prior experience teaching the concepts in our research units as these concepts are common to middle school curricula. Several months prior to the summer trials, the teachers each received detailed scripted lesson notebooks for each unit. The teachers were asked to study the notebooks prior to the first preparation meeting with the researchers. At the initial preparation meeting the researchers

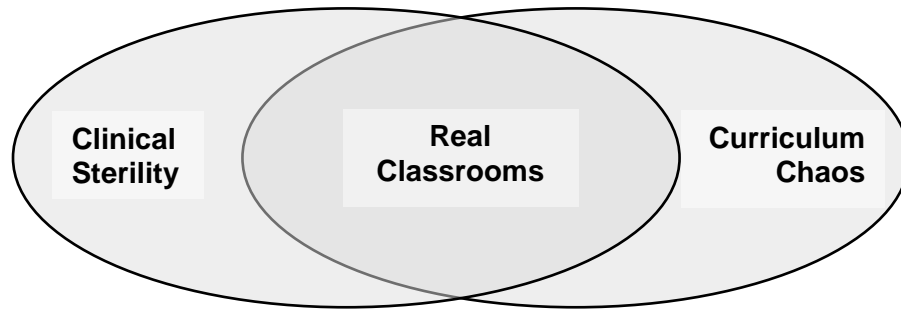
introduced the equipment and activities by demonstration, and then the teachers practiced with the equipment under the supervision of the research team. At latter meetings, the Direct and Inquiry teachers met separately with a “coach” from the research team to go through the units line-by-line. The preparation meetings came to about ten hours per unit prior to their first use. During the trials, the teachers meet after each day to debrief and prepare for the next day. At these times any lingering questions or concerns were addressed. The trials were videotaped so that the teachers could review their work prior to subsequent trials. The teachers were also given copies of the blind observation reports to consult (see below). Finally, preparation meetings continued each spring prior to the June trials so that teachers could improve their execution of the lessons.

For *verification*, teacher fidelity was monitored in three ways. After each day of a trial, the teachers posted journal notes on how the teaching went, how the students responded, and where they may have deviated from the script. Lessons were also video taped. We reviewed the tapes both as a way to monitor fidelity and for development purposes. The teachers had access to the tapes after a trial so that the tape could help the teachers prepare for the following trials. The most important way that we monitored fidelity was with independent, “blind” observers. The project contracted with an evaluation group that was accustomed to observing and evaluating science instruction, the *Science Mathematics Program Improvement* group (SAMPI) which is nationally known for its expertise at science/math curriculum and instruction evaluation (Jenness & Barley, 1999). The SAMPI observers were initially “blind” to teacher assignment to method and they entered classrooms unannounced to the teachers. After two trials, however, and because we did have good teacher fidelity to method, the observers figured out who were the inquiry and who were the direct instruction teachers. For trials 3 and 4, the SAMPI observers were given the teacher notebooks and scored the teachers specifically on fidelity to the scripted lessons. Thus, we were able to independently verify fidelity.

It is important to specify what our fidelity expectations were for the teachers, as there are limitations not just on what is possible regarding fidelity but even what is desirable. Reading the scripted lessons to the students would be like an actor reading a script. It is not what we consider true “acting”; fidelity to method is high but effectiveness is not. Effectiveness requires performance, a sense of personal ownership. One does not go to a play to see actors read a script but to see them “act” the roles described by the script. Watching actors read a script has little appeal for audiences. A teacher who “reads” the lesson as her way of teaching is just as uninteresting. Teaching is some what like a performance. We do not want the teacher to read the lesson but to perform the lesson just as an actor acts out a script as performance. The analogy is of course limited since good teaching is more than performance. It is also interaction with students. The point is that a performance has life; reading a script is wooden. The life that is in a good performance is actors with some liberty to personalize a role. The reality of classrooms thus is that one particular curriculum can be enacted very differently by two different teachers depending on just how much they personalize the curriculum.

On the other hand, excessive personalization creates chaos as the actors’ personalization of roles begins to conflict with the script’s themes. So, while in a research setting we need the liveliness of a good performance otherwise there is little hope of student attentiveness and effort, the script has to set boundaries that keep the teachers actions within the parameters of the research. The script and the themes of the play set boundaries for actor-innovation. It does not work to have Hamlet cracking jokes, but neither does it mean that any two actors will play the role of Hamlet exactly the same way. So it is with teacher fidelity to method.

As pictured below, the extremes of clinical sterility and curriculum chaos must be avoided if research findings are to be generalizable to real classrooms.



For research purposes one could take a clinical approach where the researcher attempts to control absolutely every possible variable, in other words, to create a clinical sterile environment for the research; but such an environment would be foreign to the real world of classrooms and thus diminish the potential usefulness of any findings. This fact of classroom reality is a serious limitation on all quantitative research in education. On the other hand, if no efforts are made to control the research environment, or the efforts are ineffective, the quantitative test becomes pointless. Researchers must astutely split the difference, seeking realism with neither sterility nor chaos.

All classrooms have variability. As noted above, two teachers may both be experts and still enact a common lesson differently. Even the same teacher is likely to enact the same lesson differently on different occasions. Students can be unpredictable behaving one way one day and different the next. For different subject topics, teachers may vary their teaching behaviors and students may respond differently. We refer to this as the “natural background variation” that exists in all teaching situations. Hence, our interpretation of fidelity as scored by the SAMPI observers was that under blind conditions, independent observers are able to identify instructional type within natural background variation.

There are of course other ways to deal with natural background variation. If the number of teachers and students participating in a trial are raised to a very large number (e.g., 1000 teachers and 30,000 students) then even very small differences can become statistically significant. However, it would be impossible to validate fidelity in that number of classes hence scaling up also means a substantial loss of experimental control. Moreover, in such a large study and with a small difference, one will actually have a substantial number of cases where both methods achieve comparable results, and some cases where the opposing method actually works better. Hence, true differences when small are not likely to have any practical classroom significance. The other way to reduce natural background variation is to have only one teacher, which at least reduces teacher variation. If one teacher teaches two methods to the same students, then student variation is reduced. This approach introduces new factors, for example, the experimental design would need to account for two units and two approaches, not merely two approaches. Even so, generalizing from one teacher’s experience is the real problem. This experimental design falsely assumes that natural background variation is not an issue in the real world of classrooms.

The critical consequence of natural background variation, therefore, is that any instructional approach must be robust with regard to natural background variation for it to be considered generally effective, as in “what works.” If instructional efficacy depends on exact, detailed execution then any potential gains are unrealizable in typical classroom settings due to natural background variation. Furthermore, the differential effectiveness of different pedagogical approaches must also be observable above natural background variation. Otherwise, even a true

difference between methods has no practical classroom value because the true difference can only be realized under the improbable situation of very low natural background variation. We will consider the influence of natural background variation in our discussion of findings below.

C. Objectivity

The third feature of our design is *objectivity*. Our research design embedded several areas of “blindness” in order to minimize bias during assessment development, evaluation of the assessment results, and verification of teacher fidelity to method.

- Teachers implementing the lessons were blind to the assessment, as noted in the assessment section above.
- Evaluation of student outcomes: SAMPI coded and analyzed student assessment data without knowledge of group assignments.
- Teacher fidelity: SAMPI observers, blind to teacher assignments, visited two instructional days per unit for each teacher. Each teacher was seen by two observers. The observers documented the instruction and then scored fidelity to method. Final documentation of instruction and fidelity scores were arrived at by consensus during a group meeting of the observers after the last observations were completed.

D. Transparency

The fourth critical feature of our design is *transparency*. Transparency of research is necessary for readers to know exactly what the research did and thus for any attempts at replication of the research. For example, readers of research often have access to no more than a brief description of a treatment or intervention. We have made our research as transparent as possible by placing all critical information on our webpages at: <http://www.wmich.edu/way2go/>. Specifically these pages contain the complete sets of instructional objectives and knowledge domains, instructional guides for teachers, guides for students, and the assessments. By making our units transparent, anyone can see exactly how this research implemented the theoretical ideas regarding the best practices of inquiry and direct pedagogy. Moreover, the teacher fidelity reports are also posted.

THE RESEARCH

Subjects and Setting

The subjects in the research were 342 incoming 8th graders from several Midwest urban, suburban, and rural school districts who participated in research trials run during the summers in 2006, 2007 and 2008. We used an 8-day format, 9am to noon, four days a week over the last two weeks of June. Neighboring school districts participated by sending out program announcements to current 7th graders each spring. As a result, participation in this project was a matter of “family selection,” that is, families decided whether or not a child was to participate. Our project teachers were veteran science teachers and their personal assessment was that the students attending the trials were not noticeably different from the student composition of their academic year courses in middle school science. We chose to work with a summer program so as to facilitate the implementation of the random assignment of students to treatment group. At the time our project was funded, there was concern at the National Science Foundation that research trials should be run with randomization at the subject level rather than at the classroom level.

The summer format also was also suited to the control of time-on-task.

We elected to work with 8th grade middle school students because: 1) the middle school years are transitional between elementary school and high school. As such they are critical to success at high school; 2) the middle school years begin the formal study of science including complex science concepts requiring the use of mathematics. We worked with physical science content because the physical sciences are widely found in middle school curricula and because physical science units were amenable to our two-week summer format. We choose to work with two instructional units because we do not think that it should be assumed that instructional pedagogies and outcomes are not influenced by content topic areas.

Our approach has its limitation. Students attending a summer program are there without grade pressure. Moreover, in a voluntary summer program, homework and reading assignments are unrealistic. Hence, the learning gains are solely dependent upon in-class student engagement with the lesson activities and discussions. For a number of reasons, as we will note in our discussion of findings, the learning gains in a summer arrangement such as this may not be as large as those one would expect during the regular academic year.

Research design: topic units, instructional modes, student and teacher allocations

The research involved two different topic units, each taught in direct and inquiry modes, with students being assigned at random to either inquiry or direct classes. As noted, we had recruited four experienced middle school science teachers and assigned them to teach either Direct (DI) or Inquiry (II), based on their self-assessed comfort with an instructional approach. Each teacher taught both Light and Dynamics units, in the two daily sessions. A fifth/substitute teacher became involved in the project due to temporary circumstantial necessity, and was retained to allow for the participation of additional students each year.

Respective learning gains were assessed by comparing the scores on identical pre- and post-tests for each topic. Raw gain scores are reported below (gain in percent correct), as well as normalized gain scores (percent correct over possible gain in percent correct). The normalization check is conducted for the purpose of minimizing any “gain advantage” that might be due to low prescores that stem merely from lack of exposure to the topic.

DATA AND FINDINGS

Teacher Fidelity

The SAMPI observation protocols provided a fidelity score between one and seven where a “7” indicated maximum fidelity to instructional method and a “1” indicated little or no fidelity. Our target was to have teachers in the 5 to 7 range, with anything under “4” unacceptable, where four meant that at least the method was identifiable.

One must be careful to avoid the often hidden assumption that learning will automatically improve with teacher fidelity to method. Learning gains are only positively correlated with fidelity to the extent that the method works. To put this another way, improving fidelity to a teaching method that itself is flawed cannot be expected to improve learning. We also do not know how sensitive a method is to variation of fidelity, and we certainly do not know that two methods being compared have equal sensitivities to variation of fidelity. Hence, we avoid making unsubstantiated assumptions but limiting our definition of fidelity to the method being identifiable by an independent, method-blind observer. Again we emphasize the point that a method may work very well but it is only as useful as it is robust against natural teacher

variation.

Of special interest is how well these results actually reflect the research goal of testing two “same but different” science topic lesson sets, attempting to isolate the effect of specific contrasts between Direct vs. Inquiry. The independently assessed research measure of “Teacher Fidelity to Intended Instructional Model” is of value, and yields a factor for each teacher that generally represents their degree of adherence to the Direct or Inquiry teaching models which are being tested for relative efficacy. Each teacher’s “Fidelity Multiplier” consists of the average score from four classroom observations, on a scale from 1 to 7, and the descriptive statistics for these from 2007 and 2008 are as follows:

Table 2

Fidelity scores 2007/2008	
N	9
Mean	.8333
Median	.8570
Std. Deviation	.1226
Variance	.015
Range	.429
Minimum	.571
Maximum	1.000

As noted earlier in this report, science education research must bear in mind the realities of classrooms and teachers. Hence, we did not want absolute, reductionist experimental controls and clinical sterility; but the opposite would also not have been acceptable, that is, spontaneous, uncontrolled, arbitrary and/or random classroom events. Our teacher fidelity-to-method data show a median score of 86%, which we argue is adequate for our research purposes while remaining realistic with respect to inevitable teacher differences and idiosyncrasies in real science classrooms. We also note that all classrooms showed statistically significant knowledge gains, which is consistent with the findings of Taylor et al., (2007, p. 44) that there is a “strong relationship between fidelity of curriculum implementation and student learning gains.”

However, the fidelity scores for the Direct Instruction teachers were consistently higher than the scores for the Inquiry Instruction teachers. This was not unexpected as inquiry instruction is inherently more difficult to execute than direct instruction even when the direct instruction is experientially-based as was ours. And, although it may be tempting to think that improving the fidelity of the inquiry instruction would lead to statistically better results vis-à-vis direct instruction, that temptation must be resisted. The Direct Instruction teachers also have room for improvement and so what is more likely is that improved fidelity of practice across all teachers improves student outcomes, but not the difference between the two methods. Nonetheless, teacher fidelity will once again be addressed as we prepare for our summer 2009 and 2010 trials.

Student Performance Data

In this section we provide and discuss student performance data for science content understanding, as assessed by pre- and post-instruction multiple-choice tests. Data are presented by topic, year, instructional mode and teacher. Pre- and post- mean cores are charted, along with the mean gain. We then compare performance gains to see whether and how these might depend on instructional mode, as well as on the other factors involved. We test whether any differences are statistically significant, given the spread of student scores and the variability of factors

present in a field teaching environment. Observational data and scores on teacher fidelity to instructional mode are also presented. We summarize and comment on the data and findings.

We also discuss important insights gained during the project into the multiple aspects involved in this educational research endeavor, such as instructional design considerations, materials, assessment, teacher development, classroom implementation, intended and implemented curriculum, research design and methodology, etc. This has implications for both instruction and research, and also suggests issues for further research.

Pre- and post-test results and gains

Below we present bar charts of student performance data for science conceptual understanding, as assessed by pre- and post-instruction multiple-choice tests. Data for each of the five classrooms are labeled by teacher. The mean gain is charted next to the pre-and post mean scores. Charts are presented for individual classes, labeled by teacher and mode, as well as an aggregate for each mode. Mean pre- and post- scores are shown by the bar height, with error bars spanning ± 1 standard deviation from the mean.

A. Performance data for Light 2008

The bar chart in Figure 1 shows pre- and post-test mean scores, as well as the mean gain between pre and post, for the Light unit in 2008. Data for the five classrooms are labeled by teacher. The lines on the bars are at one standard deviation above and below the mean, to give a visual indication of the spread of scores obtained. The aggregated mean scores for each instructional mode are shown in the middle of the figures.

B. Performance data for Dynamics 2008

The bar chart in Figure 2 shows pre- and post-test mean scores, as well as the mean gain between pre and post, for the Dynamics unit in 2008. The format is the same as for Light 2008.

C. Performance data for Light 2007

The bar chart in Figure 3 shows pre- and post-test mean scores, as well as the mean gain between pre and post, for the Light unit in 2007. The format is the same as for Light 2008.

D. Performance data for Dynamics 2007

The bar chart in Figure 4 shows pre- and post-test mean scores, as well as the mean gain between pre and post, for the Dynamics unit in 2007. The format is the same as for Light 2008.

E. Pooled Performance data for Light 2007-2008

Figure 5 shows the frequency distribution of raw scores (% correct) for Light Unit pooled over two trials, 2007-2008.

F. Pooled Performance data for Dynamics 2007-2008

Figure 6 shows the frequency distribution of raw scores (% correct) for Dynamics Unit pooled over two trials, 2007-2008.

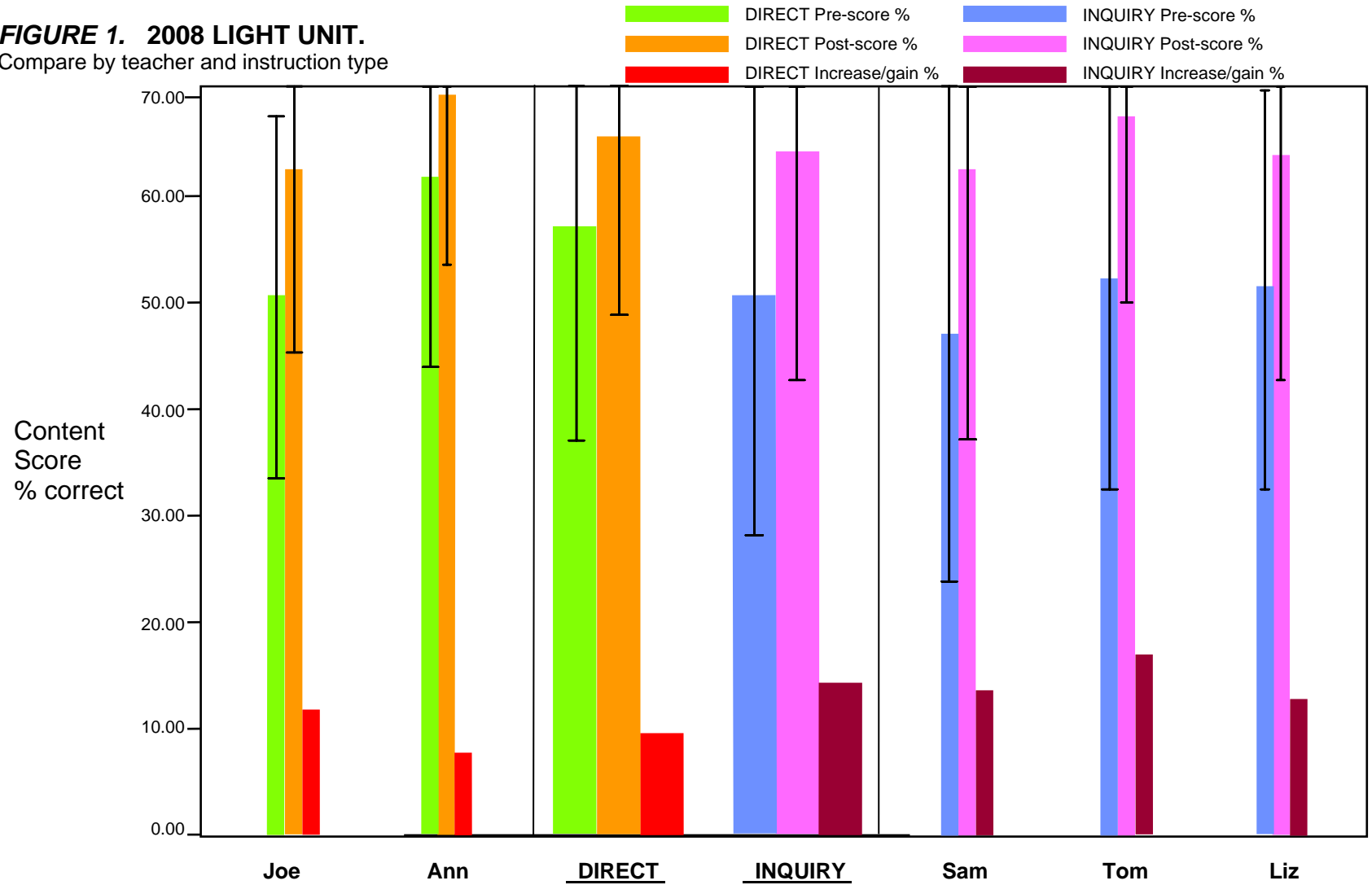
G. Pooled Comparison Performance data for Dynamics 2007-2008

The bar chart in Figure 7 shows pre- and post-test mean scores, as well as the mean gain between pre and post, for the Dynamics units in 2007 and 2008, along with the comparison of Direct and Inquiry data. The format is the same as for Light 2008.

F. Pooled Comparison Performance data for Light 2007-2008

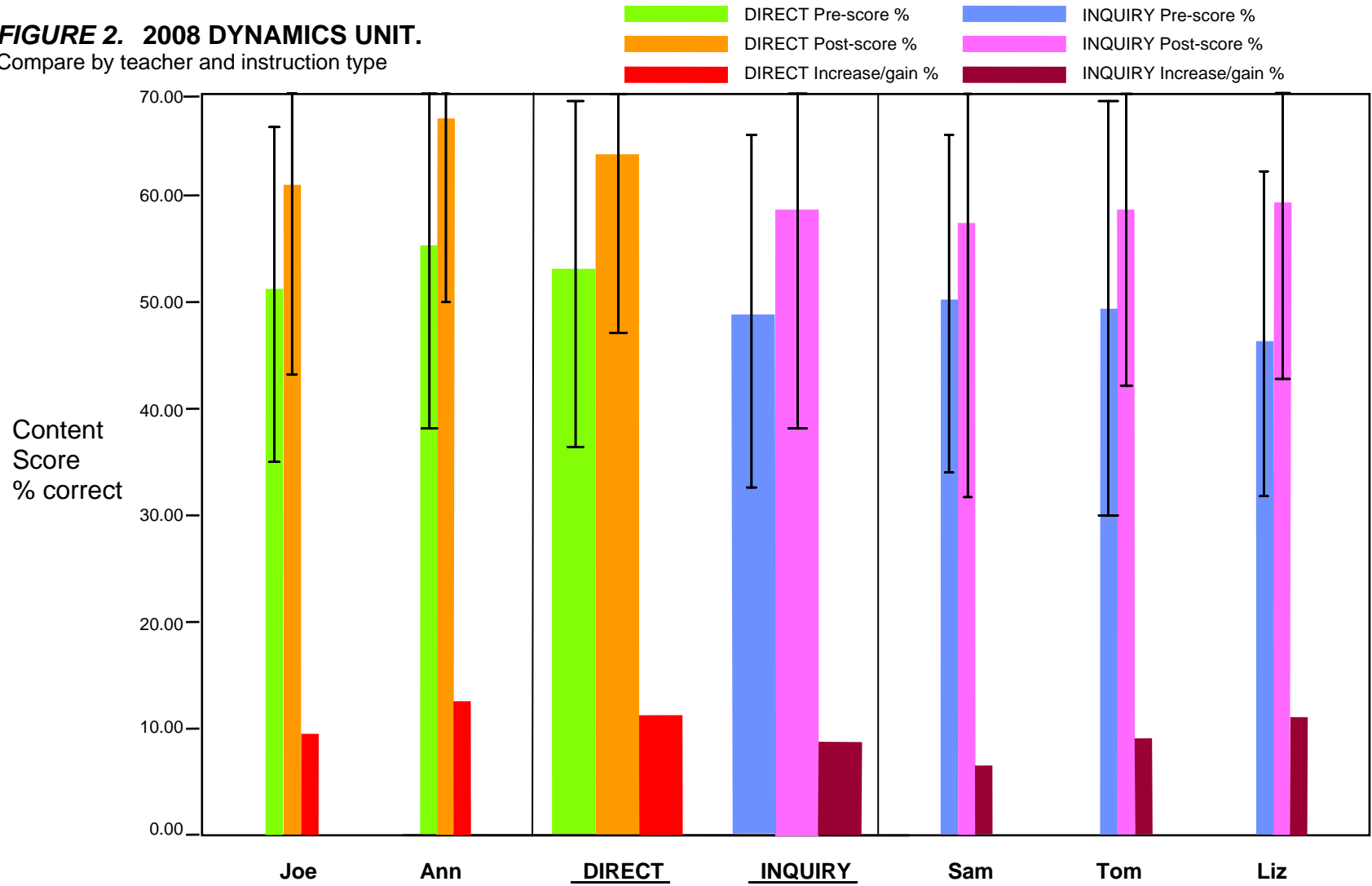
The bar chart in Figure 8 shows pre- and post-test mean scores, as well as the mean gain between pre and post, for the Light units in 2007 and 2008, along with the comparison of Direct and Inquiry data. The format is the same as for Light 2008.

FIGURE 1. 2008 LIGHT UNIT.
Compare by teacher and instruction type



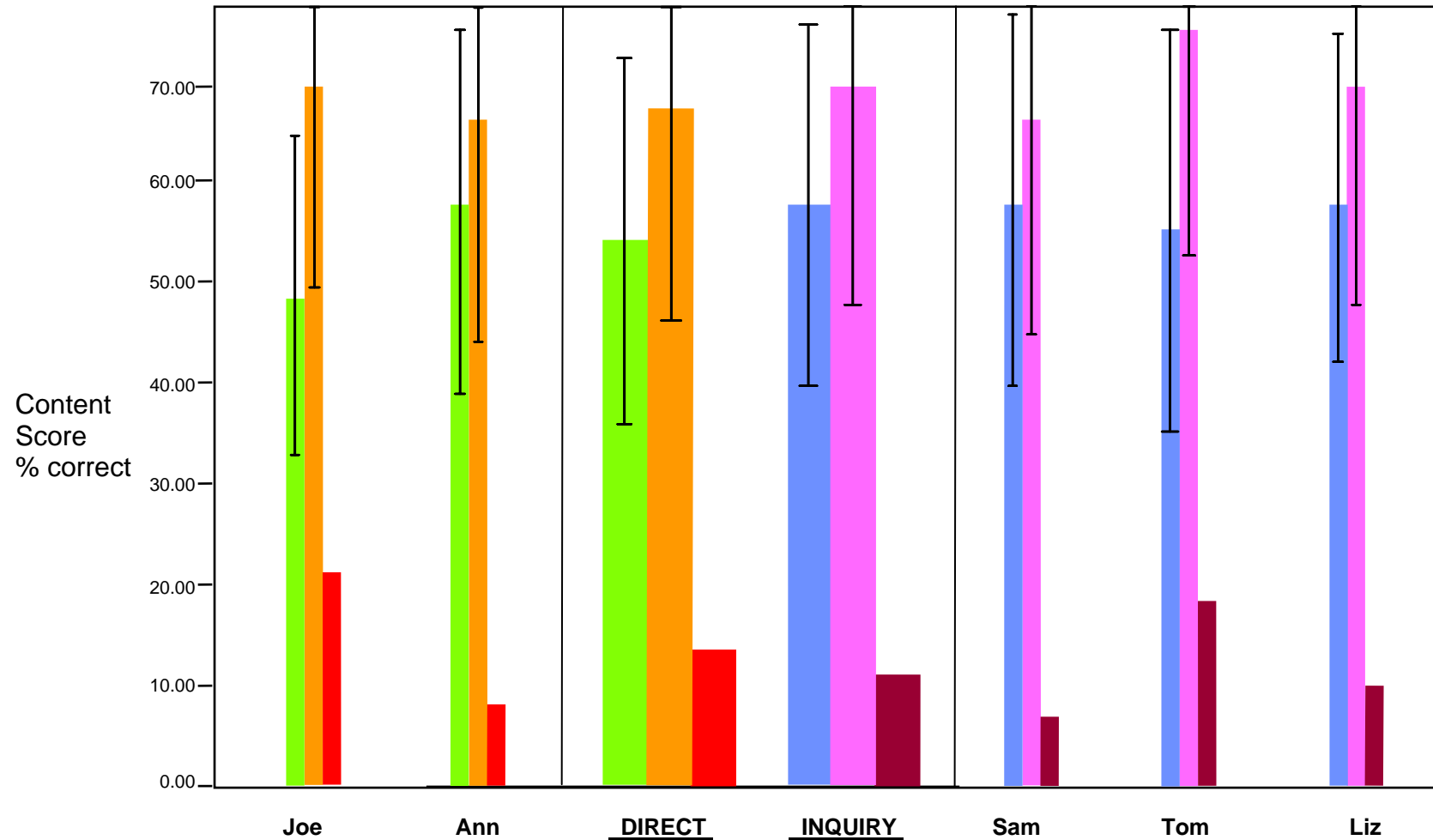
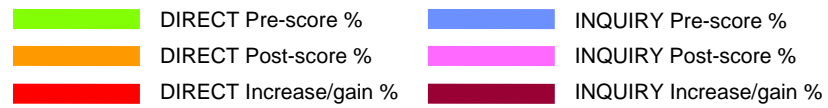
Mean (SD) ±1 SD error bar shown	(N=20)	(N=23)	(N=43)	(N=61)	(N=22)	(N=19)	(N=20)
Pre-score	50.7 (17.3)	61.3 (17.4)	56.3 (18.0)	50.7 (21.2)	49.0 (25.0)	51.7 (19.6)	51.6 (19.1)
Post-score	63.2 (18.1)	70.2 (16.6)	66.9 (17.5)	65.9 (22.6)	63.4 (25.7)	69.9 (20.0)	65.0 (22.0)
% Increase	12.5 (12.1)	8.9 (12.1)	10.6 (12.1)	15.3 (16.6)	14.5 (17.4)	18.2 (18.2)	13.4 (14.3)
Normalized gain	25.4 (26.8)	20.4 (49.6)	22.7 (40.3)	34.9 (37.7)	35.7 (40.4)	38.9 (39.6)	30.4 (34.2)

FIGURE 2. 2008 DYNAMICS UNIT.
Compare by teacher and instruction type



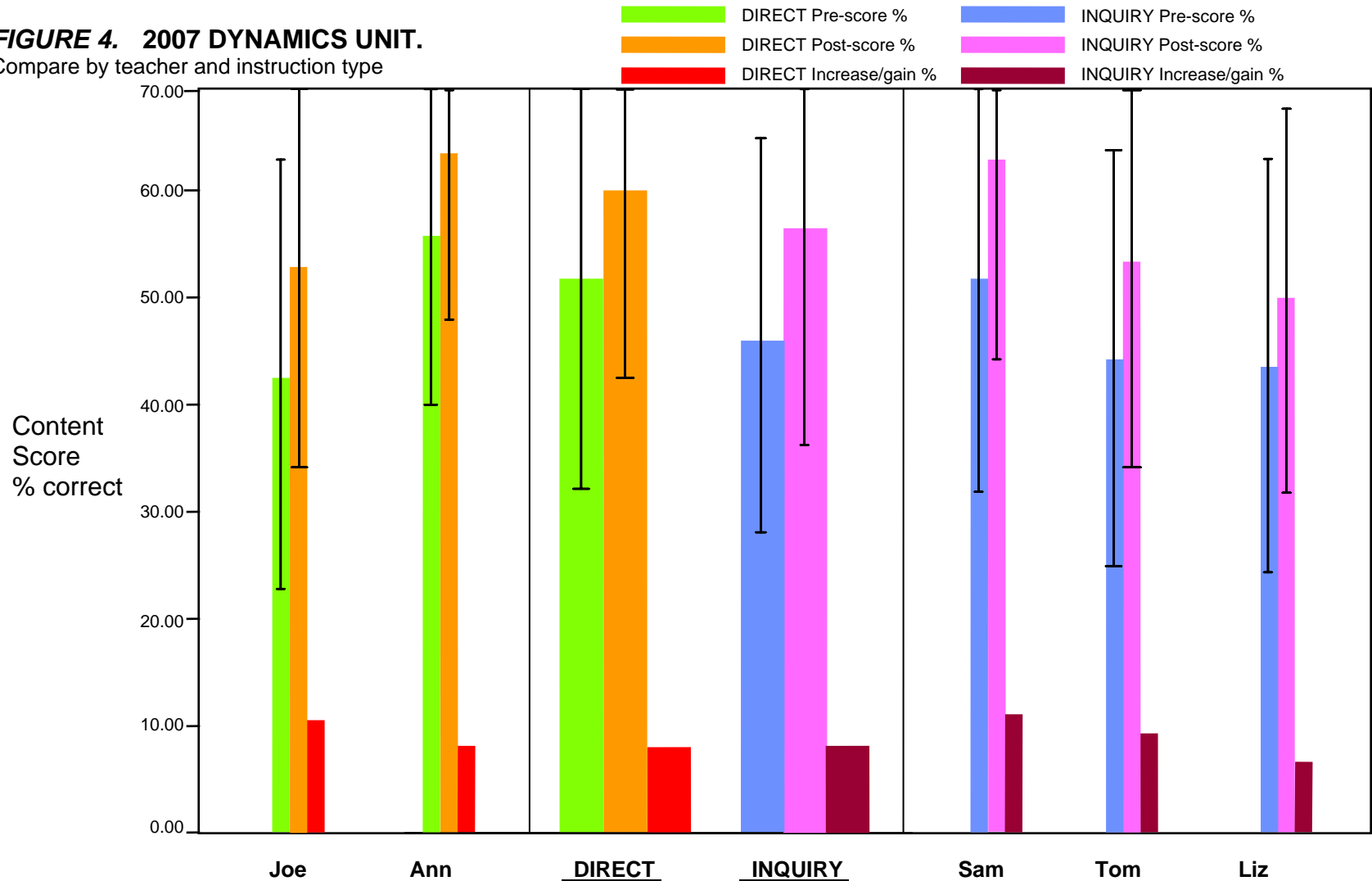
Mean (SD) ±1 SD error bar shown	(N=20)	(N=23)	(N=43)	(N=61)	(N=22)	(N=19)	(N=20)
Pre-score	51.2 (16.3)	55.3 (16.5)	53.4 (16.4)	49.3 (16.4)	50.2 (14.8)	49.9 (19.6)	47.6 (15.4)
Post-score	61.0 (17.6)	67.9 (17.5)	64.7 (17.7)	58.5 (19.8)	57.6 (25.9)	58.9 (16.3)	59.0 (15.6)
% Increase	9.8 (13.9)	12.6 (13.7)	11.3 (13.7)	9.2 (15.4)	7.4 (17.7)	9.0 (18.0)	11.4 (8.9)
Normalized gain	18.8 (30.5)	27.7 (27.2)	23.5 (28.8)	18.1 (35.4)	21.0 (43.6)	10.3 (38.5)	22.4 (19.0)

FIGURE 3. 2007 LIGHT UNIT.
Compare by teacher and instruction type



Means (SD) ±1 SD error bar shown	(N=12)	(N=18)	(N=30)	(N=47)	(N=15)	(N=16)	(N=16)
	Joe	Ann	<u>DIRECT</u>	<u>INQUIRY</u>	Sam	Tom	Liz
Pre-score	49.2 (15.7)	57.3 (18.4)	54.1 (17.6)	57.8 (17.8)	58.8 (18.1)	55.1 (19.4)	59.7 (16.7)
Post-score	69.3 (19.6)	66.9 (23.1)	67.9 (21.4)	70.4 (21.5)	66.1 (20.9)	74.1 (21.5)	70.7 (22.8)
% Increase	20.1 (14.0)	9.6 (23.8)	13.8 (20.8)	12.6 (16.3)	7.3 (14.9)	19.0 (15.7)	11.1 (16.9)
Normalized gain	43.7 (30.0)	9.2 (79.4)	23.0 (65.8)	32.7 (38.5)	22.4 (40.4)	46.9 (31.4)	28.1 (41.2)

FIGURE 4. 2007 DYNAMICS UNIT.
Compare by teacher and instruction type

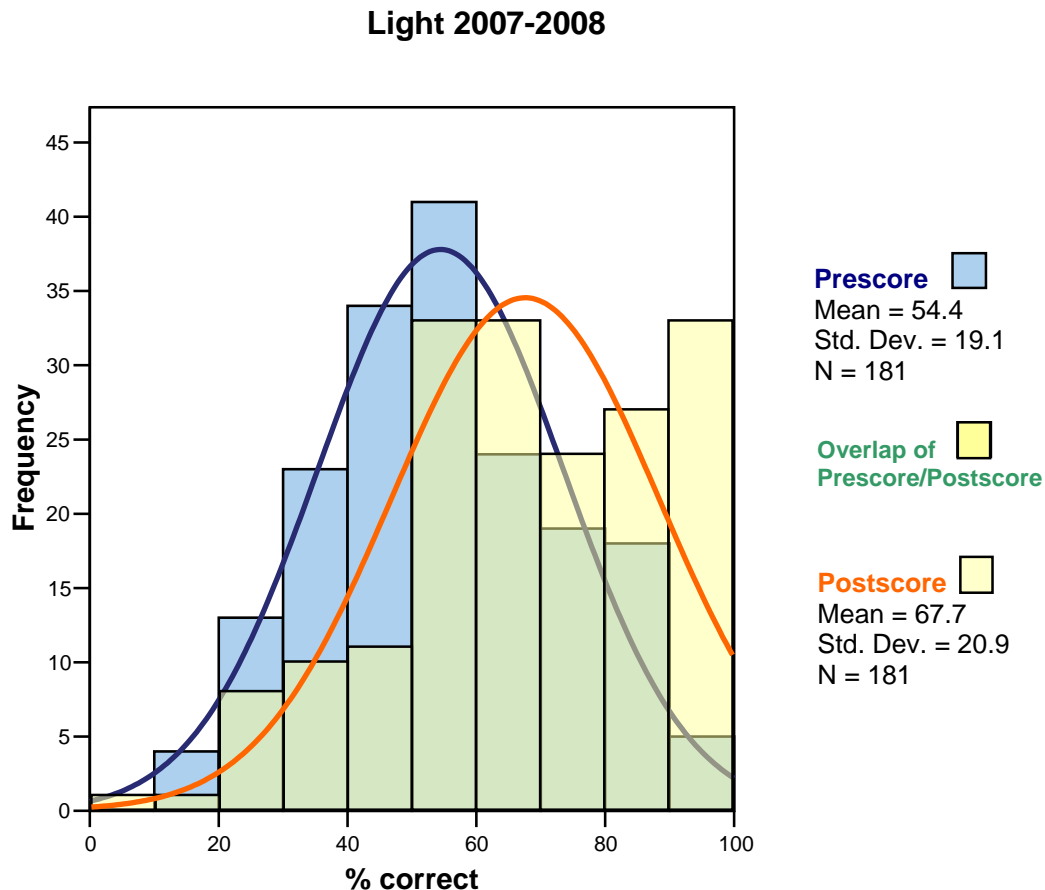


Mean (SD) ±1 SD error bar shown	(N=12)	(N=18)	(N=30)	(N=47)	(N=15)	(N=17)	(N=15)
Pre-score	42.9 (20.2)	56.3 (16.6)	51.0 (19.0)	46.9 (18.5)	52.7 (19.5)	44.5 (19.1)	43.8 (16.7)
Post-score	53.6 (19.8)	64.3 (16.4)	60.0 (18.3)	56.0 (19.0)	63.8 (18.8)	54.1 (18.7)	50.5 (18.3)
% Increase	10.7 (14.7)	7.9 (12.1)	9.0 (13.0)	9.1 (11.0)	11.1 (11.8)	9.5 (12.7)	6.7 (8.0)
Normalized gain	19.5 (29.1)	16.2 (30.3)	17.5 (29.3)	17.9 (23.0)	25.6 (21.7)	15.4 (26.4)	13.1 (19.5)

Spread in scores on pre- and post-tests

The spread in assessment scores was considerable, for both units, both pre- and post-instruction. The histogram in Figure 5 shows the score distributions for the Light unit, combining the 2007 and 2008 data. Normal curves have been overlaid corresponding to the pre and post data. Figure 6 shows the same for the Dynamics unit.

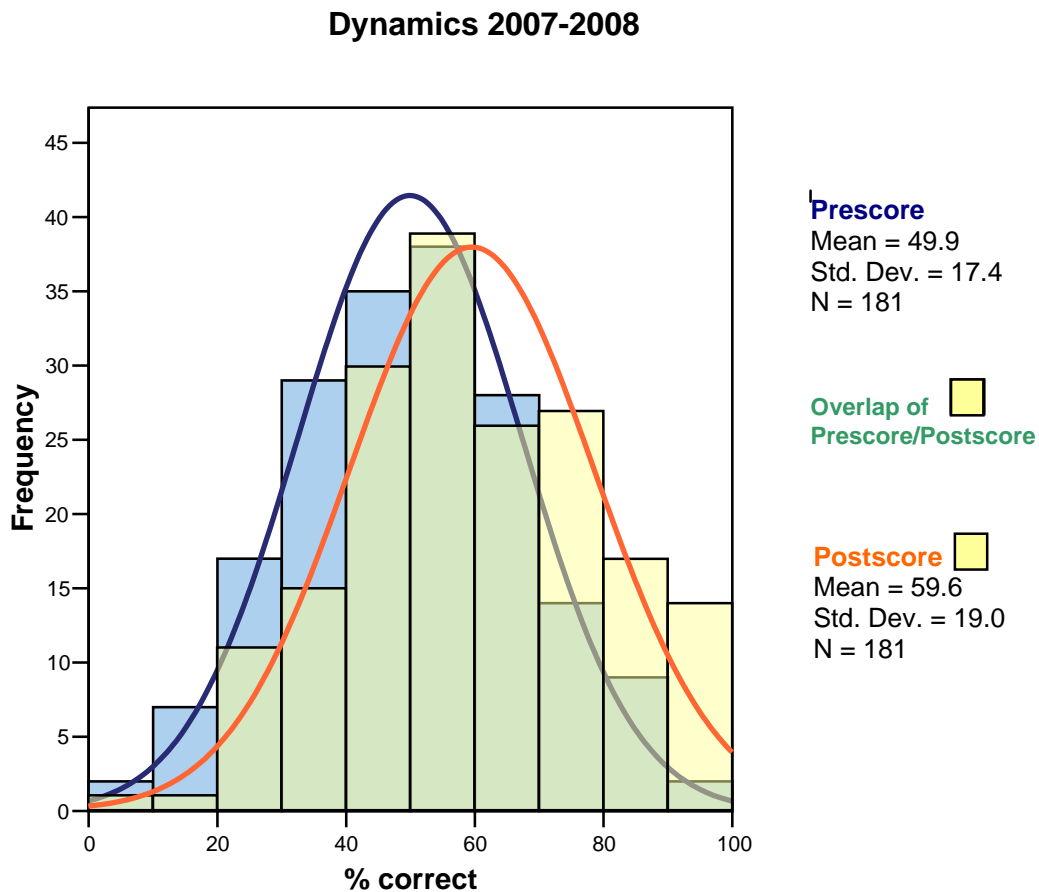
Figure 5. Frequency Distribution of raw scores (% correct) for Light Unit over two trials.



Over these two trials the difference between prescore and postscore (13.2%, SD 16.3) is statistically significant, ($t(180)=10.882, p<.001$), with an effect size (Cohen's d) of .66. A significant, negative correlation (Pearson, $r(181) = -.31, p < .001$) was found between prescore ($M = 54.4, SD = 19.1, N = 181$) and raw percentage gain ($M = 13.2, SD = 16.3, N = 181$). Since the correlation was negative (higher learning gains from lower prescores), the prescore is less likely to be indicative of student ability and more likely to be tied to initial familiarity with the topic. Raw gains were normalized to counteract apparent gain advantage that might stem from lack of prior exposure to the material. Normalized gains are individually calculated as percentage gain out of possible percentage gain. Light normalized % gain does not show significant correlation with prescores, $r(181) = .026, p = .727$. Observed mean for overall normalized gain was $M = 13.2, SD = 16.3, N = 181$.

The only statistically significant differences between teachers over two trials were between Ann and Joe on prescore ($t(71)=2.315, p=.024$), and Ann and Tom on % raw gain ($t(74)=2.334, p=.022$) and normalized % gain ($t(74)=2.228, p=.029$).

Figure 6. Frequency Distribution of raw scores (% correct) for Dynamics Unit over two trials.



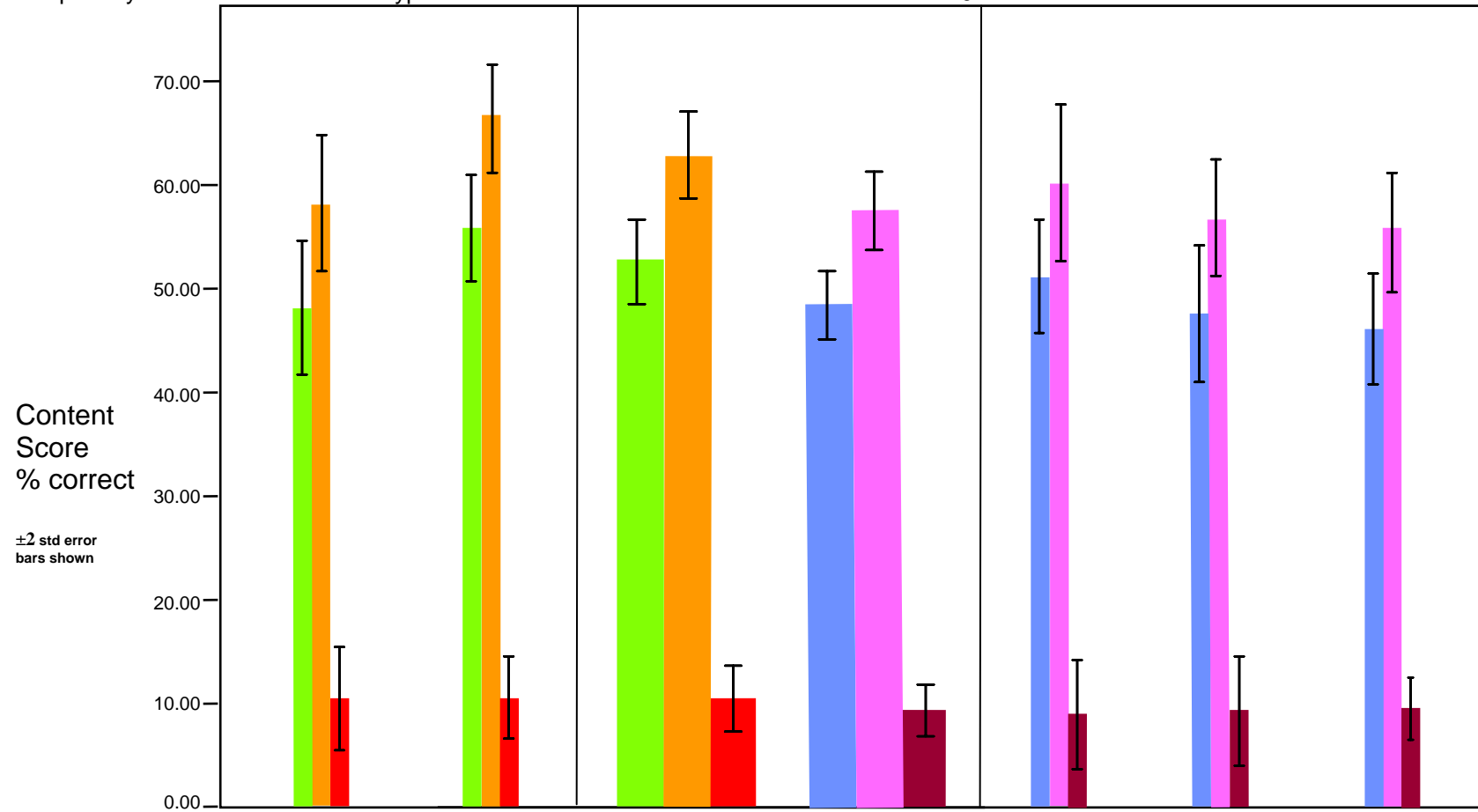
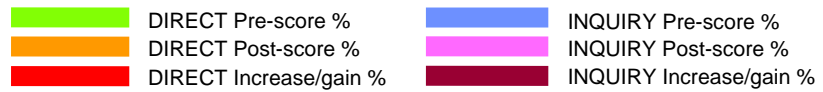
Overall mean difference between prescore and postscore (9.7%, SD 13.5) is statistically significant, ($t(180)=9.642, p<.001$), with an effect size (Cohen's d) of .53.

A significant, negative correlation (Pearson, $r(181) = -.268, p < .001$) was found between prescore ($M = 49.9, SD = 17.4, N = 181$) and raw percentage gain ($M = 9.7, SD = 13.5, N = 181$). Since the correlation was again negative (higher learning gains from lower prescores), the prescore is less likely to be indicative of student ability and more likely to be tied to initial familiarity with the topic. Raw gains were normalized to counteract apparent gain advantage that might stem from lack of prior exposure to the material. Normalized % gain does not show significant correlation with prescores, $r(180) = .008, p = .910$. Observed mean for overall normalized % gain was $M = 19.4, SD = 29.9, N = 181$.

No significant difference found between teachers on raw % gain or normalized % gain. Significant difference found between Ann and Liz on prescore ($t(74)=2.634, p=.010$), and postscore ($t(74)=2.801, p=.007$), and between Ann and Tom on prescore ($t(75)=2.068, p=.042$), and postscore ($t(75)=2.483, p=.015$).

FIGURE 7. DYNAMICS UNIT (2007-2008).

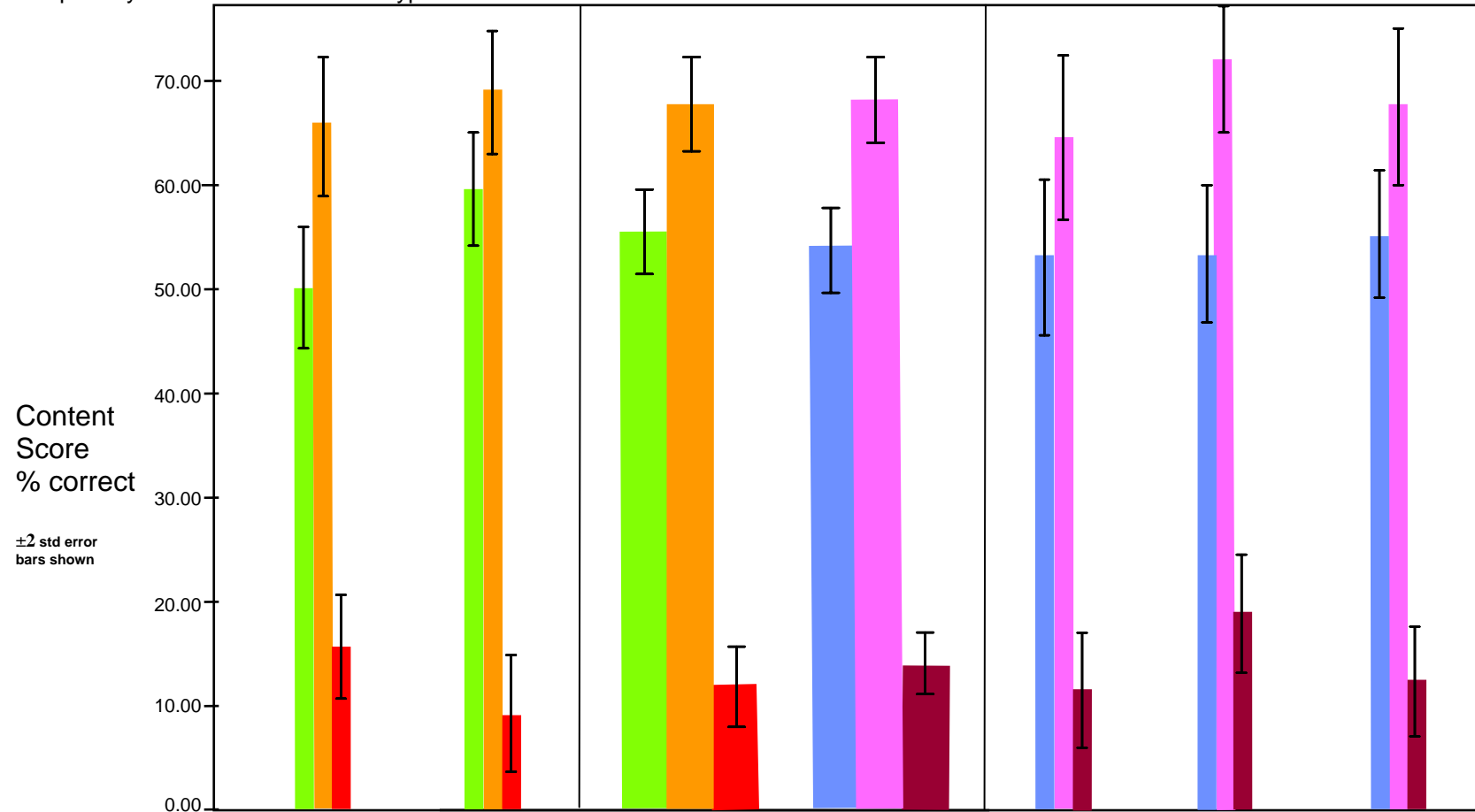
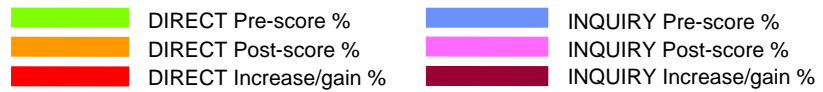
Compare by teacher and instruction type



	Joe (N=32)	Ann (N=41)	DIRECT (N=73)	INQUIRY (N=108)	Sam (N=37)	Tom (N=36)	Liz (N=35)
Means (SD)							
Pre-score	48.1 (18.0)	55.7 (16.3)	52.4 (17.4)	48.2 (17.3)	51.2 (16.7)	47.4 (19.3)	46.0 (15.8)
Post-score	58.2 (18.5)	66.3 (16.9)	62.8 (17.9)	57.4 (19.4)	60.1 (23.2)	56.6 (17.4)	55.4 (17.1)
% Increase	10.1 (13.9)	10.6 (13.1)	10.4 (13.4)	9.2 (13.6)	8.9 (15.5)	9.3 (15.5)	9.4 (8.8)
Normalized gain	19.0 (29.5)	22.6 (28.8)	21.1 (29.0)	18.0 (30.5)	22.8 (36.0)	12.7 (33.0)	18.4 (19.5)

FIGURE 8. LIGHT UNIT (2007-2008).

Compare by teacher and instruction type



	Joe (N=32)	Ann (N=41)	<u>DIRECT</u> (N=73)	<u>INQUIRY</u> (N=108)	Sam (N=37)	Tom (N=35)	Liz (N=36)
Means (SD)							
Pre-score	50.1 (16.5)	59.5 (17.7)	55.4 (17.7)	53.8 (20.0)	52.9 (22.7)	53.2 (19.3)	55.2 (18.3)
Post-score	65.5 (18.6)	68.7 (19.5)	67.6 (22.2)	67.9 (22.2)	64.5 (23.6)	71.8 (20.5)	67.6 (22.2)
% Increase	15.3 (13.2)	9.2 (17.9)	11.9 (16.2)	14.1 (16.4)	11.5 (16.6)	18.6 (16.8)	12.4 (15.3)
Normalized gain	32.3 (29.0)	15.5 (63.7)	22.9 (51.9)	34.0 (37.9)	30.3 (40.3)	42.5 (35.8)	29.4 (36.9)

Discussion

Note that the pre-test scores for the instructional topics are relatively high, with means around 50%. Furthermore the spread in scores is quite large, with standard deviation around 20%. There is a similar spread in post-test scores. In addition, for multiple choice assessment with four distracters, a pure guessing score would average 25%. Together, these aspects of the assessment represent a potential challenge for this type of research - especially when the gains obtained are not large. The gain, being the difference in pre and post means, will have a similarly large standard deviation (16%, in the case of the Light unit), and if average gain is relatively modest (say less than a standard deviation) then *differences* in such gains due to instructional mode will be even smaller. Thus mode differences found may not be statistically significant in this context, given the spread and variations due to other factors. This is what we found, for the most part.

The pre-test scores show that students clearly knew something about these particular topics when they came in. There are good 'relevance' reasons for choosing topics which are in the national science standards, in that we want to know what is likely to happen when these subjects are taught in schools, in either inquiry or direct modes. On the other hand, there is a good research case for taking the opposite approach, i.e. choosing topics which are *not* familiar to most students when they start, so that pre-scores are low with smaller spread, and post-score is likely to be directly the effect of the project instruction. For such topics the potential for obtaining larger gains with less spread should make it easier in principle to determine statistically significant gain difference between instructional modes. This notwithstanding, if one wants an indication of what will happen in the field when standards-based topics are taught, our results would provide this.

Normalized gains

It has become fairly common practice to work with normalized gain instead of raw gain as a measure of pre-post improvement. Normalized gain g is the raw gain expressed as a fraction of the maximum gain possible, as a way to take into account different pretest starting scores. Thus the defining equation for normalized gain is $g = (post - pre) / (max - pre)$

Normalized gains are thus ratios ranging between 0 and 1, with 1 being the maximum achievable. They may be reported for individual students or for class means.

We have calculated normalized gains corresponding to the various cases above, i.e. normalized gains for the two topics, two years, two instructional modes and five teachers. We also aggregated the data into direct and inquiry instructional modes. For each normalized gain value we also calculated the standard deviation and determined to what extent the normalized gain was statistically significant under the conditions of our program. We are interested in detecting normalized gain *differences*, especially between instructional modes, thus we calculated differences between cases, along with standard deviation and estimates of statistical significance.

Note that the above formula was devised for cases where gain is positive. In the less common cases where gain turns out to negative, the formula can give results which distort the actual situation, and a modification called normalized *change* has been proposed to deal with both positive and negative changes. In our study we did find cases of students whose gain was negative, hence we looked carefully at each such case, to ensure reported normalized gains represented them reasonably.

Gain Score Results

Results and data analyses from trials run in the summers of 2007 and 2008 are reported, showing both raw gain and normalized gain for the entire population of subjects/students, for each of the

two pedagogical approaches, and for each of five teachers/classrooms. Statistical analysis has been done to test whether or not there are statistically significant differences in the mean gain scores between the various groupings noted.

Descriptive statistics for each group's gain scores are shown in Table 3 and Table 4, divided by lesson topics of Dynamics and Light/Seasons, by method of instruction, and by teacher.

Table 3: 2007 Descriptive statistics – raw gain and normalized gain, by method, topic, & teacher

2007	DIRECT (30 students)			INQUIRY (47 students)			TOTAL (77 students)
	Teacher (N)	Mean SD min/max	Mean SD min/max	Teacher (N)	Mean SD min/max	Mean SD min/max	Mean SD min/max
Raw % Gain Scores	Ann (18)	7.9 12.1 -14.3/28.6	9.1 13.0 -14.3/33.3	Liz (15)	6.7 8.0 -4.8/19.1	9.1 11.0 -14.3/47.6	9.1 11.8 -14.3/47.6
				Sam (15)	11.1 11.8 0.0/47.6		
				Tom (17)	9.5 12.7 -14.3/33.3		
	Joe (12)	10.7 14.7 -9.5/33.3	17.5 29.3 -37.5/83.3	Liz (15)	13.1 19.5 -11.1/57.1	17.9 23.1 -25.0/77.8	17.8 25.5 -37.5/83.3
				Sam (15)	25.6 21.7 0.0/71.4		
				Tom (17)	15.4 26.5 -25/77.8		
Normalized % Gain Scores	Ann (18)	9.6 23.8 -63.64/40.91	13.8 20.8 -63.6/40.9	Liz (16)	11.1 16.9 -13.6/36.4	12.6 16.3 -18.2/50.0	13.0 18.1 -63.6/50.0
				Sam (15)	7.3 14.9 -18.2/27.3		
				Tom (16)	19.0 15.7 -4.6/50.0		
	Joe (12)	20.1 14.0 -9.1/40.9	23.0 65.8 -200/100	Liz (16)	28.1 41.2 -50.0/80.0	32.7 38.5 -50.0/100	28.9 50.7 -200/100
				Sam (15)	22.4 40.4 -50.0/85.7		
				Tom (16)	46.9 31.4 -7.1/100		
Raw % Gain Scores	Ann (18)	9.2 79.4 -200/100	23.0 65.8 -200/100	Liz (16)	28.1 41.2 -50.0/80.0	32.7 38.5 -50.0/100	28.9 50.7 -200/100
				Sam (15)	22.4 40.4 -50.0/85.7		
				Tom (16)	46.9 31.4 -7.1/100		
	Joe (12)	43.7 30.0 -14.3/100	23.0 65.8 -200/100	Liz (16)	28.1 41.2 -50.0/80.0	32.7 38.5 -50.0/100	28.9 50.7 -200/100
				Sam (15)	22.4 40.4 -50.0/85.7		
				Tom (16)	46.9 31.4 -7.1/100		

Table 4: 2008 Descriptive statistics – raw gain and normalized gain, by method, topic, & teacher

2008	DIRECT (43 students)			INQUIRY (61 students)			TOTAL (104 students)	
	Teacher (N)	Mean SD min/max	Mean SD min/max	Teacher (N)	Mean SD min/max	Mean SD min/max	Mean SD min/max	
Raw % Gain Scores DYNAMICS	Ann (23)	12.6 13.7 -23.8/42.9	11.3 13.7 -23.8/42.9	Liz (20)	11.4 8.9 -4.8/28.6	9.2 15.4 -23.8/42.9	10.1 14.7 -23.8/42.9	
	Joe (20)	9.8 13.9 -14.3/33.3		Sam (22) <i>(gain not sig)</i>	7.4 17.7 -23.8/42.9			
				Tom (19)	9.0 18.0 -19.1/42.9			
	Normalized % Gain Scores	Ann (23)	27.7 27.2 -41.7/71.4	24.1 28.9 -41.7/75.0	Liz (20)	22.4 19.0 -12.5/60.0	18.1 35.4 -57.1/100	20.6 32.9 -57.1/100
		Joe (20)	18.8 30.5 -37.5/75.0		Sam (22) <i>(gain not sig)</i>	21.0 43.6 -55.6/100		
					Tom (19)	10.3 38.5 -57.1/75.0		
Raw % Gain Scores LIGHT	Ann (23)	8.9 12.1 -22.7/31.8	10.6 12.1 -22.7/31.8	Liz (20)	13.4 14.3 -13.6/36.4	15.3 16.6 -22.7/59.1	13.3 15.0 -22.7/59.1	
	Joe (20)	12.5 12.1 -9.1/31.8		Sam (22)	14.5 17.4 -22.7/59.1			
				Tom (19)	18.2 18.2 -18.2/59.1			
	Normalized % Gain Scores	Ann (23)	20.4 49.6 -166.7/100	22.7 40.3 -166.7/100	Liz (20)	30.4 34.2 -27.3/100	34.9 37.7 -66.7/100	29.9 39.1 -166.7/100
		Joe (20)	25.4 26.8 -40.0/66.7		Sam (22)	35.7 40.4 -45.5/100		
					Tom (19)	38.9 39.6 -66.7/100		

Although the gain scores for both approaches were statistically significant, they were lower than what we would want to accomplish in an academic-year classroom setting. Gain scores were limited by three parameters of the study:

- Units were of high conceptual demand, with assessment to match. Assessment represented Bloom taxonomy levels 2 and 3 (comprehension and application). Larger gains are easily attainable with knowledge-based units (Bloom level 1), but this is not the goal of science education, nor is it clear how factual knowledge could be taught ‘by inquiry’. One of our pilot units, later discarded, was knowledge and process-skill based, and did lead to higher gains, but also could show no difference between instructional modes.
- A voluntary summer session provides no extrinsic motivation to do well, and thus student effort was mixed.

- Since no homework could be assigned in a summer program, out-of-class learning at a student's own pace did not occur, e.g. by doing homework problems.
- Although all of our teachers received acceptable fidelity scores, we noticed that all the teachers tended to de-emphasize or cut short certain important lesson components: application, feedback, reflection. When teachers hurried due to real or perceived time pressure, they chose to fit in all the 'action' activities, rather than take the time to make sense of them. There seems to be a natural teacher tendency to emphasize activity, to emphasize "hands-on" at the expense of "minds-on". Given that in the Direct Instruction mode students hear the concept prior to the activities, this teacher tendency is likely to impact Inquiry Instruction more than Direct Instruction.

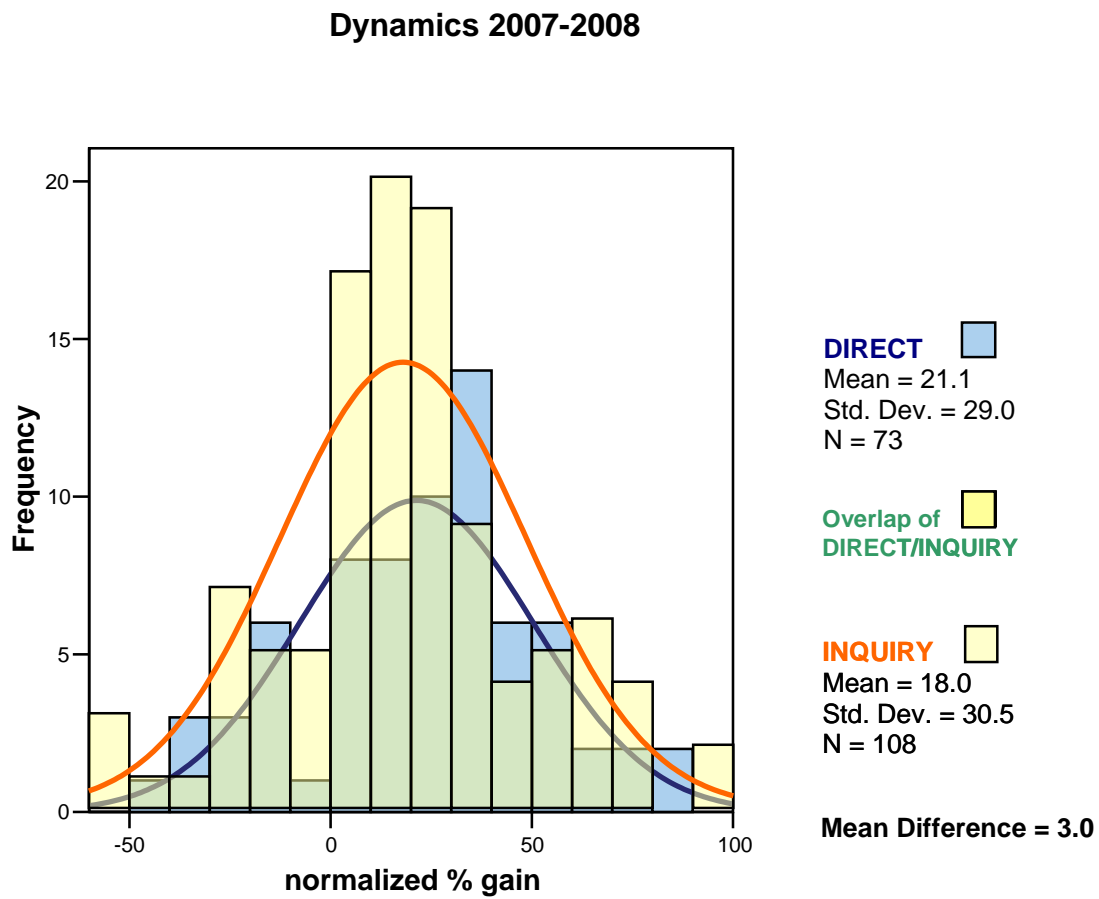
The limitations of a voluntary summer session are a fact of life that cannot be avoided. We consider the observations about teachers, however, as providing valuable insight with implications for curriculum and instruction generally, as well as for research. Hence, for our two forthcoming trials we have built into the lessons additional structures that will prompt the teachers to conclude every lesson with application, feedback, and reflection.

Discussion

One can use the sets of gain data in the tables to compare mean gains between classes/teachers, instructional modes, years and topics. While differences between such gains are evident in the tables, most differences cannot be viewed as statistically significant in this situation, given the standard deviations involved. This does not of course mean that there are no differences, just that we do not have sufficient power in this field study to detect them unambiguously. Having noted this caveat, we repeat that a mode of instruction can only be called superior to the extent that it is robust against natural classroom variation.

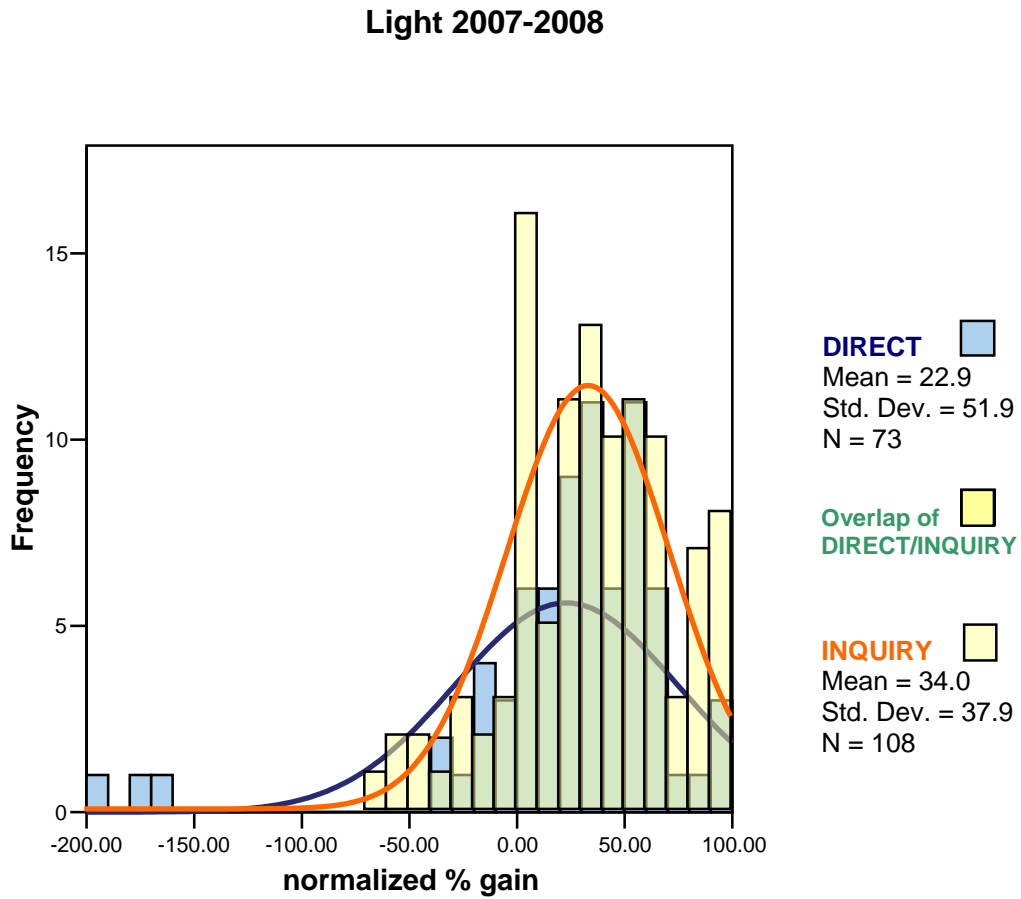
Over two trials, 85% of the raw and normalized gain scores were significant at the 0.05 level. As previously stated, prescores did not correlate significantly with normalized gains, but showed consistent (negative) correlation with raw gains. Figures 9 and 10 graphically show the Direct/Inquiry comparison for Dynamics and Light, respectively. Table 3 indicates the major areas where significant differences were and were not found.

Figure 9



No significant difference found between DIRECT and INQUIRY groups on normalized % gain ($t(179) = .665, p = .507$). (mean diff. 3.0, std. error diff. 4.5, effect size Cohen's $d = .10$)

Figure 10



Statistically significant difference not found between DIRECT and INQUIRY groups on normalized % gain ($t(179) = 1.663, p = .098$) (Mean Diff. 11.1, Std. Error Diff. 6.7, effect size Cohen's $d = .25$)

Table 5: Summary of findings

2007	Dynamics	Differences between Direct and Inquiry gains were not statistically significant.
		Differences between teacher gains were not statistically significant.
	Light	Differences between Direct and Inquiry gains were not statistically significant.
		Difference between Sam and Tom within Inquiry approach on raw gain ($t(29)=2.141, p = .041$) was statistically significant, but not on normalized gain. Difference between Sam and Joe (different approaches) on raw gain ($t(25)=2.279, p = .031$) was statistically significant, but not on normalized gain.
	Both	Differences between gains in Dynamics and Light/Seasons were not statistically significant
		Correlations between gains in Dynamics and Light/Seasons were not statistically significant
2008	Dynamics	Differences between Direct and Inquiry gains were not statistically significant
		Differences between teacher gains were not statistically significant
	Light	Differences between Direct and Inquiry gains did not reach statistical significance (raw % gains, $t(101.8)=1.672, p = .098^*$) (*equal variances not assumed, Levene's test)
		Differences between teacher gains were not statistically significant
	Both	Differences between gains in Dynamics and Light/Seasons were not statistically significant
		Correlations between gains in Dynamics and Light/Seasons were not statistically significant
2007/2008	Dynamics	Differences between Direct and Inquiry gains were not statistically significant, effect size Cohen's $d=.09$
		Differences between teacher gains were not statistically significant
	Light	Differences between Direct and Inquiry gains (2.2%, sed 2.5) were not statistically significant (raw % increase, $t(179)=.891, p = .374$), effect size Cohen's $d=.14$
		Differences in gain between teachers were statistically significant only for Ann and Tom on raw % increase ($t(74)=2.334, p = .022$), effect size Cohen's $d=.54$, and normalized % gain ($t(74)=2.228, p = .029$), effect size Cohen's $d=.52$

CONCLUSIONS

The results from our experimental study comparing specified models of inquiry and direct instruction as implemented in realistic classroom environments are that the gain differences between instructional modes were not statistically significant, as far as science conceptual understanding was concerned, within the natural variation of students, teachers and classroom situations. Note that our criterion for concept understanding is the ability to apply the concept to explain, predict or solve conceptual problems, not simply factual knowledge. Variations in gain due to other factors were at least as great as variations due to mode. This means that under our 'real' conditions, both inquiry and direct methods led to comparable student science conceptual understanding over approximately equal periods of instruction, within the limitations of our study. However note that for these same reasons the study is not able to rule out the possibility of a meaningful mode difference, if currently masked by other factors. Several factors certainly play a considerable role in student learning. Although we have tried to strike a balance between clinical control and ordinary classroom and teacher diversity, we note that different conditions and controls might be needed to pursue the question further.

In that students in both groups showed significant gain scores, the findings suggest that soundly constructed lessons, involving student engagement, and competently taught by good teachers, are as important for science concept development as whether a lesson is cast as inquiry or direct. Thus, the promotion of one mode of instruction over the other, where both are based on sound models of expert instruction, cannot be based simply on content acquisition alone. Given the composite nature of all lessons, as well as the realities of implementation in classrooms, some common claims for both direct and inquiry instruction, at least in regard to content understanding, may be seen as somewhat overstated.

Our study shows that proponents of direct instruction cannot justifiably claim that direct instruction produces superior science understanding, or does so in a significantly shorter time. Similarly proponents of inquiry cannot claim that markedly superior content knowledge will result. Claims for either method would have to be based on other grounds as well.

In regard to inquiry, most science educators feel that inquiry instruction provides 'value added' benefits in having students experience doing science for themselves and thinking much like scientists. Thus one would want to argue for example that beyond content, inquiry gives a 'feel' for science, provides appreciation of the nature of science, improves attitudes, increases transfer of knowledge to new situations, and aids long-term retention. The above-mentioned aspects, though reasonable, are of course each a new research question.

For direct instruction, it is not as clear what the further grounds might be on which to argue superiority. One is that direct is easier from the teaching point of view, particularly for inexperienced teachers. There is also some merit to the time argument, at least as far as content knowledge is concerned, but our study shows that the time aspect is not as significant as usually claimed, for specified models of good instruction either way, and the efficiency is not greater. Thus any time advantage is likely to be outweighed by missing out on the other benefits.

Finally, there is still data to be mined from our experiences with both the training trials and formal data collection trials. We will be running further trials in Kalamazoo in the next two years, in summer 2009 and 2010. For these we will switch the teachers into the opposite instructional modes, as a further test of teacher/instructional mode interactions. We will also be able to make modifications and improvements to address the educational and research issues identified and discussed in an earlier section. Further down the line, there are several aspects to

investigate in a future research project, such as the possible effects of student aptitude, science interest, time-on-task, use of application problems, improvement of teacher/student verbal discourse, and the use of new topics. Regarding the notion of ‘value added’ with inquiry instruction, we will be especially interested in investigating the growth of scientific habits of mind amongst students vis-à-vis instructional approach.

Acknowledgement

Funded by the National Science Foundation’s Interagency Education Research Initiative (IERI/NSF 04-553) Award #0437655. Any opinions, findings, conclusions or recommendations in this paper are those of the authors and do not necessarily reflect the views of the NSF.

Supplemental documents can be accessed at <http://www.wmich.edu/way2go/>.

REFERENCES

- Abd El Khalick, F., BouJaoude, S., Duschl, R., Lederman, N. G., Naaman, R. M., Hofstein, A., Niaz, M. T. D., & Tuan, H. I. (2004). Inquiry in science education: International perspectives. *Science Education*, 88(3), 397-419.
- Adams, B., Undreiu, A., & Schuster, D. G. (2007). Contrasting Inquiry and Direct Physics Instructional Designs: Examples from Dynamics. Poster presented at the winter meeting of the American Association of Physics Teachers (AAPT). Seattle, Washington.
- Adams, B., Undreiu, A., Schuster, D., & Cobern, W. W. (2008). Challenges in lesson design for testing relative effectiveness of inquiry vs. direct instruction. Paper presented at the annual meeting of the National Association for Research in Science Teaching. Baltimore, Maryland.
- Adams, B. A. J., Undreiu, A., & Cobern, W. W. (2006). The presence of inquiry in science education: What do we know? ... and what don’t we know yet? Paper presented at the annual meeting of the National Association for Research in Science Teaching. San Francisco, CA.
- Adams, G. L., & Engelmann, S. (1996). Research on Direct Instruction: 25 Years beyond Distar. Seattle: Educational Achievement Systems. 206/820-6111.
- American Educational Research Association. (2009). Definition of Scientifically Based Research [Web Page]. URL <https://www.aera.net/opportunities/?id=6790> [2009, February 28].

- American Association for the Advancement of Science (AAAS). (1990). Project 2061: Science for all Americans. Washington, DC: American Association for the Advancement of Science, Inc.
- American Federation of Teachers. (Direct Instruction [Web Page]. URL <http://people.uncw.edu/kozloffm/aftdi.html> [2003, January 6].
- Anastasiow, N. J., Sibley, S. A., Leonhardt, T. M., & Borich, G. D. (1970). A Comparison of Guided Discovery, Discovery and Didactic Teaching of Math to Kindergarten Poverty Children. American Educational Research Journal, 7(3), 493-510.
- Anderson, R. D. (2007). Inquiry as an Organizing Theme for Science. S. K. Abell, & N. G. Lederman (editors), Handbook of Research on Science Education (pp. 807-830). Mahwah, N.J.: Lawrence Erlbaum Associates, Inc.
- Atkin, J. M., & Karplus, R. (1962). Discovery or Invention? The Science Teacher, 29, 45-47.
- Ausubel, D. P. (1961). In defense of verbal learning. Educational Theory, 11(1), 15-25.
- Ausubel, D. P. (1962). In defense of verbal learning: beams in the eyes of a critic. Educational Theory, 12(4), 230-233.
- Brady, T. E. (2008). Science Education: CASSANDRA'S PROPHECY. Phi Delta Kappan, 89(8), 605-607.
- Chen, Z., & Klahr, D. (1999). All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. Child Development, 70(5).
- Committee on Prospering in the Global Economy of the 21st Century: An Agenda for American Science and Technology. (2009). Rising Above The Gathering Storm: Two Years Later. Washington, DC: National Academy of Sciences, National Academy of Engineering, Institute of Medicine.
- Craig, R. C. (1956). Directed versus independent discovery of established relations. Journal of Educational Psychology, 47(4), 223-234.
- DeBoer, G. E. (1991). A history of ideas in science education: Implications for practice. New York: Teachers College Press.
- Donovan, M. S., & Bransford, J. D. (2005). How Students Learn: Science in the Classroom. Washington, DC: Committee on How People Learn: A Targeted Report for Teachers, National Research Council, The National Academies Press.
- Education Development Center (2007). Inquiry-based Science Instruction and Students' Science Content Knowledge: A Research Synthesis. Paper presented at the annual meeting of the National Association for Research in Science Teaching New Orleans.
- Education Development Center (2008). Personal Communication.

- Erickson, F., & Gutierrez, K. (2002). Culture, Rigor, and Science in Educational Research . Educational Researcher, 31(8), 21-24.
- Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Reply to Commentators on "Scientific Culture and Educational Research". Educational Researcher, 31(8), 28-29.
- Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. Educational Researcher, 31(8), 4-14.
- Finn, C. E., & Ravitch, D. (1996). Education Reform 1995-1996. Washington, DC: Thomas B. Fordham Foundation.
- Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., DeHaan, R., Gentile, J., Lauffer, S., Stewart, J., Tilghman, S. M., & Wood, W. B. (2004). EDUCATION: Scientific Teaching. Science, 304(5670), 521-522.
- Hein, G. (2004) The Challenge of Constructivist Teaching: Adapted with permission from E. Mirochnik and D. C. Sherman (2002). Passion and Pedagogy: Relation, Creation, and Transformation in Teaching. New York: Peter Lang, pp. 197-214 [Web Page]. URL http://www.lesley.edu/faculty/ghain/papers_online/challenge_construct_2002/challenge_construct_2002.html [2009, March 8].
- Ivins, J. E. (1985). A Comparison Of The Effects Of Two Instructional Sequences Involving Science Laboratory Activities (Verification, Discovery, Information Processing). Unpublished doctoral dissertation, University of Cincinnati.
- Jenness, M., & Barley, Z. A. (1999). Observing teaching practices in K-12 classrooms: Instruments and methods (Science - version B: for multiple classroom observations to improve programming). Kalamazoo, MI: Science and Mathematics Program Improvement (SAMPI).
- Karplus, R. D. (1977). Science teaching and the development of reasoning. Journal of Research in Science Teaching, 14(2), 169-175.
- Kennedy, M. M. (1997). Defining Optimal Knowledge for Teaching Science and Mathematics. (Report No. RM10). Madison, WI: National Institute for Science Education.
- Kersh, B. Y. (1962). The motivating effect of learning by independent discovery. Journal of Educational Psychology, 53(2), 65-71.
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why Minimal Guidance during Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. Educational Psychologist, 41(1), 75-86.
- Klahr, D. (2000). Exploring science: The cognition and development of discovery processes. Cambridge, MA: MIT Press.

- Klahr, D. Paths of Learning and their consequences: Discovery Learning versus Direct Instruction in elementary school science teaching. Seminar presentation at the Pittsburgh Scientific Reasoning Supergroup 2002.
- Krajcik, J., Mamlok, R., & Hug, B. (2001). Modern Content and the Enterprise of Science: Science Education in the Twentieth Century. In L. Corno (editor), Education Across A Century: The Centennial Volume. Chicago: University of Chicago Press.
- Lamoreaux, S. K. (2001). Impressions of physics education. American Journal of Physics, 69(6), 633.
- Lawson, A. E. (2004). Preserving our Intellectual History: The History and Development of the Learning Cycle. Paper presented at the Annual Convention of the Association for the Education of Teachers in Science Nashville, TN: Association for the Education of Teachers in Science.
- Mayer, R. E. (2004). Should There Be a Three-Strikes Rule Against Pure Discovery Learning? American Psychologist, 59(1), 14-19.
- Mervis, J. (2004). EDUCATION RESEARCH: Meager Evaluations Make It Hard to Find Out What Works. Science, 304(5677), 1583.
- National Research Council. (1996). National science education standards. Washington, DC: National Academy Press.
- National Research Council. (2000). Inquiry and the National Science Education Standards: A Guide for Teaching and Learning. Washington, DC: National Academy Press.
- National Research Council. (2001). Classroom assessment and the National Science Education Standards. Washington, DC: National Academy Press.
- Novak, J. D. (1976). Understanding the learning process and effectiveness of teaching methods in the classroom, laboratory, and field. Science Education, 60(4), 493-512.
- Novak, J. D. (1977). An alternative to Piagetian psychology for science and mathematics education. Science Education, 61(4), 453-477.
- Pellegrino, J. W., & Goldman, S. R. (2002). Be Careful What You Wish For-You May Get It: Educational Research in the Spotlight. Educational Researcher, 31(8), 15-17.
- Rudolph, J. L. (2002). Scientists in the classroom: the cold war reconstruction of American science education. New York: Palgrave.
- Rutherford, F. J. (1964). The role of inquiry in science teaching. Journal of Research in Science Teaching, 2, 80-84.
- Secker, C. V. (2002). Effects of Inquiry-Based Teacher Practices on Science Excellence and Equity. Journal of Educational Research, 95(3), 151-160.

- Secker, C. V., & Lissitz, R. W. (1999). Estimating the Impact of Instructional Practices on Student Achievement in Science. Journal of Research in Science Teaching , 36(10), 1110-1126.
- Schuster, D. G. (2007). 1-4 Science, instruction and learning cycles. Chapter 1 of D. Schuster Light: Inquiry and Insights. Second Edition . Dubuque, Iowa: Kendall/Hunt Publishing Company.
- Schwab, J. J. (1962). The teaching of science as enquiry. The Inglis Lecture. Cambridge, MA: Harvard University Press.
- Schwartz, D. L., & Bransford, J. D. (1998). A Time for Telling. Cognition and Instruction, 16(4), 475-522.
- Schuster, D. G., Adams, B., & Undreiu, A. (2007). Inquiry and Direct Instructional Approaches to Conceptual Dynamics. MIAAPT Grand Rapids, Michigan: MIAAPT.
- Shulman, L. S., & Keislar, E. R. (1966). Learning by discovery: a critical appraisal. Chicago: Rand McNally.
- Shymansky, J. A., Hedges, L. V., & Woodworth, G. (1990). A Reassessment of the Effects of Inquiry-Based Science Curricula of the 60's on Student Performance. Journal of Research in Science Teaching, 27(2), 127-144.
- Shymansky, J. A., Kyle, W. C. J., & Alport, J. (1983). The effects of new science curricula on student performance. Journal of Research in Science Teaching, 20(5), 387-404.
- Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Undergraduates in Science, Mathematics, Engineering, and Technology: A Meta-Analysis. Review of Educational Research, 69(1), 21-51.
- St. Pierre, E. A. (2002). "Science" Rejects Postmodernism. Educational Researcher, 31(8), 25-27.
- Sweller, J. (2009). What human cognitive architecture tells us about constructivism. S. Tobias, & T. M. Duffy (editors), Constructivist Instruction: Success or Failure? (pp. 127-143). Routledge.
- Sweller, J., Kirschner, P. A., & Clark, R. E. (2007). Why Minimally Guided Teaching Techniques Do Not Work: A Reply to Commentaries. Educational Psychologist, 42(2), 115-121.
- Tai, R. H., & Sadler, P. M. (2009). Same Science for All? Interactive association of structure in learning activities and academic attainment background on college science performance in the USA. International Journal of Science Education, 31(5), 675-696.
- Taylor, J. A., Scotter, P. V., & Coulson, D. (2007). Bridging Research on Learning and Student Achievement: The Role of Instructional Materials. The Science Educator, 16(2), 44-50.

- Tretter, T. R., & Jones, M. G. (2003). Relationships between inquiry-based teaching and physical science standardized test scores. School Science and Mathematics, 103(7), 345-350.
- Udovic, D., Morris, D., Dickman, A., Postlethwait, J., & Wetherwax, P. (2002). Workshop biology: Demonstrating the effectiveness of active learning in an introductory biology course. Bioscience, 52(3), 272-281.
- Walberg, H. J. (1991). Improving school science in advanced and developing countries. Review of Educational Research, 61(1), 25-69.
- Welch, W. W. (1976). Evaluating the impact of national curriculum projects. Science Education, 60(4), 475-483.
- Welch, W. W., & Walberg, H. J. (1972). A national experiment in curriculum evaluation. Curriculum Evaluation, 9(3), 373-383.
- White, B. Y., & Frederickson, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. Cognition and Instruction, 16(1), 3-118.
- Wittrock, M. C. (1964). The learning by discovery hypothesis. ERIC Document # ED003340.
- Wright, C. J., & Nuthall, G. (1970). Relationships between Teacher Behaviors and Pupil Achievement in Three Experimental Elementary Science Lessons. American Educational Research Journal, 7(4), 477-491