Instructional Alignment: Searching for a Magic Bullet
Author(s): S. Alan Cohen
Reviewed work(s):
Source: *Educational Researcher*, Vol. 16, No. 8 (Nov., 1987), pp. 16-20
Published by: American Educational Research Association
Stable URL: http://www.jstor.org/stable/1175370
Accessed: 13/03/2012 05:52

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Educational Research Association is collaborating with JSTOR to digitize, preserve and extend access to *Educational Researcher*.

http://www.jstor.org

Instructional Alignment:

Searching for a Magic Bullet

S. ALAN COHEN

When critical features of instructional stimuli match those of assessment, effect sizes routinely *reach 1.2 to 3 sigma. An instructional psychologist recasts this classic problem of* stimulus control *as* instructional alignment. *This paper describes results of alignment studies that have dramatic implications for researchers and practitioners. One implication embraces the obvious validity of teaching to the test, but poses* what is worth testing *as instructional design's most awesome challenge.*

In the latest search for educational excellence, perhaps it's time to reconsider an old idea—instructional alignment. *Instructional alignment* describes the extent to which stimulus conditions match among three instructional components: intended outcomes, instructional processes, and instructional assessment (Cohen, 1984a).

The purpose of this paper is to present a new perspective of this old idea by (a) reviewing its history; (b) presenting new data demonstrating that instructional alignment generates larger effects in research and practice for less "cost" than other instructional constructs; and (c) positing implications for both school practitioners and researchers.

Historical Perspective

Carroll's claim that a fundamental component of effective instruction is the degree to which learners have a clear picture of the instructional outcome (Carroll, 1963) was consistent with the times. Those times were the early 1960s, when Skinner's ideas had generated intense interest in task analysis and behavioral objectives in instructional design.

In the applied arena, Skinner's influence on instructional design was best demonstrated in a system called *CRI* or *Criterion Referenced Instruction* (Mager & Pipe, 1974). CRI applied Pipe's servomechanism model (Pipe & Betz, 1971) in which part of output is fed back as input to modify process. According to Pipe, any instructional system must derive from a clear statement of outcome; instruction generates that outcome as demonstrated in a final assessment. The assessment results adjust the outcome, the process, or both until they equal the intended outcome.

CRI was designed to train teachers and other course designers. But it quickly became apparent that school practitioners would not abide the Pipe model. For example, CRI presents the identical task to be learned in both the instructional process as well as in the final assessment, an ideal way to insure the precise match among what is taught, what is measured, and what is intended to be learned. The effect is near perfect learning, with variation in learning rate rather than in "amount" of learning, as expressed in Carroll's model of school learning.

Unfortunately, CRI contradicted the conventional expectation of a normal distribution of assessment results. That distribution requires either poorly taught content, or assessments whose stimulus conditions differ from those taught in the instructional phase. Either option guarantees assessment score variance. CRI practically guaranteed competence, which eliminated or reduced variance, contradicting that conventional expectation.

Although talk of "criterion testing" echoed through the 1960s and 70s, the standard psychometric model nevertheless predominated, as it does today. That model requires variance for a test to demonstrate reliability and validity. Thus, a combination of psychometric necessity and a tradition of "not all shall pass through these gates" doomed CRI to economic failure in the conventional teacher training market. Mager wisely turned away from the schools to industry, government, and business, where it is routinely expected that instruction generates reduced variance. In that setting, CRI continues to flourish a quarter of a century later (Mager & Pipe, 1983).

Meanwhile, in the research arena, new instructional design models had begun to emerge from the Skinnerian bias. For example, as programmed instruction became the cutting edge of instructional psychology, Gilbert (1962) proposed that an efficient way to design effective instruction was to begin at the end. By first developing the final "frame" representing the program's criterion behavior, and working backward to the beginning of instruction, one was more certain that the intended outcome would occur. Although the term *alignment* was not used, Gilbert and his contemporaries recognized the critical role of defining criterion behaviors in terms of stimulus conditions, and that varying those stimulus conditions during instruction could be expected to cause variations from the intended outcome.

S. ALAN COHEN *is a Professor and heads the Doctoral Program in Evaluation, Research & Instructional Design, University of San Francisco, Lone Mountain Campus —Rossi Wing, San Francisco, CA 94117-1080. He specializes in research design and instructional psychology.*

By the mid 1970s, naive elements of programmed instruction had begun to disappear from the schoolbook market. However, its basic principles, especially task analysis, had become the norm for instructional design. For example, Resnick, Wang, and Kaplan (1973) published their classic task analysis of school mathematics learning. By the middle 70s, *task analysis* was a fixture of instructional design (Resnick, 1976; Resnick & Beck, 1976). Task analysts focused on two elements, the stimulus conditions of criterion behaviors, and instructional sequence. Instructional alignment applies these elements.

Although CRI failed to infiltrate the practical arena of public schools, the results of other behaviorists' task analyses caught the rising tide of federal funds targeting the disadvantaged (e.g., Cohen & Hyman, 1977; Cohen & Kaplan, 1975; Cohen & Mueser, 1972; Engelmann, 1970). Despite intense opposition by conventional educators, some American teachers got their first close look at published programs exemplifying the instructional alignment principle. However, their use was usually limited to compensatory and remedial education. These systems rarely became the school's basic programs, and as federal aid declined in the 1980s, such programs were seen less and less in the classrooms.

Thus, the term *instructional alignment* represents a well-established phenomenon in the history of instructional design. Conventional wisdom accepts the logic that effective instruction demands congruence between stimulus conditions of instruction and stimulus conditions of the criterion assessment. The assumption is that the criterion assessment is clearly the intended outcome.

Instructional Alignment Effects

We first spotted the potential of this conventional wisdom as a researchable construct while training doctoral students to routinely test their research hypotheses by predicting critical effect sizes (Cohen & Hyman, 1979, 1981). In a doctoral study of format factors of math word problems that cause difficulty, Cohen & Stover (1981) taught middle graders three types of manipulations to increase their success rates. After three 45-minute lessons, posttest differences exceeded 3.4, 2, and 1.5 sigma. The critical effect size considered educationally significant had been defined as .70 sigma. A statistically significant effect for the number of observations in this study was approximately .50 sigma. What struck us was the magnitude of the effect relative to the minimal instructional effort.

About this same time, evidence was piling up showing large effects in favor of mastery learning programs around the world (Block & Burns, 1976; Hymel, 1982). What struck us was not simply the validation of Bloom's claims about learning for mastery (Bloom, 1976), but the magnitudes of the effects.

We decided to seek a magic bullet—the most potent variable among many underlying mastery learning that contributed most to these observed effect sizes. We hypothesized that whatever its identity, it was also present in the Cohen and Stover study, in which the intervention was not intended to be an example of mastery learning. Although it is true that mastery learning tended to generate effects greater than one sigma, large effects were also common to other approaches to instruction such as tutoring (Bloom, 1984). We looked for a common thread across mastery learning, well-designed instructional experiments, and tutoring.

We noted that a critical feature of mastery learning is the creation of unit tests *before* designing the instructional program (Block, 1971, 1974; Block & Anderson, 1975). We suspected that such an outcome-driven instructional design would generate more aligned instruction than traditional approaches.

We noted that an instructional experiment done as a doctoral dissertation (as in the case of the Cohen-Stover study) would have had to survive close scrutiny by a faculty committee of instructional psychologists. The researcher would have had to satisfy the established criterion of internal validity known as *construct validity of the dependent variable* (Cook & Campbell, 1979). We suspected that dissertation review committees would be particularly sensitive to the necessary match between an experimental intervention and the measure of effect.

Finally, we noted that tutorials are generally efficient pedagogies. Time is rarely spent on classroom rituals; the outcome is defined and the tutor gets right to the task. In short, we thought instructional alignment was a common thread woven into the fabrics of all three phenomena.

We were aware of the *curriculum alignment* literature (Levine, 1982; Niedermeyer, 1979; Neidermeyer & Yelon, 1981) focusing on aligning curriculum to objectives. However, we thought our magic bullet involved a finer tuning implied in task analysis. So, we called our construct *instructional alignment* and began our studies.

Instead of studying the obvious, which had already been established in the literature on instructional "congruence" (Baddeley, 1982; Tulving & Thompson, 1973), we focused on the *degree of effect* relative to instructional effort and such other issues as: (a) the critical features of stimulus conditions that maximize alignment effects; and (b) the alignment effect compared to aptitude effect. Traditional instruction generates .25 to .50 sigma effects. Is the alignment effect as large as it looks—approximately four times this norm?[1]

New Studies in Instructional Alignment

The Koczor Study. Koczor (1984) delivered six typical fourth-grade lessons, one per day, to 25 high achievers. Each 45-minute lesson had no instructional or cognitive relationship to the other; the purpose of the six lessons was to test the alignment effect with as many different fourth-grade skills as feasible within practical limits of a single study.

Immediately after each lesson, students received a posttest, the varying formats of which represented "degree of alignment." For example, one lesson used a paired associates technique that taught how to write Arabic numerals for designated Roman numerals. In the instruction, the Arabic was always presented or written *after* the Roman numerals. One group's posttest was aligned on this factor. In contrast, the misaligned treatment group received a test in which the Arabic numeral came first, and the student had to write the Roman numeral. Most teachers would consider this a *minor* variation of the instruction's stimulus conditions. That *minor* misalignment accounted for a 40% difference in posttest raw scores. Effect sizes representing differences between aligned and misaligned conditions for the lower and average aptitude students were as high as 1.10 and 2.74 sigma.

It is important to note that these "lower" aptitude fourth graders had a mean reading aptitude test score of 4.4 grade level. The so-called "higher" aptitude group had mean aptitude scores of grade level 8.6 ($s = 1.3$). Having come to expect large effects among lower achievers, such large effects observed in very high achievers surprised us.

The Tallarico Study. Tallarico (1984) used instructional alignment to investigate testwiseness effects. With norm referenced standardized tests (NRSTs) of reading achievement, testwiseness training tries to eliminate nonreading factors that control significant amounts of test score variance. To apply the alignment construct to testwiseness instruction requires teasing out critical features of those stimuli that most contribute to this extraneous variance, and then teaching all students to cope with them. If we reduce variance caused by these irrelevances, then we increase test validity; that is, students' scores are more nearly an estimate of true reading performance because extraneous sources of variance have been reduced.

To test the effects of two extraneous variance sources revealed in a task analysis of reading NRSTs (Cohen, 1977), Tallarico randomly divided second graders into three groups. One extraneous stimulus condition, *intent consideration*, required students to choose the best correct answer when two are reasonably correct (Schuller, 1979). The first group learned intent consideration. A second group learned to preread the item stem as a comprehension cue. Both groups learned these strategies under stimulus conditions and on pages simulating NRST conditions. A third group received a placebo, equal in time and in every other respect to the two experimental groups, except lacking testwise instruction.

A three-treatment-by-two-aptitude-level ANOVA indicated that almost 15% of the total sum of squares was explained by intent consideration and stem-cue skill, over and above the reading demand.

Now consider two facts: (a) Each treatment in the Tallarico study consisted of only two 30-minute lessons, a 10-minute demonstration followed by 20 minutes of seatwork drill; and (b) most educators are aware of the learning rate differences between high- and low-aptitude students. This treatment effect exceeded half that aptitude effect in the middle and lower middle class children used in this study.

For lower achievers, the stem-cue strategy group's average score exceeded the 85th percentile of the placebo group. The intent consideration treatment caused a 1.3 sigma effect.

The Fahey Study. Ability of instruction to overcome initial aptitude differences was one goal in a study of alignment effect relative to task difficulty. Using a 3×2×3 mixed ANOVA, Fahey (1986) analyzed interactions among the effects of directed practice under three different stimulus conditions for *understanding main idea*; two levels of aptitude and three levels of alignment (test item formats: aligned with instruction, misaligned #1, and misaligned #2). The first two factors were between-group analyses; alignment effect was a repeated measures.

Community college students were stratified by aptitude and then randomly assigned to one of the three directed practice levels. The research question was not would there be a difference among three types of directed practice, but *how much* of a difference relative to alignment.

Three important findings emerged. First, alignment effect was not observed between one pair of treatment levels which were the "easy" tasks (selecting main idea statements and titles from multiple choices). These lower level demands were easily within the students' learned repertoires. But when the task difficulty increased (producing in writing one's own statement of that main idea), so did the alignment effect.

Second, as anticipated, lower aptitude students did not perform as well as higher aptitude students when test items misaligned with the type of directed practice. As we found in the Koczor and Tallarico studies, alignment is more important to lower than to higher aptitude students.

A third finding was most significant to us. On the more difficult task, alignment was so effective that lower aptitude students performed better under aligned conditions than did higher aptitude students under misaligned. It is important to note that what we structured as "misaligned" is what one normally sees in the average classroom. The observed effect size was 1.2 sigma. With only 1.5 hours of instruction, alignment made enough of a difference to eliminate the expected aptitude gap.

Fahey demonstrated that lower aptitude students can successfully perform higher cognitive tasks when we align instruction. What usually passes for normal instruction in which the stimulus conditions of teaching and testing are slightly misaligned but certainly involve the "same skill" (as it is popularly perceived) can have a deleterious effect on lower achieving students. For low achievers, a little alignment goes a long way.

The Elia Study. The degree of alignment effect was dramatically demonstrated in a fourth study of 45 low socioeconomic level, urban, low achieving fourth graders. Elia (1986) taught meanings of 24 low frequency target words under three contrasting stimulus conditions: phrases, sentences, and paragraphs. In this repeated measures design, each subject learned eight words plus four word variants (e.g., *exist, existing*) under each contrasting condition, one condition per day over three days, in a counterbalanced treatment delivery. The day after each instructional segment, one third of the students was tested with words and variants systematically varied over the three stimulus conditions. Thus, one third of the items generated an aligned condition score, and each remaining third generated scores for misaligned stimulus conditions. In addition, some words aligned with instruction, and some were variants, representing another dimension of misalignment.

A 3×3×2 mixed ANOVA tested individual and interactive effects of two types of alignment. The first three-level factor represented the three contexts or conditions under which the student was taught, words in phrases, or sentences, or paragraphs. The second three-level factor represented the test item formats, words tested in phrases, sentences, and paragraphs. The third two-level factor represented either the word taught or its variant. Thus, some kind of transfer could be demanded via the condition, or the use of a variant, or both.

Overall, Elia reported an alignment effect of .91 sigma. In the phrase condition, alignment effect reached 1.76 sigma. Alignment/misalignment accounted for 16% of the total variance, and under the phrase condition, alignment explained 23% of the total variance.

Discussion and Conclusions

So far, our work with instructional alignment has led to three conclusions:

1. Instructional alignment routinely causes the 4-to-1 Effect, effect sizes ex-

ceeding one and often two sigma, about four times what we ordinarily see in typical classrooms. We routinely observe these large effects from small amounts of instructional effort.

2. *What* to teach is a more difficult question to answer than *how* to teach, considering the fine-tuning demands of task analysis.

3. Lack of excellence in American schools is not caused by ineffective teaching, but mostly by misaligning what teachers teach, what they intend to teach, and what they assess as having been taught. We have extended these conclusions to the bold statement that, in general, most teachers are effective, but usually at the wrong things.

What may these conclusions mean for practitioners? The idea that formal instruction should test what it teaches or teach what it tests is axiomatic. In general it is not being done for four reasons.

First, the level of fine tuning required for instructional alignment is beyond the current repertoire of most teachers, not because they cannot learn the skill, but because it is neither demanded of them nor taught in teacher training.

Second, teaching and assessing have been institutionally dichotomized. Instead of being an integral part of instruction, assessment is separated institutionally as well as in practice. For example, school districts and state education departments maintain separate departments for each domain. As a result, the content of commercially published NRSTs or locally mandated criterion tests usually differ in stimulus conditions from what teachers teach in the classroom. Current tests hide behind a ''pseudo alignment'' facade by claiming to measure the same ''skills'' as those taught in the classroom. But an enormous difference exists between what most educators call a skill or an outcome, and the kind of precision implied in the performance of instructional alignment.

Third, the expectation that instruction causes a normal distribution of ability is apparently rooted in a belief in the inevitability of cognitive inequality of human beings. This belief is so all-pervading and insidious, that most teachers and administrators I talk with honestly believe that to teach what we test and test what we teach is unethical because it denies a law of nature! Apparently, to make everyone masters of calculus or appreciators of literature would be a great lie.

Fourth, educators try to avoid responsiblity for what they teach. It is safe to be for teaching ''literary appreciation,'' or ''higher cognitive skills,'' or ''aesthetic appreciation of art.'' However, it is dangerous to define these outcomes by behavioral indicators or with formal assessments making them amenable to instructional alignment. In fact, the popular view is that these fuzzies are beyond precise definition—a convenient strategy to avoid admitting to ourselves what we really mean by such lofty sounding instructional outcomes. Perhaps if practitioners realized the potency of ordinary teachers as manifest in the large effect sizes resulting from aligned instruction, they might dare to be accountable for these outcomes.

Teaching what we assess, or assessing what we teach seems embarrassingly obvious. The fundamental issue is: *What's worth teaching?* This is the same question as: *What's worth assessing?* We can either know what we're doing, or not know what we're doing, but in either case, we'll be doing something to other people's children. Do we not have an ethical obligation to know what we're up to?

The implications for researchers are equally important. Before stumping the country to promote constructs dear to our research hearts, we should consider the effect size we can expect our constructs to cause when put in practice. Presently, we find no other construct that consistently generates such large effects, which is probably why the idea of instructional alignment is so well-entrenched in the conventional wisdom of instructional designers, even if not in the programs currently found in most classrooms.

Are we saying that our alignment research is more important than what other researchers are into?

Certainly not. The purpose of scientific research is to explain phenomena. A small statistically significant effect helps us understand phenomena. Such effects support theoretical models. What we suggest is caution in disseminating information about these results to practitioners who do not appreciate the difference between *significant* effect sizes and statistically *significant* findings. As a result of this lack of appreciation, the obvious conventional wisdom of alignment gets drowned out by the cacophony of information about brain research, learning styles, and so forth, all of which are important to our sciences, but none of which may generate large effect sizes as efficiently as instructional alignment.

Notes

[1] We invented the construct ''4-to-1 Effect'' to represent this concept (see Cohen, 1984a, 1984b).

References

Baddeley, A.D. (1982). Domains of recollection. *Psychological Review, 89,* 708-729.

Block, J.H. (1971). *Mastery learning: Theory and practice.* New York: Holt, Rinehart and Winston.

Block, J.H. (1974). *Schools, society and mastery learning.* New York: Holt, Rinehart and Winston.

Block, J.H., & Anderson, L.W. (1975). *Mastery learning in classroom instruction.* New York: Macmillan.

Block, J.H., & Burns, R. (1976). Mastery learning. In L.S. Shulman (Ed.), *Review of research in education* (chap. 1). Itasca, IL: Peacock Publishers and the American Educational Research Association.

Bloom, B.S. (1976). *Human characteristics and school learning.* New York; McGraw-Hill.

Bloom, B.S. (1984). The search for methods of group instruction as effective as one-to-one tutoring. *Educational Leadership, 41,* 4-17.

Carroll, J.B. (1963). A model of school learning. *Teachers College Record, 64,* 723-733.

Cohen, S.A. (1977). Tests are not all that bad (they're worse!). *Reading World, 16,* 219-222.

Cohen, S.A. (1984a). Implications of instructional psychological research on mastery learning. *Outcomes, A Quarterly Newsletter of the Network of Outcome-Based Schools, 2,* 18-25.

Cohen, S.A. (1984b). June, '84—a researcher's end-of-year reflections. *Outcomes, A Quarterly Newsletter of the Network of Outcome-Based Schools. 4,* 7-11.

Cohen, S.A., & Hyman, J.S. (1977). *The reading house series from Random House.* New York: Random House.

Cohen, S.A., & Hyman, J.S. (1979). How come so many hypotheses in educational research are supported? A modest proposal. *Educational Researcher, 8,* 104-109.

Cohen, S.A., & Hyman, J.S. (1981, April). *Testing research hypotheses with critical EFFECT SIZE instead of statistical significance in educational research.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Cohen, S.A., & Kaplan, J. (1975). *High intensity learning systems—math.* New York: Random House.

Cohen, S.A., & Mueser, A.M. (1972). *High intensity learning systems—reading.* New York: Random House.

Cohen, S.A., & Stover, G. (1981). Effects of teaching sixth grade students to modify variables of math word problems. *Reading Research Quarterly, 16,* 175-200.

Cook, T.D., & Campbell, D.T. (1979). *Quasi experimental designs in field settings.* Boston: Houghton Mifflin.

Elia, J.S.I. (1986). *An alignment experiment in vocabulary instruction: Varying instructional practice and test item formats to measure transfer with low SES fourth graders.* Unpublished doctoral dissertation, University of San Francisco.

Engelmann, S. (1970). *Distar reading I, II, III: An instructional system.* Chicago: Science Research Associates.

Fahey, P.A. (1986). *Learning transfer in main ideas instruction: Effects of instructional alignment and aptitude on main idea test scores.* Unpublished doctoral dissertation, University of San Francisco.

Gilbert, T.F. (1962). Mathetics: The technology of education. *Journal of Mathetics, 1,* 7-73.

Hymel, G. (1982). *Mastery learning: A comprehensive bibliography.* New Orleans: Loyola University Center of Educational Improvement.

Koczor, M.L. (1984). *Effects of varying degrees of instructional alignment in posttreatment tests on mastery learning tasks of fourth grade children.* Doctoral dissertation, University of San Francisco.

Levine, D. (1982). Successful approaches for improving academic achievement in inner-city schools. *Phi Delta Kappan, 63,* 523-526.

Mager, R.F., & Pipe, F. (1974). *Criterion-referenced instruction.* Los Altos, CA: Mager.

Mager, R.F., & Pipe, F. (1983). *Criterion-referenced instruction.* Atlanta, GA: Center for Effective Performance.

Niedermeyer, F.C. (1979).. *Curriculum alignment—A way to make schooling more understandable* (SWRL Professional Paper no. 41). Los Almitos, CA: SWRL Educational Research and Development.

Niedermeyer, F.C., & Yelon, S. (1981). L.A. aligns instruction with essential skills. *Educational Leadership, 38,* 618-620.

Pipe, P., & Betz, P. (1971, January). Approximation in programmed self-instruction for dentists. *Journal of Dental Education.*

Resnick, L.B. (1976). Task analysis in instructional design: Some cases from mathematics. In D. Klahr (Ed.), *Cognition and instruction.* Hillsdale, NJ: Erlbaum.

Resnick, L.B., & Beck, I.L. (1976). Designing instruction in reading: Interaction of theory and practice. In J.T. Guthrie (Ed.), *Aspects of reading acquisition.* Baltimore: Johns Hopkins Press.

Resnick, L.B., Wang, M.C., & Kaplan, J. (1973). Task analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum. *Journal of Applied Behavioral Analysis, 6,* 679-710.

Schuller, S.M. (1979). *A large scale assessment of an instructional program to improve testwiseness in elementary school children.* New York: Educational Solutions.

Tallarico, I. (1984). *Effects of ecological factors on elementary school student performance on norm referenced standardized tests: Nonreading behaviors.* Doctoral dissertation, University of San Francisco.

Tulving, E., & Thompson, D.M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352-373.

AERA

Winter Training Institute
January 18–19, 1988
see p. 31

DELTA AND
UNITED AIRLINES
NOW OFFER
AERA MEMBERS
SPECIAL NEGOTIATED
CONVENTION AIR FARES

For the 1988
Annual Meeting
New Orleans
April 5–9

DELTA AIRLINES

40–75% saving on fare to New Orleans. Call 1-800-241-6760 and refer to file no. D0084. Delta is also holding a contest drawing, with a prize of two free round-trip coach tickets, for persons who make reservations through the Delta toll-free number.

UNITED AIRLINES

5% off any fare for which you qualify (40–75% off normal coach fares) OR a minimum of 40% off fares that have no advance purchase or minimum stay required. Call 1-800-521-4041 and refer to AERA account no. 8016-C.

★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★

If you purchase your ticket from a travel agent, have him or her place your reservations through the airline's toll-free number to obtain the special AERA fare.