THE LIKELIHOOD OF INSTRUCTIONALLY BENEFICIAL, TRIVIAL, OR NEGATIVE RESULTS FOR KINDERGARTEN AND FIRST GRADE LEARNERS WHO COMPLETE AT LEAST HALF OF HEADSPROUT® EARLY READING

Janet S. Twyman
University of Massachusetts Medical School/Shriver Center

T. V. Joe Layng & Zachary R. Layng
Headsprout

ABSTRACT: Both kindergarten and first grade students using an internet-delivered reading program showed instructionally beneficial results on standardized test measures when compared to control groups in an unbiased control study. When kindergarten and first grade students completed at least 41 lessons (M = 67) of the 80-lesson Headsprout® Early Reading program, they showed a greater improvement on reading measures than students not receiving the Headsprout® program. Hopkins’ (2007) statistical evaluation method was used to provide a basis for understanding the likelihood of obtaining an instructionally beneficial, trivial, or negative result. The probability that the partial to full completion of Headsprout® Early Reading would produce an instructionally beneficial effect ranged from 2 to 1 (likely to be beneficial) as measured by ITBS subtest scores, to 1735 to 1 (almost certainly beneficial) as measured by Woodcock–Johnson III subtest scores.

Key words: Headsprout® Early Reading, practical significance, instructional effects, Hopkins confidence interval, early reading

With the current emphasis on evidence-based reading instruction, it has become essential that creators and vendors of educational programs submit their programs to rigorous evaluation, most often using group or statistical comparisons. In the case of research employing comparison group designs, statistical significance is conventionally used to ascertain whether one group, typically the group receiving the program, performed better than a control that did not receive the program, or received a different program. Whereas statisticians and educational researchers may be content with using statements of statistical significance, educators faced with

AUTHORS’ NOTE: The authors thank many individuals at Headsprout® for their participation in the implementation of this project and for comments on an earlier version of the paper. They also thank Marta Leon for her editorial assistance in the preparation of this manuscript. Correspondence concerning this article should be addressed to: Dr. Janet S. Twyman, UMass Medical School/Shriver Center, 333 South Street #13E806, Shrewsbury, MA 01545; Email: Janet.twyman@umassmed.edu
classroom decisions may be less so. This is quite understandable. On the one hand, research investigators are interested primarily in whether the average score of the treatment group will be higher than the average score of the control group, and whether such a difference is greater than chance alone would have allowed. On the other hand, educators typically are faced with the question, “Will this program be a sound instructional investment for my students?” The information that students, on average, perform better after using a program than they might otherwise have, is often of little help. It doesn’t let them know for which students, under what conditions, this statement is true. Moreover, educators know that differences that are statistically significant may be instructionally meaningless, and that some students in the treatment group may have done no better than others in the control group and, further, a substantial number of students in the experimental group may have done much worse than others in the control group. Thus, for educators, there is an important practical difference between statistical significance and instructional significance.

In the interest of real-world instructional effectiveness questions faced by educators, we shall consider the concept of instructional, or practical, significance. Recently, Batterham and Hopkins (2006) and Hopkins (2002, 2007) have suggested that the question of “practical significance” (clinical or instructional) be taken seriously by those whose primary concern is determining the real-world effectiveness of a particular treatment. Batterham and Hopkins (2005) argued that statistical significance, as indicated by p-value, fails to provide “information about the direction or size of the effect or, given sampling variability, the range of feasible values” (p. 7). Criticism of the p-value’s use as a predictive tool has been raised due to its high variation over replication, nearly regardless of sample size, since knowing [an obtained] p value . . . “gives only the most hazy notion of what the next replication p will take” (Cumming, 2008, p. 293). As an alternative to p-value, Hopkins advocates the use of confidence intervals combined with qualifying, probabilistic information on magnitude of effect. Whether an effect is harmful, trivial, or beneficial is determined by establishing the threshold for smallest important effect (Batterham & Hopkins, 2006). Unlike the p statistic, however, there is designedly no convention for determining the important effect (e.g., alpha = .05). Instead, a gain amount that would make a particular investment in one program more instructionally beneficial than another is decided upon prior to determining effectiveness. Gains at or greater than that amount have practical significance; lesser or no gains (or losses) would have trivial or even harmful effects.

Paradoxically, Hopkins (2002, 2007) noted that programs that may show statistically significant results may, using his methods, be considered instructionally trivial, whereas programs shown not to produce statistically significant results may, in fact, have a relatively good chance of being instructionally helpful (for criticisms of the use of statistical significance see Batterham & Hopkins, 2006; Cohen, 1994; Cumming, 2008; Hopkins, 2007; Kirk, 1996; Murphy & Myers, 2004; Rigby, 1999).

To illustrate a Hopkins test of instructional effectiveness, consider an example from the present study. A gain of a two-month grade level improvement over typical curriculum, after seven months of program use, was identified to indicate that the
program would be instructionally beneficial. That is, a true mean difference $\mu_{\text{diff,ITBS}} > 2$ indicates a beneficial effect as measured by ITBS scores, $-2 < \mu_{\text{diff,ITBS}} < 2$ (i.e., the value of the group mean difference falls between $-2$ and $2$) is said to be trivial and $\mu_{\text{diff,ITBS}} < -2$ indicates a negative effect, and for the Woodcock Johnson $\mu_{\text{diff,WCJ}} > 2$ indicates a beneficial effect, $-2 < \mu_{\text{diff,WCJ}} < 2$ is said to be trivial and $\mu_{\text{diff,WCJ}} < -2$ indicates a negative effect. In addition, we indicate the likelihood that the true population mean lies within each region of effect, using percentages and qualitative terms such as “unlikely,” “almost certainly,” and so on, that make it easier for education consumers to understand.

In this way, reasoned decisions can be made concerning such issues as whether or not scarce resources should be allocated for a particular program or set of practices, and whether or not program deployment would be professionally ethical in the first place. Such an empirical analysis is always important, but perhaps even more so as fiscal resources remain scarce in many educational settings throughout the country and the world. To illustrate this perspective, instructional outcomes in this report are expressed in terms of the calculated odds that the Headsprout® Early Reading program (hereafter referred to as HER) will have instructionally beneficial, trivial, or negative effects.

HER is an internet-based reading program that is designed to teach the essential skills and strategies required for reading success. The program’s formative evaluation (see Layng, Stikeleather, & Twyman, 2006) included testing with children who produced over 250 million data points that were collected and analyzed in order to continually research, develop, test, and refine the program through the application of a nonlinear, behavior-analytic design process (see Twyman, Layng, Stikeleather, & Hobbins, 2004). Children proceed through lessons designed to teach essential reading skills through multiple interactions with cartoon-based episodes set in the environs of Space World, Dinosaur World, Undersea World, and Jungle World. The program’s infrastructure uses patented technology (Layng et al., 2003) to teach the phonemic awareness, phonics, fluency, vocabulary, and comprehension skills and strategies necessary to decode words and read with understanding. The program automatically and continuously tracks each learner’s performance and adjusts the instruction based on analysis of a learner’s individual responses, including patterns of errors and correct responses. Table 1 summarizes the nine teaching routines that form the core of the program. A more detailed description of the procedures for teaching reading skills and strategies can be found in Layng, Twyman, and Stikeleather (2003, 2004a, 2004b).

HER is designed to provide its greatest effect for those students who complete all 80 lessons and who read all 80 print stories (Layng, Twyman, & Stikeleather, 2004b). However, although many schools may set out to complete the entire 80-lesson program, for a variety of reasons all lessons may not be completed by all students (Clayton et al., 2008; Haas et al., 2008; Webb et al., 2008). This study investigated whether HER can produce beneficial instructional outcomes in reading performance when fewer than 80 lessons are completed.
Table 1. Nine Teaching Routines Incorporated into *Headsprout Early Reading*

<table>
<thead>
<tr>
<th>Routine Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Establishing routines</td>
<td>Teach sound–letter correspondence and sight words through explicit instruction</td>
</tr>
<tr>
<td>Adduction routines</td>
<td>Teach skills through a discovery learning method</td>
</tr>
<tr>
<td>Vocal potentiation routines</td>
<td>Teach learners to speak out loud and become their own listeners</td>
</tr>
<tr>
<td>Blending and segmenting routines</td>
<td>Teach learners to blend sounds together into words and to segment words into their individual sounds</td>
</tr>
<tr>
<td>Sentence and story routines</td>
<td>Teach skills such as reading from left to right and reading for meaning</td>
</tr>
<tr>
<td>Fluency routines</td>
<td>Involve guided, timed reading practice</td>
</tr>
<tr>
<td>Motivation routines</td>
<td>Involve both extrinsic and intrinsic reward components</td>
</tr>
<tr>
<td>Application routines</td>
<td>Involve applying skills and strategies to new words, stories, and contexts</td>
</tr>
<tr>
<td>Overall sequencing</td>
<td>Designed to develop an interlocking set of skills and strategies that cumulative build upon previous components</td>
</tr>
</tbody>
</table>

**Method**

**Setting**

This investigation was conducted in a public elementary school located in a large, extremely diverse urban school district located in the northeastern United States. Approximately 800 pre-kindergarten through fifth grade students attended the school, with over 90% of the students qualifying for free or reduced-price lunch. Of the students attending the school, 74% were African American, 23% Hispanic or Latino, 2% American Indian or Asian, and 1% Caucasian. Typically, 70% of the fourth grade students in the school had not demonstrated reading proficiency.

The computer-based online instruction used in this investigation occurred in the school’s computer lab, fully equipped with modern computers (described below) made possible by a corporate donation. The computer lab housed the computers, server, teacher’s desk, bookshelves and cabinets, and two small tables for one-on-one or small group work. The lab was accessed during eight 40-minute periods
across the school day, from 7:50 a.m. to 3:20 p.m. In addition, the lab was used in the school’s “after school” and Saturday morning instructional activities.

**Materials**

Thirty internet-connected Apple® eMac computers equipped with 256 Mb RAM, Safari web browser with the Macromedia Flash plug-in, a mouse, the HER online program, and 80 printed stories were used in this study. The online program resided on Headsprout servers and was presented whenever the learner accessed the program via the internet. All learner responses were recorded and uploaded to the Headsprout servers. All computers were located in a single room.

**Participants**

Seventy kindergarten students (39 male and 31 female), and 55 first grade students (32 male and 23 female) participated in this study. Participants came from six kindergarten and four first grade classrooms. Although all 80 lessons of the program were available for the experimental group to complete, the average number of lessons actually completed was 67 (range 41–80). Only those participants who completed at least half the program (i.e., the first 41 lessons or more) are included for comparison for two reasons. First, the design of HER focuses on building skills and strategies in a cumulative and accretive, as opposed to a linear, way (Layng et al., 2004a). Second, formative evaluation of the program had shown that the first effects measurable by typical standardized tests occur after completing the first half of the program (Layng et al., 2004b).

**Procedures**

Existing kindergarten and first grade classrooms were assigned to either experimental or control groups. Assignment was unbiased, with a flip of a coin determining to which group classrooms were assigned. Three kindergarten and two first grade classrooms were assigned to the experimental group, and the same number to the control group. The experimental group thus included 35 kindergarten students (20 male and 15 female) and 30 first grade students (20 male and 10 female). The control group included 35 kindergarten students (19 male and 16 female) and 25 first grade students (12 male and 13 female).

All participants were pretested on the Woodcock Johnson III-R Letter-Word Identification subtest (Schrank, Mather, & Woodcock, 2006; hereafter WJ-LW). Kindergarten students were pretested using the prescribed subtests for beginning kindergarten from the Iowa Test of Basic Skills (Hoover et al., 2001; hereafter ITBS): Word Analysis, Reading Words, and Reading Comprehension. First grade students were pretested on the Word Analysis, Reading Words Part I, and Reading Comprehension Part II subtests of the ITBS standardized assessment as prescribed for beginning first graders.

At the end of the school year, both experimental and control groups were administered the WJ-LW and ITBS subtests as posttests. Kindergarten students were
tested on the Word Analysis, Reading Words, and Reading Comprehension subtests of the ITBS standardized assessment as prescribed for the end of kindergarten. First grade students were tested on the Word Analysis and Reading Comprehension Part I and Part II subtests of the ITBS standardized assessment as prescribed for the end of first grade.

An independent assessment team recruited from City University of New York administered all pretests and posttests. The assessment team did not know the participants or their teachers and had never previously visited the school. Throughout the experiment was blind as to which group the participants belonged (experimental or control).

All participants received 180 minutes a day of reading/language arts instruction. For participants in the experimental group, 40 minutes of the 180 were allocated to HER instruction in the computer lab. Each online HER lesson averaged 20 minutes in length. With transition and login time, the time away from the classroom averaged approximately 40 minutes. The control group students remained in their regular classroom for reading/language arts instruction and received other supplementary phonics and reading instruction during that time period. Once a program-prescribed number of HER lessons had been completed, a printed story based on what had been learned to date was available to be read by the participant, for a total of 80 printed stories. Printed HER stories were available continuously in the program. The actual number of stories read by the experimental group was not recorded, but interviews with the teachers indicated that participants read fewer than the recommended number of stories.

Results

A test for differences that existed prior to treatment was conducted to expose any unintended significant differences between groups even though we employed an unbiased assignment. These tests were not employed to evaluate post-treatment effects. A two-tailed t test revealed no significant pre-treatment differences (\( \bar{d} = 0.05 \)) for WJ-LW raw scores between students in the experimental group who completed at least 41 of the 80 lessons and students in the control group. The pre-treatment difference in raw scores between kindergarten students in the experimental (M = 11.3) and control (M = 9.5) groups was not significant, t (66) = 1.99, p > .05. The difference in raw scores between first grade students in the experimental (M = 23.9) and control (M = 19.5) groups was also not significant, t (54) = 2.00, p = 0.054.

Similarly, a two-tailed t test showed no significant pre-treatment differences (\( \bar{d} = 0.05 \)) for ITBS Word Analysis standard subscores between kindergarten students in the experimental group who completed at least 41 lessons (M = 112.3) and kindergarten students in the control group (M = 112.6), t (66) = 1.99, p = 0.823.

Two-tailed t tests were conducted comparing pre-treatment scores on the ITBS Word Analysis, Reading Comprehension, and Reading Words subtests for first grade students in the experimental group who completed at least 41 lessons and first grade students in the control group. There were significant differences in Word Analysis scores between experimental (M = 133.9) and control (M = 123.4) groups, t (57) =
2.00, p = 0.0008. There were significant differences in Reading Comprehension pre-treatment scores between experimental (M = 137.8) and control (M = 130.6) groups, t (53) = 2.00, p = 0.02. Finally, there were significant differences in Reading Words scores between experimental (M = 137.3) and control (M = 131.7) groups, t (57) = 2.00, p = 0.01.

Although post-treatment testing of first grade participants indicated substantial differences on the ITBS subtests between the experimental and control groups favoring the experimental group, these data will not be presented due to pre-existing differences in pre-treatment performance. Accordingly, only WJ-LW results are provided for first grade.

**Posttest Group Comparison**

As noted earlier, even if statistical significance is often used to provide evidence of instructional effectiveness, questions have been raised as to the appropriateness of these measures when instructional decisions about program effectiveness are being considered. Data on instructional effectiveness can also be considered in terms of the likelihood of producing an instructionally beneficial result as indicated by posttest assessments (Batterham & Hopkins, 2006; Hopkins, 2007). To review, a true mean difference $\mu_{\text{diff,ITBS}} > 2$ indicates a beneficial effect as measured by ITBS scores, $-2 < \mu_{\text{diff,ITBS}} < 2$ is said to be trivial and $\mu_{\text{diff,ITBS}} < -2$ indicates a negative effect, and for the Woodcock Johnson $\mu_{\text{diff,WCJ}} > 2$ indicates a beneficial effect, $-2 < \mu_{\text{diff,WCJ}} < 2$ is said to be trivial and $\mu_{\text{diff,WCJ}} < -2$ indicates a negative effect. A two-month difference was chosen based on the following considerations: In a well-known national study Borman et al. (2007) reported highly significant effects as measured by the Woodcock Johnson assessment of a two-month difference between control and experimental groups for ten months of full implementation. Since the current program duration lasted only seven months, a two-month difference in instructional outcome should be considered quite meaningful. Additionally, each group received extensive and intensive instruction for nearly three hours per day in reading and closely related language-arts skills. One would expect that an intervention that occurred on average less than 60 minutes per week would have little effect given so much intensive instruction. In that context, a difference of two months during a seven-month cycle was considered to be quite beneficial.

Tables 2 and 3 show the results of this study in terms of a 95% confidence interval with data calculated according to formulas provided by Hopkins (2007). For the WJ-LW subtest, the likelihood of an instructionally beneficial effect of at least a two-month difference in grade-level performance was 97% for kindergarten participants (“very likely”) and 100% for first grade students (“almost certain”). Though substantial gains for kindergarten students were not quite so certain on the three ITBS subtests, the likelihood of substantial gain was still 85% for Word Analysis, 65% for Reading Words, and 75% for Reading Comprehension.

Tables 4 and 5 show the probability (after Hopkins, 2007) that the true effect could be instructionally beneficial, trivial, or negative for all tests for both kindergarten and first grade students. The likelihood that HER would have a
beneficial effect for those students who completed at least the first 41 or more of 80 lessons ranged from a low of 2:1 to a high of 1735:1. The ITBS subtests produced the lowest likelihood of showing an instructionally beneficial effect, and the WJ-LW indicated the greatest likelihood of an instructionally beneficial effect. While even the lowest odds (2:1) are high that partially completing HER will be instructionally beneficial, it is also important to note that the odds of partially completing HER having a negative effect are extremely low (the odds of a negative effect range from 1:699 for kindergarten participants tested on the ITBS Reading Words subtest to 1:2,094,712 for first grade participants tested using the WJ-LW).

Table 2. Comparison of Kindergarten Experimental Group Participants Who Completed at Least 41 Lessons with Control Group Participants

<table>
<thead>
<tr>
<th>Compared Groups Kindergarten</th>
<th>Score Differences in Means &amp; 95% Confidence Limits</th>
<th>Likelihood (Percent; Qualitative) of Substantial Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Woodcock Johnson III–R: Letter–Word Identification subtest</td>
<td>4.60; ± 2.8*</td>
<td>97; very likely</td>
</tr>
<tr>
<td>Iowa Test of Basic Skills: Word Analysis</td>
<td>4.04; ± 3.9**</td>
<td>85; likely</td>
</tr>
<tr>
<td>Iowa Test of Basic Skills: Reading Words</td>
<td>2.53; ± 2.7**</td>
<td>65; possibly</td>
</tr>
<tr>
<td>Iowa Test of Basic Skills: Reading Comprehension Part I</td>
<td>3.12; ± 3.3**</td>
<td>75; likely</td>
</tr>
</tbody>
</table>

*Raw score differences
**Standard score differences
### Table 3. Comparison of First Grade Experimental Group Participants Who Completed at Least 41 Lessons with Control Group Participants

<table>
<thead>
<tr>
<th>Compared Groups</th>
<th>Score Differences in Means &amp; 95% Confidence Limits</th>
<th>Likelihood (Percent; Qualitative) of Substantial Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Grade</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woodcock Johnson III-R: Letter–Word Identification subtest</td>
<td>8.63; ± 3.9*</td>
<td>99.9; almost certain</td>
</tr>
</tbody>
</table>

*Raw score differences

### Table 4. Likelihood that the True Value of the Statistic for Kindergarten Learners is:

<table>
<thead>
<tr>
<th></th>
<th>Instructionally Beneficial</th>
<th>Instructionally Trivial</th>
<th>Instructionally Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>prob (%)</td>
<td>odds</td>
<td>prob (%)</td>
</tr>
<tr>
<td>Woodcock Johnson III–R: Letter–Word Identification</td>
<td>97</td>
<td>30:1</td>
<td>3</td>
</tr>
<tr>
<td>Iowa Test of Basic Skills: Word Analysis</td>
<td>85</td>
<td>6:1</td>
<td>15</td>
</tr>
<tr>
<td>Iowa Test of Basic Skills: Reading Words</td>
<td>65</td>
<td>2:1</td>
<td>35</td>
</tr>
<tr>
<td>Iowa Test of Basic Skills: Reading Comprehension</td>
<td>75</td>
<td>3:1</td>
<td>25</td>
</tr>
</tbody>
</table>
Table 5. Likelihood that the True Value of the Statistic for First Grade Learners is:

<table>
<thead>
<tr>
<th></th>
<th>Instructionally Beneficial</th>
<th>Instructionally Trivial</th>
<th>Instructionally Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>prob (%)</td>
<td>odds</td>
<td>prob (%)</td>
<td>odds</td>
</tr>
<tr>
<td>Woodcock Johnson III–R: Letter–Word Identification</td>
<td>99.9</td>
<td>1735:1</td>
<td>0</td>
</tr>
</tbody>
</table>

Discussion

The practical significance evaluation method (Hopkins, 2007) provides a basis for understanding the likelihood of obtaining instructionally beneficial, trivial, or negative results. Results analyzed in this way may help ensure that programmatic decisions will be made on the basis of the likelihood of a successful educational outcome rather than simply on the basis of a statistically significant result that might, nonetheless, be instructionally trivial, or a result that may not be statistically significant but that might be instructionally beneficial. Furthermore, it provides an indicator of the likelihood of a negative effect. If a negative effect is highly unlikely, then taking a chance on a beneficial effect may be potentially advantageous. Practical significance and confidence intervals are relevant in that they describe the extent to which study results are applicable in general—that is, beyond the participants involved in the study (Wolfe & Cumming, 2004). One of its greatest advantages is that it is also a method of reporting that can be easily understood by education consumers, educators, and families alike.

In the case of HER, the lowest odds that the partial completion of the program would produce an instructionally beneficial effect on a posttest measure were 2 to 1. Given that the greatest odds of a negative outcome were 1 in 699, the likelihood of a beneficial outcome far outweighs the likelihood of a negative outcome. In the case of the WJ-LW test, the likelihood of an instructionally beneficial outcome was more than “likely.” Indeed, it was “very likely” for kindergarten students and “almost certain” for first grade students, with the likelihood of a negative effect almost nonexistent.

Unbiased assignment was performed at the level of the classroom. That is, each classroom was assigned to either experimental or control conditions, but individual participants within each classroom were not. Nevertheless, such unbiased assignment, even at the level of the classroom, provides greater confidence that the results obtained in this study were a function of the treatment conditions. It is also
important to note that the evaluators who pretested and posttested the students were independent and blind to the conditions. They did not know the students, and perhaps most importantly, did not know from which group, experimental or control, the learners were drawn.

**General Discussion**

Students who progressed through the program and completed at least 41 lessons of HER (M = 67) showed an instructionally beneficial effect on every reading test measure as compared to those students who did not receive the HER program. This is particularly important, as the acquisition of skills in HER does not occur in a linear, equal interval fashion. Learners who use HER typically take longer (more lessons) to acquire new skills in early lessons and learn more skills across fewer lessons as they progress through the program. A learner accumulates new skills and accelerates learning while proceeding through the program, resulting in even greater reading performance in later lessons (Layng et. al., 2004a). Accordingly, the reading skills taught through lesson 67 are not nearly as complete or robust as those presented through lesson 80. These results suggest that students who receive fewer than the recommended 80 HER lessons may still show an “instructionally beneficial” effect, and further, they show that there is almost no likelihood of an “instructionally negative” effect.

Another striking aspect of these results is that both the experimental and control groups, in both kindergarten and first grade classrooms, received at least 180 minutes a day of reading and language instruction. This intensive instruction was part of a city-wide effort to improve reading outcomes. The HER time was carved out of the 180 minutes of instruction; thus, both groups received the same number of total minutes per day of reading instruction. The students began to use HER only late in the fall and experienced a long break from the program for the holidays. Thereafter, beginning in January, they averaged about four days of use per week, for about 20 or 30 minutes a day, and typically did not complete the program. Even so, the experimental group still outperformed the control group in standardized reading tests. These data strongly suggest that HER can improve student outcomes even when (1) control participants receive intensive reading instruction, and (2) experimental group participants do not complete the program. Whether these results are a function of an interaction between intensive core instruction and HER, or HER alone, is not clear from this study. Formative evaluation data, however, suggest that students using HER will show substantial gains under a variety of conditions (Layng et. al., 2004a, 2004b).
References


LIKELIHOOD OF INSTRUCTIONAL BENEFIT


