EDUCATIONAL PSYCHOLOGIST, *39*(4), 203–212

Methodological Alignment in Design-Based Research

Christopher M. Hoadley
College of Education, and School of Information Sciences & Technology
Penn State University

Empirical research is all about trying to model and predict the world. In this article, I discuss how design-based research methods can help do this effectively. In particular, design-based research methods can help with the problem of methodological alignment: ensuring that the research methods we use actually test what we think they are testing. I argue that our current notions of rigor overemphasize certain types of rigor at the expense of others and that design-based research provides an opportunity to select different inferential trade-offs. I describe how 1 design-based research trajectory evolved over time in a way that helped ensure that the learning theories being studied were well represented by the planned interventions and that the interpretation of outcomes was grounded in an understanding of not only the research design, but how the research played out in practice when enacted in real classrooms.

The word *rigor* may be the most used and most effective weapon in scientific controversies. We all labor as researchers under the assumption that the question of whose interpretation of the world is right is independent of agendas, personalities, or politics; rather, we put our faith in listening to data and taking an empirical stance. But agreeing to be empirical does not always resolve the conflict of ideas. Sociologists of science point out that all science, as a human endeavor, is filtered through our politics, our biases, our worldview, and so on (Kuhn, 1962; Latour, 1987). We persist in working toward a rigorous science as a laudable, if unattainable, goal. Data itself may be unassailable, but the ways the data are collected and interpreted most certainly are. This is the heart of rigor.

RESEARCH, RIGOR, AND EXPERIMENTS

Experimentation is one attempt to proceduralize rigor. One type of rigor is avoiding bias; this is accomplished by producing experiments we believe to be replicable and describing our methods in detail. We also use random assignment and experimental control to avoid misinterpreting confounds or covariates. Of course, we recognize that there is stochasticity in the system; for this reason, we use sophisticated statistical techniques to help predict how likely it is that the results we see are due to chance. These statistical inference techniques presume things like normally distributed results (presumably due to measurement error, but possibly due to inherent indeterminacy in the system). All of these efforts are a means to increase the rigor with which we make inferences using an experimental paradigm.

This paradigm is not without drawbacks. Experimental control is difficult in the complexity of a real classroom, and we know that we cannot control for learners' prior experiences, a demonstrably significant covariate. Nor can we ensure that a treatment is identical across situations. In medicine, our understandings of chemistry suggest that any drug with the same chemical structure is in all ways equivalent to other drugs with the same chemical structure. In this context, it makes sense to attempt double-blind studies to reduce the role of interpretation by a researcher who may unconsciously affect the results. In educational research, the notion of a double-blind study has limited use; how do we prevent teachers (the treatment administrators) from knowing what treatment they are administering? What is the equivalent of chemical structure that will let us know when one pedagogical treatment is in all important ways equivalent to another?

This special issue is concerned with another paradigm—that of design-based research (Brown, 1992; Collins, 1992; Design-Based Research Collective, 2003; Hoadley, 2002; Kelly, 2003). Design-based research is in many ways a complement to experimentation. For simplicity's sake, I treat design-based and experimental research as distinct methods, although in reality research may be seen to fall on a continuum between strictly experimental paradigms and design-based paradigms or as an interleaving of these paradigms (cf. Brown, 1992).

Requests for reprints should be sent to Christopher M. Hoadley, College of Education, and School of Information Sciences & Technology, Penn State University, 314D Keller Bldg., University Park, PA 16802. E-mail: edpsychol04@tophe.net

Design-based researchers treat as fundamental the problem of context. Much as cultural anthropology cannot be conducted experimentally, when we do design-based research, we acknowledge the difficulty in educational research of ensuring control and assuming universality. Instead, design-based research views outcomes as the culmination of the interaction between designed interventions, human psychology, personal histories or experiences, and local contexts. All four impact the outcomes (which include the enacted, as opposed to designed, interventions). Hence, design-based researchers recognize the difficulty of experimental control, as dozens (if not millions) of factors interact to produce the measurable outcomes related to learning. Perhaps the most important commitment of design-based researchers is in understanding that treatments may not go as planned. In a very important sense, in design-based research, the enacted intervention is a dependent, not an independent, variable. " … the intervention is the outcome … in an important sense" (Design-Based Research Collective, 2003, p. 5).

Design-based research, therefore, proceeds in a very different manner than experimental research. For one, the research program often involves a tight relationship between researchers and teachers or implementers, blurring the "objective" researcher–participant distinction. A second distinction is the use of tentative generalization; results are shared without the expectation that universality will hold. Third, although planned comparisons do occur, the design-based researcher frequently follows new revelations where they lead, tweaking both the intervention and the measurement as the research progresses. Fourth and finally, the design-based researcher, to treat enacted interventions as an outcome, often documents what has been designed, the rationale for this design, and the changing understanding over time of both implementers and researchers of how a particular enactment embodies or does not embody the hypothesis that is to be tested. In short, the treatment's fidelity to theory is initially, and sometimes continually, suspect. This leads to a broad documentation of the intervention (to catch all relevant, but unanticipated, consequences of the design on the enactment).

This new notion of how to do research raises many questions related to rigor. How can we ensure that we have adequately characterized an intervention we did not entirely control? How can we document interactions and outcomes when these outcomes are not known in advance? Will the results obtained in one context generalize to another, and under what conditions? How can we characterize the second- and third-order effects of a designed intervention as it is enacted in a particular context and the implications for not only the learning but the context itself? Will others be able to implement similar interventions in their own local contexts in ways that are similar enough to include the benefits of any "successful" intervention, allowing enough adaptation to a local context to permit the benefits while not allowing the "lethal mutations" (Brown, 1992) that might prevent the intervention from being similar in theory and effect to the proposed intervention?

Although the notion of how to ensure certain kinds of rigor in the experimental paradigm are relatively well-specified, the notion of rigor is being developed in design-based research. Some of the tenets of experimental research are violated (such as changing treatment protocols midimplementation). This might lead the casual observer to believe that design-based research is less rigorous than experimental research. However, through the following example, I propose that design-based research is more rigorous in certain ways. In particular, design-based research is strong at helping connect interventions to outcomes through mechanisms and can lead to better alignment between theory, treatments, and measurement than experimental research in complex realistic settings like the classroom.

ALIGNMENT

The notion of alignment is essential to our understanding of research validity. Usually, when people discuss validity, they are referring to *measurement validity*, or the ability to ensure that our measurements accurately reflect the constructs that we are trying to measure. However, validity has a larger sense: The *validity* of a study is the likelihood that our interpretation of the results accurately reflects the truth of the theory and hypotheses under examination. In this sense, we need to be concerned with two other kinds of validity in research. We need to ensure that we have *treatment validity*—that is, that the treatments we create accurately align with the theories they are representing. Finally, we need *systemic validity*—that is, the whole research endeavor must not only create a fair test of the theories, but those theories must be communicated in a way that is true to the inferences used to prove them. These theories must then be applicable to decisions based on the research (*consequential validity.*) For instance, a field-wide presupposition of $p < .05$, along with widespread understanding that these p values are an indication of likelihood of chance occurrences (and not a guarantee of truth), yields a certain kind of systemic validity.

In experimental research, the presupposition is usually that as long as good experimental hygiene is performed, aligning theories to treatments (being able to say that a treatment represents such-and-such a situation in the theory) is relatively unproblematic. Likewise, for well-specified theories, the "if … then" or predictions implied by the theories should be relatively unproblematic. Finally, although measurement is often treated as problematic in experimentation, the notion of consequential validity (or how the interpreted results of the experiment will be applied in practice to future prediction and implementation) is not. Debates around construct validity of this measure or that ("Is this really measuring what we think it does?") often hinge on tacit differences in the understanding of the concomitant consequential valid-

ity ("Do we have the same understanding of what the construct is in terms of what we do about it?"). Systemic validity is what we are really after: Does the research and the inferences drawn from it inform the questions that motivated the research in the first place?

To achieve true systemic validity as educational researchers, our studies must inform our theories, which must inform practice. Educational research has been found wanting by many in this sense of systemic validity. In particular, Robinson (1998) highlighted that some of the attempts to maintain objectivity by distancing theory, research, and practice in fact yield disconnects between research and practice. Lagemann (2002) termed this the need for *usable knowledge* and discussed how, historically, the debate between Deweyan and the more behaviorist paradigm advocated by Thorndike led to the loss of Dewey's original model of educational research as tightly linked to educational practice (Lagemann, 2000).

In design-based research, the process of forcing the same people to engage the theory, the implementation of interventions, and the measurement of outcomes encourages a greater degree of methodological alignment. Design-based research is, at its heart, an attempt to combine the intentional design of learning environments with the empirical exploration of our understanding of those environments and how they interact with individuals. On the one hand, this appears to diminish replicability by increasing greatly the responsibility of the individual researcher to document what happened in unbiased terms and not to selectively attend to data that confirm prior understandings. On the other hand, forcing individuals to carry ideas all the way from explanation to prediction to falsification to application seems like the missing link in educational research that will ensure our theories have practical implications. Indeed, we may have been deceiving ourselves all along, in that we never really had a handle on whether our treatments really represented the theory-interpreted conditions they were standing in for. In situations where the relevant variables for learning are multitudinous (thousands of contextual, individual, and group factors; myriad teacher decisions made on the fly) and hard to control out, being intimate with the research setting and linking on an extremely fine scale, the designed and enacted intervention may be our best hope for relevance.

This relevance comes at a cost. First of all, design-based research is based on the idea that universality is rare in educational phenomena, and because the method takes tentative steps by first examining individual contexts, design-based researchers generalize their findings only tentatively, making this a *local science* (diSessa, 1991). Second, because the researchers are participant–observers who intervene deliberately in the settings they study, it is incumbent on the researcher to describe and monitor ways that their own agenda is responsible for the results. A researcher may produce a successful outcome due to a wonderful theory or an effective treatment or through unintended aspects of her or his own participation in the situation. Design-based researchers must not only document their perspective or starting point, but must also document any plausibly relevant interventional strategies used not only by participants observed, but also by the researcher herself or himself. On the negative side, it seems a little unusual to demand such self-reflection, straying close to introspective methodology. On the positive side, one can view this as the need to document design strategies, and who better to do this than the designers? By documenting what it is like to try to make learning happen from the point of view of those who would foster learning, we may be edging toward a more usable, and hence more valid, form of research.

It is also important to realize that no method should be allowed to stand in isolation. Much as a quantitative experimentalist might turn to a qualitative, ethnographic study or a simulation to help make sense of her or his own results, the design-based researcher can interleave methods as long as the systemic validity of the activity holds. When a design-based researcher is uncertain if their outcomes are simply the by-product of an (unknown) aspect of their own involvement, they can turn to an experimental paradigm to help ascertain what is causing what (while drawing on the rich contextual knowledge formed by their engagement in the setting). In contrast with others' descriptions of the method (Brown, 1992; National Research Council, 2002), design-based research need not be seen as "prescientific" or merely as hypothesis generation. In areas where controlled experimentation may be used to adequately test a hypothesis, the experimental paradigm is a powerful means for inferring causal relations. But if, as is the case in many educational research endeavors, the assumptions of this method are violated (such as universality, control, or treatment replicability), experimental results may be at best difficult to interpret (e.g., each study may seem to generate conflicting results due to uncontrollable covariates) or at worst meaningless or misleading (such as when a "proven" intervention proves useless in practice because what the intervention means in varying contexts proves more opaque than expected).

In the sections that follow, I describe a research trajectory that illustrates how doing design-based research can help the researcher align not only measurements, but theories, treatments, and interpretations in a manner than lends itself to usable knowledge. This article is about some designs of technologies and activities that fostered collaborative aspects of learning, predominantly in the Knowledge Integration Environment (KIE) research project (Bell, Davis, & Linn, 1995; Hoadley & Bell, 1996), which developed software for Internet-based middle school science education. I participated as a designer, a researcher, and a teacher in the project.

Designing a Technology-Enhanced Environment for Collaborative Science Learning

Research on collaboration adds design complexity; it is particularly sensitive to variations in context, and any intervention reverberates through the setting changing both the individuals

and the social context. Time is required to see how the intervention settles into a more stable state as both individuals' practices and the group practices adapt to the new tools and possibly reach equilibrium. Here, I give a description of work that provided rich contexts for studying how technology could scaffold learning with a pair of tools in a variety of contexts: university researchers, engineering undergraduate and graduate courses, and middle school science classes.

Researching the Multimedia Forum Kiosk and the SpeakEasy Discussion Tool

Our initial design problem was straightforward: allow productive discussion around multimedia by people who were not in a single location at the same time. Like many design problems, this one capitalized on the potential of technology to make possible what had previously been impossible. We designed our initial prototype in HyperCard and dubbed it the Multimedia Forum Kiosk, or MFK. We examined prior interfaces such as Internet newsgroups (at that time, primarily an academic communication medium) and e-mail mailing lists. Another important example we considered was Scardamalia and Bereiter's tool, Computer Supported Intentional Learning Environment (CSILE; Scardamalia & Bereiter, 1992, 1994; Scardamalia, Bereiter, McLean, Swallow, & Woodruff, 1989). Theoretically, we were aligned with their theories of collaborative knowledge building, but we wanted to incorporate a more general discursive model than CSILE's (which was primarily science focused) to foster a sense of community or awareness of others in the dialogue (CSILE did not directly support social awareness) and to integrate video into discussions (Hoadley & Hsi, 1993). In this sense, we were already working on treatment validity toward the idea of "computer supported collaborative learning," which CSILE had previously instantiated.

Our tool had many now-common features (including a top-level organization by topic and threaded discussion) and several features that made it unique (Hoadley, Hsi, & Berman, 1995). First, it provided two collaboration spaces—one, the *opinion area*, allowed one comment per person on the topic that could be revised over time, whereas the second, the *discussion area*, allowed threaded discussion, but did not allow revision of prior comments, only response. Second, the tool made use of *semantic labels*, or labels from a fixed set of choices (we borrowed this idea from Scardamalia and Bereiter, but used categories from a more general model of small-group discussion; see Bales, 1969). Third, we made extensive use of social cues throughout the interface based on a theory of social representations. All comments were represented by face icons, and all topics were introduced by a topic author. This tool underwent at least three major redesigns, with at least two incarnations as the MFK and at least two incarnations as the Web-based tool SpeakEasy.

SpeakEasy was one of the first two Web-based threaded discussion tools (along with HyperNews) that are so familiar to Internet users today, predating the introduction of the first Netscape browser. In our final study, our implementation of SpeakEasy discussion doubled the prevalence of correct science conceptions in our student population and significantly improved partially correct conceptions (Hoadley, 1998; 1999a; Hoadley & Linn, 2000).

Our design process was principled and relied on a specific, tentative model of how collaboration would foster knowledge building. We recognized that poorly implemented collaboration could hinder learning as much as help (Linn & Burbules, 1993). Our model of productive discussion (Hsi & Hoadley, 1997, after Pea, 1992) dovetailed with the knowledge integration approach taken elsewhere in our research program. We faced two challenges: first, to ensure participation in discussion; second, to ensure the discussion was productive—meaning that it demonstrated the features hypothesized to be necessary (and possibly sufficient) for learning via discussion. Briefly, these features are *inclusiveness* and *participation* (all members of the discussion are able to participate), the *externalization* of a repertoire of understandings or models of the domain (often different initial viewpoints), *differentiation processes* (where old models lead to new variants), *linking* (consideration of which models are coherent or incoherent), and *selection* (privileging or selecting the models that have the most explanatory power and coherence). In addition, as a component of a larger set of interventions—initially, the Computer as Learning Partner microcomputer-based laboratories (Linn & Hsi, 2000) and later the KIE suite of tools and activities—we had a responsibility to contribute to the overall goals of the project. We explicitly tried to help students develop their scientific epistemology through a coherent curriculum that included real-world experiences, laboratory experiences and inquiry, and critical examination of information resources from the Internet. Eventually, we succeeded in all these goals, although it took two dissertations to develop and implement a workable set of tools and activities, ensure that our tools were actually fostering productive discussion (Hsi, 1997), and demonstrate how this productive discussion leads to individual learning (Hoadley, 1999a).

Usability versus context of use. Naively, we assumed that usability would be the primary indicator of success in our design. After creating the initial prototype, we tested the tool with participants from an education research department using think-aloud analyses, time-usage analyses, and interviews. Our initial analysis did in fact demonstrate that the tool was usable—our test participants were given no instruction and still managed to uncover and use every feature of the system, from reading and navigating comments to contributing their own comments in both the opinion area (nonthreaded) and discussion area (threaded). In one case, the think-aloud protocol provided direct evidence that suggested our semantic types prompted reflective thinking and prevented a "flame." By usability metrics, our system was a

success already; people quickly figured out what it was for and how to use it, even people who had not used Internet newsgroups or, indeed, any online discussion tools other than e-mail (Hsi, Hoadley, & Schwarz, 1992). This was one of the first instances in which we had poor systemic validity, however, as we confused usability with likelihood of adoption. Later, we saw that usability does not always lead to use.

Designing functional activities and implementing a context for use. Initially, we took our tool into a research department lounge and engineering classrooms on several college campuses, both graduate and undergraduate. At this time, we also started installing the tool elsewhere: a self-paced study center for undergraduates, a museum, and the lobby of a college building. Partially through discussions with users, partially through comments students left in the system, and partially by quasi-experimentally comparing participation in the different settings, we realized that there were important contextual preconditions for use (Hsi & Hoadley, 1994). The public installations turned out to be too idiosyncratic for us to understand what made some people use them and other people not, but the classroom experiences started giving us some consistent messages. First, we realized that students' use of the tool was directly related to their ability to access the kiosk running the software (remember, this was prior to widespread use of even the Mosaic browser), the degree to which the topics were perceived as relevant and interesting, and the degree to which the tool was integrated with their course (Hsi & Hoadley, 1994). These findings seem obvious in hindsight, but addressing them is easier said than done and involved significant exploration in our contexts. For instance, we thought of classrooms and public spaces as easy to access, but they were not because of the social discomfort caused by working on the kiosk in these spaces. Instead, laboratories provided a much more approachable venue, because students were used to being collocated with other students working on independent activities. Regarding integration with the course, we saw different instantiations of integration that supported the tool via the course and vice versa. In some cases, students felt they were better able to solve homework problems if they read and participated in the online discussion because the topics closely paralleled the technical content in class, and in other cases, students participated because the instructor summarized comments in class and reacted to them, indicating a strong interest on the part of the professor. In many cases, anonymity played a big role in the participation, as students had few if any ways to communicate anonymously with their instructors besides our system. In some other cases, the asynchronous nature of the communication medium proved important; for instance, students with limited proficiency in English were able to participate in the discourse by taking extra time to read comments and prepare responses in English. The integration with the course also took some interesting twists. Although some instructors actually provided participation grades for contributing comments to the system, we had nearly equivalent participation when an instructor read, summarized, and responded to student comments in class (this was a large course with nearly 100 students, and other opportunities to influence instruction were rare). The kiss of death, however, was superficial integration with the course—even if students were introduced to the system in class, if the instructor never mentioned the system again and did not give grades on it, most students would opt not to participate. The few who did participate in these circumstances, of interest, were often women or minorities. Without the in-class discussions and one-on-one interactions the kiosk provoked, the kiosk itself would have been a different intervention. Identifying the nature and scope of the intervention when the cultural changes provoked by our tools and activities were coconstructed simultaneously with use of the tools and activities made traditional before-and-after testing less meaningful. This coevolution of phenomena proved to pose a methodological challenge that would crop up repeatedly, one that is probably intrinsic to the problem of studying collaboration (Barab, Hay, & Yamagata-Lynch, 2001; Roth, 2001). In this way, we began to explore what our "treatment" really consisted of. What we thought was a tool-as-intervention began to become, for us, a tool plus activities in a favorable context as our intervention.

Enactment as a joint product of context and designed intervention. Later in the development of the system, we began experimenting with our discussion tools in the Computer as Learning Partner middle school science classroom with 12- to 13-year-old students (Linn & Hsi, 2000). Initially, this experimentation began with the MFK technology and science-oriented topics (Hoadley, 1999a; Hsi, 1997). There were important interactions between our tool and the culture of the classroom, interactions that evolved as tools influenced use and use influenced culture. Some elements of the local culture already supported use. For instance, students in this classroom (which had a 2 to 1 ratio of students to computers) were familiar with computers, and each student had some prior experience working on a computer. Likewise, the teacher had previously started a tradition of coming in to work on labs or computer work during lunch and immediately before and after school; the system benefitted from these practices. Other aspects of the culture evolved in ways that we would not have predicted. For instance, the fact that the system was based on a sole kiosk (we actually had two computers in a single kiosk, but each student had an account on only one of the two machines) led to some interesting cultural outcomes. Initially, the single kiosk enhanced interest and face-to-face collaboration—students would gather around the kiosk and read over each others' shoulders as comments were made. The relative rarity of the kiosk machines made them more attractive, and soon "kiosk groupies" would frequently visit the machine as a social group outside of class time. Unfortunately, the emergence of these groupies

began to erode access to the discussion for other students; the stronger the social bond between the groupies became, the harder it was for those not in the clique to access the machine. The teacher, who was aware of the problem, began to try different ways to ensure access, including a signup sheet for time on the kiosk and strategic shooing when clumps of people began to form around the machine. The teacher did not dissuade all groups from clustering around the machines, but rather based his actions on who else was in the room and whether they were likely to be encouraged or dissuaded by the current group near the kiosk (Hsi, 1997). This type of very nuanced design activity was only possible because the teacher was aware of activity around the machine (in part with the help of the researchers), the goals of the research, and the intervention and had a number of techniques to try to encourage equitable access. It is likely that, in other circumstances, different social issues would have arisen and required different interventions to allow all students to participate in the online discussion. Eventually, by moving to a Web-based system, we eliminated the problem of a single point of access, but we raised other issues about who could access the Internet where.

Another aspect of our intervention coevolving with culture happened later, as the culture of technology changed outside the school. When we switched to the SpeakEasy tool from the MFK software (mid-semester), our students brought their prior practices easily to the networked version of the tool, and student participation rates escalated slightly but insignificantly. We found no differences in student comment length or quality. At that time in the early 1990s, few students (less than 10%) had any experience with the Internet at all. In an after-school session lasting about an hour, we gave some students an introduction to the Web that included instruction on what hyperlinks looked like, how to click on them, and how to use the "Back" and "Forward" buttons to retrace their prior steps. The rest of the students got an abbreviated version of this tutorial and were encouraged to seek help from peer guides.

When students began to use the online discussion tool, they often perceived it to be a completely different social setting, with different expectations than their familiar face-to-face counterparts such as in-class time or on the playground. Hsi (1997) documented how this worked to our advantage, as students expressed amazement not only that their peers could discuss science topics with them, but also that their peers had different ideas than they did about scientific phenomena. This eye-opening experience was described by many students in clinical interviews, and many students contrasted the rules of the new space with those in other social spaces, explicitly denying that they would ever have the same conversations with the same people (their peers at the school) face to face (in class or out). The ability of the teacher to "stake out" this new social territory as being for intellectual, student-centered, science-oriented discussion was a powerful point of leverage on the students' social interaction (Hoadley, 1999a; Hsi, 1997).

Over time, this advantage dissipated due to changes in the cultural surround. Within 3 years of this initial run, the Internet went from being unknown to being ubiquitous. Not only did a majority of students come to class with knowledge of hyperlinks and browsers, they had favorite search engines, Web sites, and deeply held beliefs about Internet usage. Our initial training needs decreased, and student access from home and from the popular nearby library skyrocketed. However, students came to class with strong expectations about what online discussion was like. Increasingly, students would mention AOL chat rooms, e-mail, and other online discussions in their interviews about the SpeakEasy, and it became more and more difficult to ensure that students held to the norms we tried to set in SpeakEasy. The teacher spontaneously began to differentiate the tool when introducing it to the class by describing how special it was, how experimental, and so on, and by explicitly contrasting it with AOL chat. Maintaining the sense of our online discussions as new social territory required deliberate effort.

Likewise, we were aided by invoking cultural norms specific to the classroom environment. Students might not have had a good idea of what scientific explanation, argument, and questions looked like before coming to this course, but this was a genre the teacher could invoke as the students learned these concepts during the semester. This prospect in particular suggests how delicately intertwined the nature of the cultural practices and the nature of the tool itself are and how locally (and temporally) specific they are. This example shows how enactments are a product of both the design and the context, mediated by the teacher and researchers, illustrating why treatment validity may be difficult to pin down.

Systemic and consequential validity: Equity and anonymity. Equity is an important issue, especially for middle school science, where girls, who have higher achievement than boys in the primary grades, begin a downward trend compared to their male peers, presumably due to social factors. In particular, girls are often disadvantaged in classroom talk (American Association of University Women Educational Foundation, 1992). Because this is a recognized problem in participation, and because inclusiveness is an important component of our model of productive discussions, we had a deliberate goal of ensuring equitable participation by members of both genders. In our engineering work, we saw that the ability to communicate asynchronously, without needing to interrupt or take the floor to contribute, was an important force toward inclusiveness. (Asynchronous, text-based communication was also anecdotally related to the ability of non-native speakers of English to participate in the discussions in our engineering work.) We also saw that anonymity was important for participants who might not have social status but wished to express their views. This, in particular, conflicted with earlier theories that had driven our work: specifically, a theory that representations that included social context information and were socially engaging would

promote ownership of ideas and motivate participation. It was for this reason that we had initially included face icons as part of the initial MFK system and had carried that feature through each iteration. However, we also heard that students were making use of anonymity in support of their participation, which would suggest that less social representations might be better. This became an important question for us as we investigated the role of identity in online participation and how our system affected both genders.

The initial MFK system had a limited set of pseudonymous identities that people could use to contribute anonymously, such as Minnie Mouse. These icons were initially created to allow users to participate who had not been previously set up in the system. We also saw the possibility that they could be used to contribute anonymously and therefore made it possible to contribute using one of these pseudonymous identities even after logging in as oneself. Initially, we questioned whether consistent pseudonymity was important, and several versions of the MFK were designed so that each person, when commenting anonymously, was given a separate anonymous identity, making it possible to identify which anonymous comments were made by the same or different individuals, even if the individual could not be identified. We did find in surveys that participants appreciated the ability to contribute anonymously. Some discussions were heavily anonymous (especially those discussing sensitive topics such as classroom atmosphere in the college engineering courses), whereas others had less anonymity. Of interest, in one semester with the four engineering instructors, we noticed much less anonymity in the discussions of the two courses led by female professors than in the two courses led by male professors. Gender certainly seemed to be playing some role in the participation structures.

Hsi and I undertook a more careful comparison in the middle school science classroom. Students were given free choice of anonymity, and girls contributed significantly more of the anonymous comments than boys (Hsi & Hoadley, 1997). Interviews with boys and girls revealed that the girls cited social safety (avoiding embarrassment) as the primary reason that online discussion was better than offline discussion. In what was expected to be a replication, we varied whether students were forced to attribute their comments to their real names and identities or were forced to not attribute their comments. Surprisingly, we saw no significant differences between participation in the two groups and no interactions between treatment group and gender (Hsi & Hoadley, 1997).

How could we explain these findings? In interviews with girls and boys in later semesters (with free choice of anonymity), girls often mentioned the option of anonymity as an important social feature that increased their comfort level in the discussion. Surprisingly, many of the girls who mentioned this never made anonymous comments in any discussions. As designers, we found this to be an exceptionally poignant example of a finding that would not have been uncovered without iterative design. We had created an interface feature that had important benefits for the collaboration without even being used! If use of the anonymity feature was independent of how the feature affected social comfort, how could we explain why some students used the anonymity feature while others did not?

It was around this time that we probed student beliefs about anonymity and attribution further. We surveyed, interviewed, and observed students to ascertain how they might view or use attribution in navigating or understanding student comments. Approximately half of the students navigated the comments in the discussion (chose which ones to read or in which order to read them) on the basis of attribution, and students frequently stated that they liked being able to tell who had contributed a comment before and after reading the contribution. Many students explicitly said that they avoided reading anonymous comments. This contradicted the impression held by many girls that anonymity was an important safety valve to allow students to honestly and safely express ideas to their peers. It appeared that students were less likely to read anonymous comments, which defeated the inclusivity purpose of the anonymity feature, one of the central aspects of our theory of productive discussions. Students might feel empowered to contribute to the discussion if they could do so anonymously, but their ideas were not being heard by other students. Around this time, we switched from the stand-alone MFK system to the Web-based SpeakEasy.

We got our big break by examining who was making anonymous comments. We found that rates of anonymity were surprisingly consistent for any given individual over time. That is to say, the percentage of comments made anonymously by a person in one discussion correlated very highly with the percentage of comments made anonymously by the same person in a later discussion. Also, the percentage of comments made by a person in a discussion correlated with rates of anonymity for other students in the same discussion. Thus, some discussions had a large amount of anonymous participation by many individuals, whereas others did not (Hoadley, 1999b). This was data we had previously collected but not examined in this way.

We finally uncovered a large part of the reason for anonymous contribution through informal observation and discussion with students in the classroom. Many students (not surprisingly) would skim the comments already in the discussion before contributing their initial opinion. If the students encountered mostly (or entirely) anonymous opinions, they themselves would contribute anonymously. This happened quite frequently because we had learned to seed discussions with comments to avoid an intimidating "blank slate" discussion. To avoid presenting these views as authoritative (coming from us as researchers), we added them anonymously. This anonymity would be perpetuated as increasing numbers of anonymous opinions accumulated, further discouraging students from contributing their views under their own name. The reason that some discussions had escaped this fate was that some students preferred to contribute before

reading others' comments. These students were basing their decisions about comment attribution on their own sense of confidence rather than on the prior contributions.

Responding to this realization, we designed a simple intervention that would encourage students to participate with attribution. Resurrecting an interface design we had employed earlier, we changed the system to force students to contribute their opinion on the topic before browsing others' opinions. We had dropped this feature when we had introduced it previously because users were reluctant to state their views without exploring the topic (especially for science topics that were new to them), but we found this reluctance could be overcome. We also emphasized in our oral introduction to the system that students should revise their opinions as often as their views changed, even during their first login session, if change was warranted. The new feature and the new instructions had three benefits: Students were less likely to comment anonymously (because they were basing their decisions on their own confidence rather than peer pressure exerted by the fictitious contributors of the seed comments), students were encouraged to develop a habit of revising their opinion-area comments, and we as researchers got the beneficial side-effect of having a true student pretest for the topic (which was ultimately part of the data collection technique for our individual learning measures.) Overall, student participation—reading and writing comments—remained equally high as without the new feature (actually trending toward an increase), gender balance of contributions remained high (with trends favoring girls), and anonymity (which had inhibited other students from reading the comments) dropped significantly.

In this way, through a design stance and a close involvement with the classroom, we short-circuited what might have been a long series of expensive studies that would have misled us about how anonymity could benefit the discussion. Indeed, our view on anonymity in discussion changed from believing anonymous participation was evidence of our theoretical notion of inclusiveness to believing it was a threat to inclusiveness. By designing a new technology feature and some new activities around the feature, we were able to maintain the sense of safety in the discussion by allowing the option of anonymous participation while greatly reducing the negative impact heavy use of that option previously implied. Consequential validity was vastly improved when we realized anonymous contribution was evidence of failure, not success.

Had we simply scattered our software to the four winds and tested outcomes, we might never have realized what conditions of use needed to be met nor would we have been able to proliferate those conditions as a theme and variations in a wide variety of contexts. When testing new tools, as we were, any sort of research on effectiveness would have been meaningless without giving the tools a chance to succeed by helping establish best practices of use. This point bears repeating. Certainly, although one may study the outcomes of any intervention in all the naturally occurring variations of use that might arise in the field (e.g., in the way some reform efforts are), these studies may not answer the question we really want to know, which is: What will happen if the reform really takes root? Without understanding the relation between designed intervention and enactment, we might get a lot of data, but it does not address meaningful questions about how best to educate or support learning.

Consider how differently this research might have unfolded if we had instead conducted only laboratory experiments. Certainly, because the discussions represented sustained effort on the part of the students, we would have had to make use of a demandingly long research protocol. The investment in participant hours required to run the experiment would have probably encouraged us to carefully pilot and then fix a particular set of instructions and a particular version of the interface. The iteration we conducted on a time scale of several years would have been far less likely. There is every likelihood we would have misinterpreted the role of gender and anonymity in the interface. Even if, by some miracle, we had uncovered the inconsistencies between girls' attitudes as a result of the presence of the anonymity option versus the effects of use of the anonymity option, we would not have had the informal observation that led us to not only a sensible explanation, but an easy remediation. This is the power of design-based research.

In this example, I have described how a particular discussion tool coevolved with various activities in a context of learning science. The moral of this story is not about the particulars of the design of an online discussion system (this is another interesting story told elsewhere, as in Hoadley & Linn, 2000; Hsi & Hoadley, 1997). Rather, it serves as an example of the crucial interrelation between the collaborative tool and the ways in which the tool is construed and embedded in local participants' activity structures. It also shows how a detective-like attentiveness to details and causes of social phenomena by participants (in this case, by the researchers and teacher) allows for a much greater degree of robustness, as idiosyncratic barriers to producing an effective instructional environment can be sniffed out and addressed through (sometimes trivially easy) intervention.

CONCLUSION

By engaging in design on both a technical and a social level, we were able to arrive at valuable insights in how to foster computer-supported collaborative learning. This central point has been argued by others at a theoretical level (Koschmann, 1996); here, I argue it from the point of view of our research on electronic discussion tools.

All empirical methods are faced with similar challenges for rigor—namely, to generate empirically consistent understandings and apply them appropriately with true consequential validity. Different research paradigms manage the need for rigor in different ways based on their different assump-

tions; naturalistic inquiry is inductive and (because it takes context as a primary independent variable) situation-specific, focused on developing and refining both an individual researcher's intimate understanding of the activities and practices through participation in the context. Interpretation is the core challenge. Experimental research, on the other hand, worries more about how to insulate the researcher's perspective from the work, thereby emphasizing (and, to some extent, limiting itself to) understandings that are generalizable across a wide variety of contexts. Control and comparison are often the core challenges. In the case of design-based research, the researcher is both a participant in a particular context and an agent for trying to generalize to other contexts. Here, implementation is one of the core challenges because the design-based researcher recognizes that any findings are composed of the interaction between design and enactment, between the general and the local. Iteration and replication are not checks against dishonest researchers or chance coincidences, but rather the fundamental mechanism for exploring how local and global interact, for probing the edges of design-oriented understandings. The downside is that design-based researchers hesitate to generalize across contexts until many designs and enactments are allowed to occur and to be studied formally. The upside is that the knowledge generated is applicable from the very beginning, a strong indication that we are progressing toward usable knowledge.

Only time will tell how this endeavor will fare in its ability to ferret out the kinds of knowledge that has been demanded of educational research. It will take further exploration to fully understand the trade-offs involved in design-based research related to bias, relevance, and rigor and even how design-based research might change the dissemination of educational research to match the assumptions in a design-based paradigm. However, the promise of having better alignment in research—certain and sure links from theories to hypotheses to interventions to data gathering activities to interpretation and application—should be a strong incentive to continue to pursue the design-based research approach.

ACKNOWLEDGMENTS

This work was supported by Grants EEC-9053807, MDR-9155744, RED-9453861, and EHR-9554564 from the National Science Foundation; Grant 200100273 from the Spencer Foundation; and a grant from the Evelyn Lois Corey Fellowship Program. The work was conducted while the author was at a variety of institutions, including the University of California at Berkeley, Graduate Group in Science and Mathematics Education; Stanford University School of Education, Learning Design and Technology Program; SRI International's Center for Technology in Learning; and Mills College, Department of Mathematics and Computer Science. Opinions expressed are those of the author and not the funders or employers. Portions of this article were previously published in *Proceedings of Computer Support for Collaborative Learning 2002.* The involvement of collaborators on the SpeakEasy research, especially Sherry Hsi, Doug Kirkpatrick, and Marcia C. Linn, is gratefully acknowledged; as are the productive discussions about design-based research with the members of the Design-Based Research Collective (http://www.designbasedresearch.org/).

Finally, feedback and encouragement from Lawrence Friedman, Bill Sandoval, and the anonymous reviewers are gratefully acknowledged.

REFERENCES

American Association of University Women Educational Foundation. (1992). *How schools shortchange girls*. Washington, DC: Author.

Bales, R. F. (1969). *Personality and interpersonal behavior.* New York: Holt, Rinehart & Winston.

Barab, S. A., Hay, K. E., & Yamagata-Lynch, L. C. (2001). Constructing networks of action-relevant episodes: An in situ research methodology. *Journal of the Learning Sciences, 10*(1), 63–112.

Bell, P., Davis, E. A., & Linn, M. C. (1995). The Knowledge Integration Environment: Theory and design. In S. Goldman & J. Greeno (Eds.), *Computer supported collaborative learning '95* (pp. 14–21). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2*(2), 141–178.

Collins, A. (1992). Toward a design science of education. In E. Scanlon & T. O'Shea (Eds.), *New directions in educational technology* (pp. 15–22). New York: Springer-Verlag.

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8.

diSessa, A. (1991). Local sciences: Viewing the design of human-computer systems as cognitive science. In J. M. Carroll (Ed.), *Designing interaction: Psychology at the human-computer interface* (pp. 162–202). Cambridge, England: Cambridge University Press.

Hoadley, C. (1998). Shaping social interactions for knowledge integration through technology. In B. K. Nichols, A. C. Kemp, & D. Jackson (Eds.), *71st NARST annual meeting* (p. 166). San Diego, CA: National Association for Research in Science Teaching.

Hoadley, C. (1999a). *Scaffolding scientific discussion using socially relevant representations in networked multimedia.* Unpublished doctoral dissertation, University of California, Berkeley.

Hoadley, C. (1999b, January). *Social text: Learning in online peer discussion in science.* Paper presented at the Winter Text Processing conference, Jackson Hole, WY.

Hoadley, C. (2002). Creating context: Design-based research in creating and understanding CSCL. In G. Stahl (Ed.), *Computer support for collaborative learning 2002* (pp. 453–462). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hoadley, C., & Bell, P. (1996, September). Web for your head: The design of digital resources to enhance lifelong learning. *D-Lib Magazine, 2*(8).

Hoadley, C., & Hsi, S. (1993). A multimedia interface for knowledge building and collaborative learning. In, *Adjunct proceedings of the International Computer Human Interaction Conference (InterCHI) '93* (pp. 103–104). Amsterdam: ACM Press.

Hoadley, C., Hsi, S., & Berman, B. P. (1995). The Multimedia Forum Kiosk and SpeakEasy. In P. Zellweger (Ed.), *Proceedings of the third ACM international conference on multimedia* (pp. 363–364). San Francisco: ACM Press.

Hoadley, C., & Linn, M. C. (2000). Teaching science through on-line, peer discussions: SpeakEasy in the Knowledge Integration Environment. *International Journal of Science Education, 22*, 839–858.

Hsi, S. (1997). *Facilitating knowledge integration in science through electronic discussion: The Multimedia Forum Kiosk.* Unpublished doctoral dissertation, University of California, Berkeley.

Hsi, S., & Hoadley, C. (1994). *An interactive multimedia kiosk as a tool for collaborative discourse, reflection, and assessment.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hsi, S., & Hoadley, C. (1997). Productive discussion in science: Gender equity through electronic discourse. *Journal of Science Education and Technology, 10*(1), 23–26.

Hsi, S., Hoadley, C., & Schwarz, C. (1992). *Scaffolding constructive communication in the Multimedia Forum Kiosk.* Unpublished course report for EMST 291B, University of California at Berkeley, Education in Math, Science, and Technology.

Kelly, A. E. (2003). Research as design. *Educational Researcher, 32*(1), 3–5.

Koschmann, T. D. (1996). *CSCL, theory and practice of an emerging paradigm*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lagemann, E. C. (2000). *An elusive science: The troubling history of education research*. Chicago: University of Chicago Press.

Lagemann, E. C. (2002). *Usable knowledge in education: A memorandum for the Spencer Foundation Board of Directors*. Chicago: Spencer Foundation. Retrieved October 14, 2003, from http://www.spencer.org/publications/usable_knowledge_report_ecl_a.htm

Latour, B. (1987). *Science in action*. Cambridge, MA: Harvard University Press.

Linn, M. C., & Burbules, N. C. (1993). Construction of knowledge and group learning. In K. G. Tobin (Ed.), *The practice of constructivism in science education* (pp. 91–119). Washington, DC: American Association for the Advancement of Science Press.

Linn, M. C., & Hsi, S. (2000). *Computers, teachers, peers: Science learning partners*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

National Research Council. (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Pea, R. (1992). Augmenting the discourse of learning with computer-based learning environments. In E. deCorte, M. Linn, & L. Verschaffel (Eds.), *Computer-based learning environments and problem solving* (pp. 313–343). New York: Springer-Verlag.

Robinson, V. (1998). Methodology and the research-practice gap. *Educational Researcher, 27*(1), 17–26.

Roth, W.-M. (2001). Situating cognition. *Journal of the Learning Sciences, 10*(1), 27–61.

Scardamalia, M., & Bereiter, C. (1992). An architecture for collaborative knowledge building. In E. deCorte, M. C. Linn, H. Mandl, & L. Verschaffel (Eds.), *Computer-based learning environments and problem solving* (Vol. 84, pp. 41–66). Berlin, Germany: Springer-Verlag.

Scardamalia, M., & Bereiter, C. (1994). Computer support for knowledge-building communities. *Journal of the Learning Sciences, 3*(3), 265–283.

Scardamalia, M., Bereiter, C., McLean, R. S., Swallow, J., & Woodruff, E. (1989). Computer-supported intentional learning environments. *Journal of Educational Computing Research, 6*(1), 55–68.