OVERLINE

Instructional Complexity and the Science Needed to Constrain It

Kenneth R. Koedinger[1]*, Julie L. Booth[2], David Klahr[3]

Science and technology have had enormous impact on many areas of human endeavor. Consider travel. Over the past two centuries, we have gone from horse-drawn carriages and unpaved roads to a vast and complex system of "planes, trains, and automobiles". However, during that same period, science and technology have had surprisingly little effect on another crucial area of the human enterprise: Education *(1)*. Although our nation has made great strides toward the goal of universal K-12 schooling, progress on the quality and content of that schooling has been inadequate. As Slavin *(2)* critically notes, at the "dawn of the 21st century, educational research is finally entering the 20th century" (p. 15). Classrooms look much the same today as they did when the automobile and the steam engine were invented, and many large-scale field trials of science-based technological innovations in education have yielded scant evidence of improvement in student learning *(3, 4)*, and the evidence in even the best studies is mixed (*5*).

To return to our transportation analogy: The design and engineering of our complex airline system required the concurrent solution of many problems: from the aerodynamics of control surfaces and the efficiency of engines, to the organization and coordination of the air traffic control system, the financial structure of the airline industry, and the management of a complex set of regulatory agencies. The challenge of improving education is similarly complex and interactive, and advances in the efficacy and quality of our educational systems will not be straightforward. The full complexity of education involves many important issues, such as cultural questions of values, but our focus is on the part of the full complexity that involves instructional decision-making in the context of determined instructional goals.

We demonstrate how instructional complexity implies that the typical "binary" debates in education are inherently unproductive. For too many years, the ability to translate good psychological science into effective educational practice has been hampered by the so called "reading wars", "math wars", and most recently "science wars" *(6, 7)*. In the following sections we describe a few of the battles and skirmishes from these wars and explain why – although often based on very solid empirical and theoretical work -- they are unlikely to lead to solutions. We then elaborate on the factors that define the complexity of instructional design. Finally, we attempt to quantify that complexity and suggest a way to manage it by principled isolation of primary functions of instruction – returning again to our analogy of the development of the air transportation system. We close with recommendations for ways to advance research on learning and to apply it effectively to the field of education as a whole.

[1]Human Computer Interaction Institute and Psychology Department, Carnegie Mellon University.
[2]Psychological, Organizational, and Leadership Studies, Temple University.
[3]Psychology Department, Carnegie Mellon University
*Correspondence to: koedinger@cmu.edu

Two Sides to Every Educational Debate?

Researchers have come up with many dichotomies to describe educational methods *(8)* (see Table S1). Typically, such dichotomies contrast a focused approach -- involving more support and more emphasis on the "basics" (e.g., traditional instruction; direct instruction) against a more open-ended approach, involving greater student engagement and more emphasis on understanding (e.g., reform instruction; constructivism). This widespread dichotomization of instructional methods is troubling for several reasons. One is that the there is no consensus on the meaning of the terms of interest. For example, a particular instance of instruction that one person calls "direct instruction" may be called "inquiry" by another researcher. Such deep ambiguities derive from a widespread tendency to apply these labels to vaguely described procedures, rather than to clear operational definitions of instructional practices *(9, 10)*.

Another fundamental problem with such dichotomies is that even when the practices are reasonably well defined, there is not yet a strong evidential base for deciding which of the two choices is optimal for learning. To the contrary, empirical investigations of general instructional methods, including controlled laboratory experiments in the fields of cognitive and educational psychology, often fail to yield such consensus. For instance, although some researchers have produced credible empirical evidence for learning benefits of *immediate* feedback *(11)* others have done the same for *delayed* feedback *(12)*. Similar controversy exists with data supporting the use of concrete materials *(13)* vs. abstract materials *(14)* to promote learning and transfer. Even for those practices for which a consensus appears to be emerging, the evidence in favor of the purported principles is not typically strong. For example, of the nine scientifically-based recommendations in the IES Practice Guide on organizing instruction and study *(15)*, only two are considered to have strong evidence in their favor. Three recommendations have low support, coming from just a few studies in a limited range of domains, and often with a small, homogeneous population (though new studies are accumulating (cf., *16*)). The four remaining recommendations have moderate support, indicating that the findings are not necessarily generalizable to a wider population.

Further complicating the picture is the fact that in cases where principles have been tested across a wide variety of content domains or student populations, results often vary. For example, instruction that is effective for simple skills has been found to be ineffective for more complex skills *(17)*, and techniques such as prompting students to provide explanations (a strong recommendation in *(15)*) may not be universally effective (*18*). The many positive demonstrations have targeted domain principles or laws (e.g., in mathematics and science), but there are negative results for procedural skills or more probabilistic or sparse-featured categories *(9)*. The effectiveness of different approaches is often contingent on student population or level of prior achievement or aptitude; in fact, some interventions may be particularly important for low-achieving students *(19, 20)*. Thus, while relying on binary instructional decisions may be useful at the level of the *individual* student (e.g., will this student learn better right now if I give her feedback or if I let her grapple with the material for a while?), the search for *general* methods of instruction that optimize the effectiveness, efficiency, and level of student engagement is far more challenging.

Complexity of Instructional Design

Of the many factors inherent in real-world learning situations that might have an impact on instructional design we describe three of particular importance — type of instructional technique, subsequent decisions about dosage, and timing of the intervention—and discuss how each factor independently contributes to the complexity of instructional design. Our approach is to (a) first indicate the vast size of the space of potential combinations of different levels of all of these factors, and (b) then suggest a way to avoid the combinatorial explosion by conceptualizing the essential processes in terms of a "function space" of the learning, knowledge, and assessment outcome functions of instruction.

Instructional techniques

It is tempting to think that the key to improving how students learn is to devise research-based instructional methods and to apply them broadly and consistently. However, researchers in education and psychology have, in fact, already proposed a wide array of instructional techniques, and many experiments have demonstrated how such techniques improve student learning beyond that produced by comparable instruction without the targeted technique. Just how many such techniques are there? One can easily find many lists of learning or instructional principles that suggest instructional techniques and point to supporting research *(15, 16, 21-28)*. Each of these sources lists between 3 and 25 principles, and the overlap is far from complete. An in-depth synthesis of 9 such sources yielded an estimate of 30 independent instructional principles (see Table 1; a detailed mapping of each principle to the 9 sources can be found in External Database S1).

Dosage and implementation choices

Other significant sources of uncertainty and complexity in instructional decision-making stem from the fact that many of these distinctions are not binary, but rather have multiple possible strength values or else are continuous (e.g., the ratio of examples to questions/problems given in an assignment, the time spacing between practices, or the time delay of feedback). These variables substantially increase the number of options to consider in instructional design and in a general theory of learning and instruction. Compounding this problem of complexity is the fact that these choice dimensions are mostly compatible with each other – that is, almost all of them can be combined or sequenced together with any other *(29)*. Figure 1 illustrates how choices on one dimension can be independently combined with choices on other dimensions to produce a vast space of reasonable instructional choice options. The path of the thicker black arrows shown in Figure 1 indicates one set of choices on six instructional dimensions: The interval between practices is gradually widened, the practices involve retrieval tasks or problems to solve rather than examples to study, the problems are presented in an abstract form, delayed feedback (after a full problem solution is attempted) is provided, the problems interleave different knowledge goals or topics, and students are prompted to "self-explain" the principles that justify their problem solving steps. As indicated in Figure 1, this particular path is just one of the 6^3 that could be traversed in this instructional design space.

Intervention timing

The number of instructional options is further increased given evidence that the optimal choice may not be the same early in learning as it is later in learning. Research on the use of worked examples during problem solving has demonstrated that, for novice students, extensive use of worked examples, in place of many problems, enhances learning. However, as students develop expertise, shifting to pure problem solving practice is more effective *(30)*. Another example from research on the spacing effect is that more narrow spacing is optimal for beginning learning, but gradually wider spacing is optimal as the learner advances *(31)*. More generally, many researchers have suggested that effective instruction should provide more structure or support early in learning or for more difficult or complex ideas and fade that assistance as the learner advances *(32-35)*.

Quantifying the Complexity of Instructional Design

The three factors described above—instructional techniques, dosage and implementation choices, and timing of intervention—can be combined to estimate the size of the space of instructional choices. With 30 instructional techniques and three levels of dosage, we estimate 3^{30} or *over 205 trillion instructional choices*. If we consider only 15 instructional techniques (because some are similar to each other), but include a simple intervention-timing factor, namely that the optimal choice may be different early than late in instruction, we get 3^{15*2} or 205 trillion once again.

To the extent that some of the combinations are not possible *(36)*, or may not make sense in a particular content area (e.g., what would faded practice of abstract problems look like in a History class?), this formulation may appear to over-estimate the size of the choice space. However, we believe that, in fact, it under-estimates for three reasons. First, many dimensions have more than three possible values such as the time between spaced practices or the ratio of worked examples to problems. Second, there may be more than two time points where the instructional optimum changes. Third, different knowledge needs in different domains often require a different optimal combination. For example, the optimal practice scheduler used by Pavlik and Anderson *(31)* continually adjusts the spacing interval for each student on each practice of each knowledge component. As another example, when the target knowledge is simple facts (or students are more advanced), requiring recall and use of knowledge produces more robust learning, but for complex problem-solving skills, a significant amount of worked example study is better (*3*; see rows 2 and 3 in Table S1). Considering these additional factors, it is likely that the size and complexity of the full space of instructional choices is much greater than our estimate suggests.

The vast size of this space sheds light on the research-practice gap. Our analysis reveals that the all-too-common binary debates about educational policy choices -- in the scientific literature as well as in the public forum -- tend to ignore and obscure the complexity that a productive science and engineering effort must address.

Five Scientific Recommendations for Taming Instructional Complexity

What can be done to address instructional complexity? We make five recommendations to advance instructional theory and to maximize its relevance to educational practice.

1. Searching in the function space. Our first recommendation takes us back to the analogy between educational systems and modern transportation systems, in this case to the invention of the airplane. In a cogent analysis of how the Wright brothers managed to win the race to construct the first powered, controllable, and human-carrying flying machine, Gary Bradshaw (*37*) identified a crucial difference between the Wrights' strategy and those of their competitors. Most of the inventors of that time were working in what Bradshaw calls the "design space", in which each new design represented a choice of one of the several values from each of the potentially relevant variables. As he puts it "… to these men, the airplane consisted of a set of structures, such as wings, fuselage, propulsion plant, etc. Developing an airplane meant exploring the set of possible designs." (p 247). However, as Bradshaw insightfully notes, "Design-space search is inherently inefficient for two reasons: The design space is large, and global measurements (time and distance in flight) provide little guidance in moving through the space." (p 247). Rather than searching this huge space, the Wrights used what Bradshaw calls a "function-space search". They hypothesized that the key functions were relatively independent, and could be studied successfully without simultaneously varying all the other factors. "Lift could be addressed without regard for lateral control, and vice versa." (p 248). Function space search is much more tractable.

Are the Learning Sciences poised to approach the challenge of instructional complexity by following a strategy analogous to the Wrights' function space search? If so, what would the analogous function space be? We suggest that a recently developed conceptualization, organized around a three part taxonomy of knowledge, learning processes, and instructional methods – called the "KLI framework" (*21*) -- provides a powerful answer to this question. The KLI framework specifies three layers of functions of instruction from the more distal and observable to the more proximal and hidden: 1) Instruction is intended, ultimately, to yield better assessment outcomes, that is, enhanced learner performance in activities and tasks in future work or academics. 2) To do so, instruction must change learners' knowledge base in ways that produce enhanced future performance. 3) For the knowledge base to change, the learners' minds must execute learning processes or mechanisms that implement these changes. Using KLI, we can specify different functions to be achieved at each layer. The most distal, but observable, functions of instruction are assessment outcomes: long-term retention, transfer to new contexts, or desire for future learning. More proximal, but unobservable, functions of instruction are to change different kinds of knowledge (malleable mental structures or processes): facts, procedural skills, principles, learning skills, or learning beliefs and dispositions. The most immediate and unobservable functions of instruction are to support different kinds of learning processes or mechanisms: memory and fluency building, induction and refinement, or understanding and sense making (*38*).

Using these distinctions, we can create a function space that reduces the instructional design space. For instance, instead of asking what instructional choices optimize learning in general, the functions of instruction at each layer suggest more focused questions: Which instructional choices best support memory to increase long-term retention of facts? Which are best for induction of general skills that produce transfer of learning to new situations? Which are best for sense making processes that produce learning skills and higher learner self-efficacy toward better future learning?

Indeed, we can associate different subsets of the instructional design dimensions with individual learning functions. For example: spacing enhances memory, worked examples enhance induction, and self-explanation enhances sense making (see External Database S1 for more examples of mappings of instructional methods to learning functions). We do not pretend to have solved the instructional complexity problem with this function space proposal. Nevertheless, we offer it as a productive path forward. The success of the approach depends on partial decomposability (*39*), that is, on some independence of effects of instructional variables: Design configurations that are optimal for one function (e.g., memory) should not be detrimental to another function (e.g., induction). To illustrate the plausibility of (conditional) independence, consider that facts (or skills later in acquisition) require memory but not induction, thus a designer can focus on the subset of instructional variables facilitating memory, independent of other variables.

We need theoretical work to better understand the possibility of this kind of decomposability and to gain insight for when main effects of instructional methods can be combined without concern for interaction effects. One powerful method for theory development is creating computational models of learning that can engage in learning like human students do (*31, 40-42*). Such computational theories not only provide precise, replicable predictions, but can also be used to compare alternative instructional methods – essentially as instructional crash test dummies. Examples include a computational learning study showing that providing a curriculum that separates instruction into "bite size" pieces ("one subgoal at a time") enhances learning (*40*) over one that does not, or another that shows that interleaving problems of different kinds enhances learning (*41*).

2. Experimental tests of instructional function decomposability. In addition to theoretical investigations of instructional function decomposability, there is an urgent need for empirical tests of decomposability. One important case is multi-factor studies of how the nature of the to-be-acquired knowledge can change the optimal instructional choice. For example, we have noted (*21*) how the contradiction between the testing effect (*43*) recommendation--that instruction should include more questions to answer--and the worked example effect (*32, 44, 45*) recommendation--that instruction should include more question-answer pairs to study--may be resolved by distinguishing the nature of the targeted knowledge (i.e., specific facts or general problem-solving skills). If the instructional goal is long-term retention (an outcome function) of a fact (a knowledge function), then better memory processes (a learning function) are required and more testing than study will optimize these functions. However, if the instructional goal is transfer (a different outcome function) of a general skill (a different knowledge function), then better induction processes (a different learning function) are required and more worked example study will optimize these functions. The ideal experiment to test this hypothesis is a two-factor study that varies both the knowledge content (fact learning vs. general skill) and the instructional strategy (example study vs. testing). That other optimal instructional choices may be function-specific is suggested by conflicting results across studies of other instructional methods where different results appear dependent on the nature of the knowledge goals. These include studies of prompting for self-explanations (*18, 46*), spaced practice (*47*), and blocked vs. random practice schedules (*48*). We need more experiments that directly test these possible function-specific outcomes.

3. Massive online multi-factor studies. It is now possible to run massive learning experiments online that can involve thousands of participants and can vary many factors at once. For example, a recent experiment with a fraction numberline game involved 7000 students and varied five factors (with 2 to 25 levels each) (*49*). Such studies (*49, 50*) can accelerate the accumulation of empirical facts that can drive instructional theory development. The point is not to build up to multi-factor experiments to address the full instructional complexity space discussed above. Instead, the point is to test hypotheses suggested by the function space -- in particular, to identify in context of a particular instructional function, what instructional dimensions can or cannot be treated independently of each other.

So far, these studies have emphasized experimenting with system features (usually elements of the graphical user interface) rather than with instructional methods (*51*). Designing such studies to vary instructional methods would not be hard, but there is a major unsolved problem in convenient and consistent access to ideal outcome variables in such studies. Proximal variables measuring student engagement and local performance are easy to collect (e.g., how long/much a game or technology is played or used, proportion correct within game context); however, measures of students' local performance and their judgments of learning are sometimes not only unrelated, but even negatively correlated with desired robust learning outcomes (*52*).

Solutions to this robust learning metric problem would open this methodology to a huge potential for productivity. For example, if data from multi-factored instructional alternatives within widely used online tutors or games (e.g., for learning scientific inquiry in 6th grade) were tied to downstream data on relevant student performance of students (e.g., science achievement in the 7th grade common Core assessment), data-driven decision theoretic methods (e.g., *53*) could be used to select these alternatives based on robust learning outcomes.

4. Learning data infrastructure. A huge opportunity exists to develop an infrastructure to support more and better course-based data collection and experimentation. Massive instructional experiments are essentially going on all the time in schools and colleges across the country. Because collecting data on such activities has historically been expensive, all but a very few are not monitored (i.e., by tracking variations in instruction and the resulting changes in student computes). However, technology is increasingly providing low cost ways to instrument the learning experience for data collection. These technologies include both ever-better video and audio recording methods and software and web-based environments in which students interact, such as online courses, educational simulation games and simulations, online homework systems, intelligent tutoring systems, and discussion boards.

More broadly, investment is needed in infrastructure to facilitate large-scale data collection efforts in schools, particularly in urban and low-income school districts. Such efforts require improved technology for managing and delivering large databases, so that data collected in the real world can be accessed by decision-makers and researchers and analyzed in a user-friendly way. Two such current efforts include LearnLab's huge educational technology data repository (*54*) and the Gates Foundation's Shared Learning Infrastructure (*55*).

5. School-researcher partnerships. To make effective use of a learning data infrastructure, we need a corresponding collaborative problem-solving infrastructure that can facilitate interaction between researchers, practitioners, and school administrators. When school cooperation is well-managed and most or all of an experiment is computer-based, large well-controlled experiments can be run in courses with substantially less effort than an analogous lab study. For example, LearnLab's school-collaboration support led to many, large *in vivo* experiments, which are principle-testing controlled studies in the context of classrooms or online courses.

We need more *in vivo* experimentation to better address issues of context, including the knowledge content being taught, differences in student prior knowledge, in student cultural background, in teachers, and in school environment. Testing instructional principles in different course contexts benefits theory development. Cross-context comparison of studies of prompting students to self-explain indicate that verbal, reflective reasoning enhances learning in contexts involving more discrete rule-like generalizations, which tend to occur in math and science domains (*56*, *57*). However, more implicit, non-verbal learning mechanisms are better for more diffuse, probabilistic categories, which tend to occur in language and perceptual domains (*18*, *58*).

More practically, a lab-derived principle may not scale to real courses because non-manipulated variables may change in moving from the lab to a real course and the change in the background conditions may change learning results. In an *in vivo* experiment these background conditions are not arbitrarily chosen by the researchers, but instead determined by the existing context. Thus, they detect limits to generalization more quickly before moving to long, expensive randomized field trials.

School-researcher partnerships are useful not only for facilitating experimentation in real learning contexts, but also for designing and implementing new studies that address practitioner needs. One such effort, the Strategic Education Research Partnership (SERP), establishes long-term research relationships with school systems. Interdisciplinary teams of researchers are recruited to work with practitioners on projects to solve "use-inspired" problems in education (*59*). More generally, the Institute of Education Sciences at the U.S. Department of Education has recently released a call for researcher-practitioner partnerships (*60*).

In addition to school administrators and practitioners, effective partnerships must be comprised of a variety of critical research perspectives, including a combination of domain specialists (e.g., Biologists, Physicists, etc.) who bring a deep understanding of the content to be taught, Learning Scientists (Psychologists, HCI experts), who bring an understanding of the nature and development of learning and cognition, and education researchers (Physics and Math educators), who bring an understanding of teaching practices, standards, teacher professional development, and constraints evident in the education system. As part of these partnerships, it will be important to forge compromises between the control desired by social science and STEM domain researchers (*61*) and the flexibility demanded by the realities of real-world classrooms. One such opportunity may come from practitioners and education researchers involving more domain specialists and psychologists in ongoing design-based research (DBR) efforts, in which iterative changes are made to instruction in a closely observed, natural learning environment in order to examine the effects of multiple factors within the classroom context (*62*).

A Call to Action

The endeavors laid out in our recommendations are no small charge, and would require the necessary stakeholders (schools, practitioners, education researchers, cognitive scientists/psychologists, domain specialists) to devote time and energy and to seriously reexamine their assumptions about the types of research that are useful. However, we believe these efforts are necessary in order to constrain the challenge of instructional complexity. We see great signs of promise including sustained science-practice infrastructure funding programs, creation of new learning science programs at universities, and emergence of new fields and associated conferences, such as Society for Research on Educational Effectiveness (*63*) and Educational Data Mining (*64*), that bring science and technology to bear on the challenge of optimizing educational outcomes.

To be sure, today's air transportation system was created not by a single key insight but rather by a long period of accumulated science and engineering, in disciplines ranging from from physics to chemical engineering to human factors. Analogously, for science and technology to have a major impact on our vast and complex educational system, we need to recognize the tremendous scientific challenge inherent in that goal, and continue to bring together influences from multiple disciplines in the learning and education sciences.

References and Notes:

1. Note: Recent efforts and investment in the science of learning is a sign of progress *(65)*
2. Slavin, R. (2002). Evidence-based education policies: transforming educational practice and research. *Educational Researcher, 31*, 15-21.
3. Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., Means, B., ..., & Sussex, W. (2007). *Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort.* Report provided to Congress by the National Center for Education Evaluation.
4. http://ies.ed.gov/ncee/wwc/reports/hs_math/cog_tutor/index.asp
5. http://ies.ed.gov/ncee/wwc/interventionreport.aspx?sid=87
6. Klein, D. (2007) A quarter century of US 'math wars' and political partisanship. *Journal of the*

British Society for the History of Mathematics, *22*. 22-33.
7. Anderson, K. (2000) *The Reading Wars: Understanding the Debate Over How Best to Teach Children to Read (Book Review)*, L.A. Times (June 18, 2000). SSRN Link
8. Note: For example, *(66)* experimentally contrast classes "taught either by traditional or by interactive learning style".
9. Klahr, D. & Li, J. (2005). Cognitive research and elementary science instruction: From the laboratory, to the classroom, and back. *Journal of Science and Educational Technology, 4*, 217-238.
10. Klahr, D. (in press). What do we mean? On the importance of not abandoning scientific rigor when talking about science education. *Proceedings of the National Academy of Sciences*.
11. Corbett, A. T. & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In *Proceedings of ACM CHI'2001 Conference on Human Factors in Computing Systems*, 245-252
12. Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207-217.
13. Paivio, A. (1965). Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 4*, 32-38.
14. Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math. *Science*, *320*, 454-455.
15. Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K.R., McDaniel, M., and Metcalfe, J. (2007). *Organizing Instruction and Study to Improve Student Learning* (NCER 2007-2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.
16. Dunlosky, K.A., Rawson, E.J., Nathan, M.J., Willingham, D.T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1): 4-58.
17. Wulf, G., & Shea, C.H. (2002). Principles derived from the study of simple motor skills do not generalize to complex skill learning. *Psychonomic Bulletin and Review, 9,* 185-211.
18. Wylie, R., Koedinger, K. R., & Mitamura, T. (2009). Is self-explanation always better? The effects of adding self-explanation prompts to an English grammar tutor. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 1300–1305). Austin, TX: Cognitive Science Society.
19. *Goska*, R. E., & *Ackerman*, P. L. (*1996*). An aptitude-treatment interaction approach to transfer within training. *Journal of Educational Psychology, 88,* 249-259.
20. Kalyuga, S. (2007). Enhancing instructional efficiency of interactive e-learning environments: A cognitive load perspective. *Educational Psychology Review*, *19*, 387-399.
21. Koedinger, K. R., Corbett, A. C., & Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36,* 757-798.
22. Mayer, R. E. (2001). Multimedia learning. New York: Cambridge University Press.
23. Graesser, A. C. (2009). Inaugural editorial for Journal of Educational Psychology. *Journal of Educational Psychology, 101*(2), 259-261.
24. http://www.learnlab.org/research/wiki/index.php/Main_Page
25. Asleitner, H. (2005). Principles of effective instruction: General standards for teachers and instructional designers. *Journal of Instructional Psychology, 32*, 3-9.
26. National Research Council (1999). *How People Learn: Brain, Mind, Experience, and School.* J.D. Bransford, A.L. Brown, & R.R. Cocking, Eds. Washington, D.C.: National Academies Press.
27. Merrill, M. D. (2001). Components of instruction toward a theoretical tool for instructional design. *Instructional Science, 29*(4-5), 291-310.
28. Gagné, R. M. (1965). *The conditions of learning and theory of instruction* (1st ed.). New York, NY: Holt, Rinehart & Winston.
29. Note: A few exceptions do exist. For example, the timing of feedback is not relevant on study trials or examples where there is no student response to which to provide feedback.
30. Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1) 23–31.
31. Pavlik, P.I. & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied, 14*, 101-117.
32. Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Combining fading with prompting fosters learning. *Journal of Educational Psychology, 95*, 774-783.
33. Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the Learning Sciences*, 14, 69-110.
34. Pimsleur, P. (1967). A memory schedule. *The Modern Language Journal*, *51*(2), 73-75.
35. Wozniak, P.A., & Gorzelanczyk, E.J. (1994). Optimization of repetition spacing in the practice of learning. *Acta Neurobiologiae Experimentalis, 54*, 59-62.
36. Note: Another potential source for reducing the estimation is the hope that the learning outcomes of choices combine independently (i.e., they produce "main effects" without "interactions"). In this hopeful scenario, there would be 30*3*2 = 180 choices (not considering the other sources of complexity mentioned next). More studies involving a factorial combination of two or more dimensions (e.g., a 2x2 design) are needed, as is better theory of dependencies between dimensions.
37. Bradshaw, G. (1992). The airplane and the logic of invention. In R. Giere & H. Feigl (eds.), *Cognitive Models of Science*. University of Minnesota Press.
38. Note: The categories of memory, induction, and sense making are described in greater depth in (*21*) and with some further elaboration on sense making in (*67*).
39. Simon, H. A. (1969) Sciences of the artificial. Cambridge, MA: M.I.T. Press.
40. VanLehn, K. (1987). Learning one subprocedure per lesson. *Artificial Intelligence, 31*, 1–40.
41. Li, N., Cohen, W., & Koedinger, K. (2012). Problem order implications for learning transfer. Proceedings of the 11th International Conference on Intelligent Tutoring Systems. Chania, Greece.
42. Li, N., Schreiber, A., Cohen, W., & Koedinger, K. (2012). Efficient complex skill acquisition through representation learning. *Advances in Cognitive Systems, 2*, 149-166.
43. Karpicke, J.D. & Blunt, J.R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 11*, 772-775.
44. Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*, 59–89.
45. Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7*(1), 1-39.
46. Williams, J.J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. Cognitive Psychology, 66, 55-84.
47. Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect. Journal of Applied Psychology, 84, 795–805.
48. Wulf, G. & Shea, C.H. (2002). Principles derived from the study of simple skills do not generalize to complex. Psychonomics Bulletin and Review, 9, 185-211.
49. Lomas, D., Forlizzi, J., & Koedinger, K.R. (in press). Optimizing challenge in an educational game using large-scale design experiments. To appear in ACM Conference on Human Factors in Computing Systems (CHI 2013).
50. Andersen, E., Liu, Y., Snider, R., Szeto, R., and Popovic, Z. (2011). Placing a value on aesthetics in online casual games. ACM Conference on Human Factors in Computing Systems (CHI 2011), May 7–12, 2011, Vancouver, BC, Canada.
51. Note: It is worth highlighting that technology-based experiments have not only addressed cognitive issues, but broader non-cognitive issues including motivation (*68*), values (*69*), metacognition (*70*), classroom dialogue and argumentation/epistemic skills (*71*).
52. Bjork, R. A., Dunlosky, J., & Kornell, N. (2012). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417-44.
53. Chi, M., K. VanLehn, D. J. Litman, and P. W. Jordan (2011). Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. User Model. User-Adapt. Interact. 21.1-2, pp. 137–180.
54. http://www.learnlab.org/technologies/datashop/index.php
55. https://www.inbloom.org
56. Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26(2).*
57. Hausmann, R. G. M., & VanLehn, K. (2007). Explaining self-explaining: A contrast between content and generation. In R. Luckin, K. R. Koedinger & J. Greer (Eds.), *Proceedings of AI in Education*, pp. 417-424. Amsterdam: IOS Press.
58. Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: overgeneralization in the face of exceptions. Journal of Experimental Psychology: General. Advance online publication.
59. http://www.serpinstitute.org/about/overview.php
60. http://ies.ed.gov/funding/ncer_rfas/partnerships.asp
61. Note: This may be an easier jump for scholars in domains which have greater tolerance for noisy data (e.g., Geosciences).
62. Barab, S. A. (2006). *Design-Based Research: A methodological toolkit for the learning sciences*. In K. Sawyer (ed.) Handbook of the Learning Sciences (pp. 153-170), Cambridge , MA: Cambridge University Press.
63. https://www.sree.org/
64. http://www.educationaldatamining.org/

65. Meltzoff, A.N., Kuhl, P.K., Movellan, J., Sejnowski, T.J. (2009). Foundations for a new science of learning. *Science, 325*, 284-288.
66. Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science, 13,* 862-864
67. Resnick, L., Asterhan, C., & Clarke, S. (Eds.) (2013). *Socializing Intelligence through Academic Talk and Dialogue.* Washington, DC: American Educational Research Association.
68. Bernacki, M. L., Nokes-Malach, T. J., & Aleven, V. (2013). Fine-grained assessment of motivation over long periods of learning with an intelligent tutoring system: Methodology, advantages, and preliminary results. In R. Azevedo and V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (pp. 629-644). NY: Springer.
69. Walkington, C., Petrosino, A., & Sherman, M. (in press). Supporting algebraic reasoning through personalized story scenarios: How situational understanding mediates performance and strategies. Mathematical Thinking and Learning.
70. Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. Learning and Instruction, 21, 267-280.
71. Rosé, C. P. & Tovares, A. (in press). What Sociolinguistics and Machine Learning Have to Say to One Another about Interaction Analysis, in Resnick, L., Asterhan, C., Clarke, S. (Eds.) Socializing Intelligence Through Academic Talk and Dialogue, Washington, DC: American Educational Research Association.
72. Dempster, F.N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology, 79*, 162-170.
73. Bahrick, H.P., Bahrick, L.E., Bahrick, A.S., & Bahrick, P.E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, *4*, 316-321.
74. Pashler, H., Rohrer, D., Cepeda, N.J., & Carpenter, S.K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review, 19,* 187-193.
75. Pavlik Jr., P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied, 14*(2), 101-117.
76. McDaniel, M.A., Roediger, H.L., and McDermott, K.B. (2007). Generalizing test enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 200-206.
77. Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory and Cognition, 36*, 438–448.
78. 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*, 75–86.
79. Dean, D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education*, *91*, 384-397.
80. de Jong, T. (2005). The guided discovery principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 215–228). New York: Cambridge.
81. Chi, M.T.H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439-477.
82. McNamara, D.S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38*, 1-30.
83. Craig, S.D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). The deep-level-reasoning-question effect: The role of dialogue and deep-level-reasoning questions during vicarious learning. *Cognition and Instruction, 24*, 565-591.
84. Mathan, S. A. & Koedinger, K. R. (2005). Fostering the Intelligent Novice: Learning from errors with metacognitive tutoring. *Educational Psychologist, 40*(4), 257-265.
85. Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science, 14*, 511–550.

Acknowledgements: The preparation of this paper was supported in part by National Science Foundation Grant SBE-0836012 to Carnegie Mellon University and by Department of Education Grants R305A100074 to Temple University and R305A100404 to Carnegie Mellon University. The views expressed are those of the authors and do not represent those of the funders. Thanks to Vincent Aleven, Anna Fisher, Nora Newcombe, Suzanne Donovan, and Tim Shipley for comments and suggestions on earlier drafts.

10.1126/science.1238056

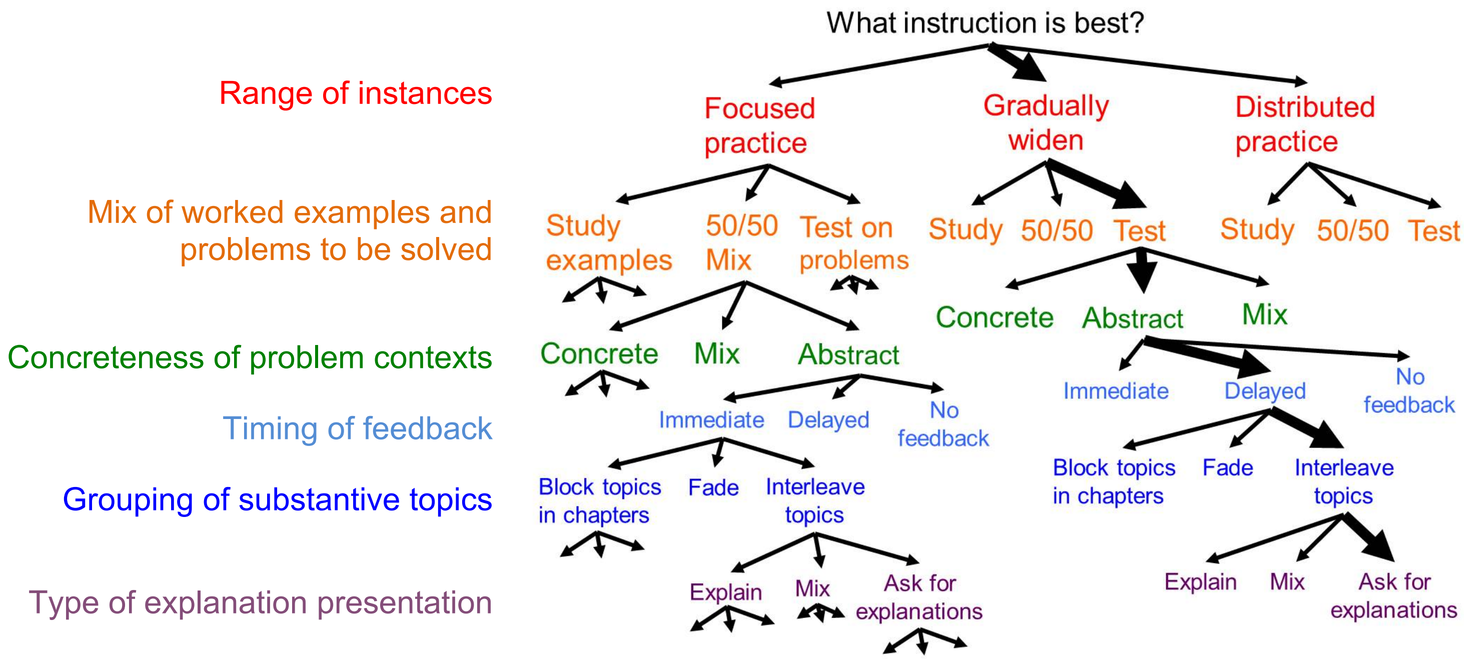
Table 1. List of principles for Instructional Design. Principles address three different functions of instruction: A) Memory/Fluency, B) Induction/Refinement, and C) Understanding/Sensemaking

Figure 1. Different choices along different instructional dimensions can be combined and produce a vast set of instructional options. The path with thicker arrows illustrates one set of choices for six dimensions of choice among the 729 different paths that could be selected among the six three-valued instructional dimensions shown here. We estimate trillions of such options given the 30 instructional choice dimensions that learning research has identified.

Table 1. List of principles for Instructional Design. Principles address three different functions of instruction: A) Memory/Fluency, B) Induction/Refinement, and C) Understanding/Sensemaking

		Principle	Description
A. Memory/Fluency	1	Spacing principle	Spaced practice > massed practice
	2	Scaffolding principle	sequence instruction towards higher goals > no sequencing
	3	Exam Expectations principle	students expect to be tested > no testing expected
	4	Testing principle	Quizzing > no quizzing
	5	Segmenting principle	lesson presented in learner-paced segments > as a continuous unit.
	6	Feedback principle	provide feedback during learning > no feedback provided
B. Induction/Refinement	7	Pre-Training principle	students know main concepts prior to lesson > don't know main concepts
	8	Worked example principle	Worked examples + problem solving > problem solving alone
	9	Concrete grounding principle	grounded representations > abstract representations
	10	Guided attention principle	words include cues about organization > no organization cues
	11	Linking principle	integration of instructional components > no integration
	12	Goldilocks principle	instruction at appropriate level > instruction that is too hard or too easy
	13	Activating Pre-Conceptions principle	student's prior knowledge activated for lesson > no connection to prior knowledge
	14	Corrective feedback principle	immediate feedback on errors > delayed feedback
	15	Interleaving principle	content spread out through lesson > blocked presentation of content
	16	Application principle	practice applying new knowledge > no application
	17	Comparison principle	comparing multiple instances > only one instance
	18	Variability principle	comparing varied instances > comparing similar instances
C. Understanding/Sensemaking	19	Coherence principle	extraneous words, pics, sounds excluded > included
	20	Redundancy Principle	verbal descriptions presented in audio or written > both together
	21	Temporal contiguity principle	corresponding information presented close together in time > far apart
	22	Spatial contiguity principle	corresponding information presented close together in space > far apart
	23	Multimedia principle	Graphics + verbal descriptions > verbal descriptions alone
	24	Modality principle	verbal descriptions presented in audio > written verbal descriptions
	25	Anchored learning principle	real world problems > abstract problems
	26	Metacognition principle	metacognition supported > no support for metacognition
	27	Explanation principle	Prompt for self-explanation > no prompting
	28	Questioning principle	time for reflection & questioning > instruction alone
	29	Cognitive Dissonance principle	considering incorrect/alt. perspectives as part of instruction > only correct instances
	30	Interest principle	instruction that is relevant to the student>instruction that isn't relevant

Figure 1. Different choices along different instructional dimensions can be combined and produce a vast set of instructional options. The path with thicker arrows illustrates one set of choices for six dimensions of choice among the 729 different paths that could be selected among the six three-valued instructional dimensions shown here. We estimate trillions of such options given the 30 instructional choice dimensions that learning research has identified.



Supplementary Materials:

Table S1

External Database S1

Supplementary Materials:

Table S1. Sample dichotomous educational debates. Italicized entries indicate the condition has experimental support.

More instructional assistance	**More challenge during instruction**	Sources
Massed practice	***Spaced practice***	See recommendation 1 in *(15)* for a review. Some exemplary studies include *(72-74)*. Note that despite the overall evidence for spaced practice enhancing long-term retention, more massed practice is warranted early in learning of particular knowledge components *(75)*.
Study (recognition practice)	***Tests (retrieval practice)***	See recommendation 5 in *(15)*. Some exemplary studies include *(43, 76, 77)*. Although the consensus of the literature on this so-called "testing effect" is in favor of testing (asking students to answer questions) over study (giving students example answers to questions), this dimension is essentially the same as the next one *(21)*, but makes the opposing recommendation (i.e., substantial example study is beneficial).
Examples to study	Problems to solve	See recommendation 2 in *(15)*. Some exemplary studies include *(32, 44, 45)*. This dimension is essentially the same the testing effect above, but the consensus of the literature on the so-called "worked example effect" is in the opposite direction. A content-based resolution of this contradiction has been proposed *(21)*. Note research indicating a "reversal" (all problem solving practice is

		better) for learners who have become to develop expertise *(30)*.
Direct instruction	Discovery learning	See *(9)* and *(78)*. But note, some researchers argue for discovery learning *(79)* or an intermediate level of "guided discovery" *(80)*.
Re-explain	***Prompt for self-explanation***	See recommendation 7 in *(15)*. Some exemplary studies include *(81-83)*. Note, however, indicating a disadvantage of self-explanation for grammar learning *(17)*.
Immediate feedback	***Delayed feedback***	Some researchers have produced credible experimental evidence for learning benefits of *immediate* feedback *(11, 84)* whereas others have done the same for *delayed* feedback *(12)*.
Concrete materials	***Abstract materials***	Controversy exists with data supporting the use of concrete materials *(13)* vs. abstract materials *(14)* to promote transfer of learning. See also *(33, 85)*.