Read Writ (2013) 26:539–563
DOI 10.1007/s11145-012-9418-z

Four methods of identifying change in the context of a multiple component reading intervention for struggling middle school readers

Jan C. Frijters · Maureen W. Lovett · Rose A. Sevcik · Robin D. Morris

Published online: 9 October 2012

Abstract The results from controlled intervention research have indicated that effective reading interventions exist for children with reading difficulties. Effect sizes for older struggling readers, however, typically have not matched the large effects demonstrated with younger children. Standardized effect sizes for intervention/control comparisons obscure important individual differences within intervention and control groups—differences potentially relevant to the *who* and *why* of intervention success. The present study reports the outcomes of *PHAST Reading*, a research-based multiple component reading intervention. Participants were 270 Grade 6, 7, and 8 students reading significantly below age-level expectations, who participated in a year-long intensive small-group intervention. Four methods were applied to characterize individual change: (a) normalization relative to age-appropriate standards; (b) statistically-reliable pre–post change using the Jacobson–Truax index; (c) individually-estimated growth rates using hierarchical linear modeling; and (d) change to a fixed criterion across multiple measures. Each method was evaluated for its ability to identify intervention outcomes, replicate traditional group-based effect size metrics, and characterize individual differences across participants depending on whether change was demonstrated. Each method replicated traditional group-based effect sizes, with advantages in consistency and predictive power for the reliable change index and growth curve approaches.

J. C. Frijters (✉)
Departments of Child and Youth Studies and Psychology, Brock University, 500 Glenridge Avenue, St. Catharines, ON L2S 3A1, Canada
e-mail: jan.frijters@brocku.ca

M. W. Lovett
Department of Paediatrics, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada

R. A. Sevcik · R. D. Morris
Department of Psychology, Georgia State University, Atlanta, GA, USA

Keywords Reading intervention · Outcomes · Treatment evaluation · Reading disability · Middleschool children · Adolescents · Dyslexia

Introduction

There is converging evidence that basic reading skills can be substantially improved for younger readers with reading disability (Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Lovett et al., 2000; Morris et al., 2012; Olson, Wise, Ring, & Johnson, 1997; Torgesen, Wagner, Rashotte, Alexander, & Conway, 1997; Vellutino et al., 1996; Wise, Ring, Sessions, & Olson, 1997). Attention is turning to interventions for older struggling readers and current evidence indicates that these are associated with smaller effect sizes (Reed & Vaughn, 2010; Swanson, 1999; Edmonds et al., 2009). Standardized effect sizes for intervention-control comparisons obscure important individual differences within intervention and control groups—differences potentially relevant to the *who* and *why* of intervention outcomes. The main limitation of effect sizes is that they generalize to the intervention-control group difference, and provide no information about individuals within either group. With the smaller effect sizes associated with older struggling readers, addressing this limitation becomes more important for tuning of intervention techniques and optimizing type of intervention for particular struggling readers. Recent research has investigated cognitive and linguistic individual differences factors that moderate degree of remedial response (Al Otaiba & Fuchs, 2002; Fletcher, Stuebing, & Barth, 2011; Frijters et al., 2011; Nelson, Benner, & Gonzalez, 2003). One approach to answering these questions is to search for subsets *within* intervention conditions, identifying who changes and what characterizes members of that subset. To embark on this venture effectively, there is a need to understand the adequacy of different methods to identify changing individuals within a larger group participating in an intervention. There also is a need to understand performance of such metrics relative to traditional group-based measures of effect size. Finally, the relative performance of each method needs to be characterized in their productive capacity to formulate typologies of outcomes. Across all methods reviewed below, it is assumed that in order to claim that a struggling reader's proficiency has changed the following conditions must be met: (1) change must be demonstrated; (2) change must be positive; (3) change must be meaningful.

Methods to identify such subgroups have been driven by several broad frameworks. In one set of methods, a participant must achieve a post-intervention criterion that places them within the normative range for their age-group. This is represented by the normalization method considered below. In another, a participant must make a gain or change at a rate that represents a statistically reliable improvement, with no gain or change as the comparator. This is represented by the reliable change index and growth curve methods considered below. In yet another, a participant must demonstrate change in outcome scores that exceeds a fixed criterion, such as 1 SD for the sample or 1 SEM for the test. This is represented by the 'within-individual gains, replicated over tests' (WIGROT) method considered

below. The present study reviews these specific methods, with a focus on research design, data requirements, strengths, and limitations. This is followed by a report on the application of each method to outcome data from an intensive research-based intervention for middle-school readers with reading disability (RD).

A brief note on the demonstration of change versus response is warranted at this point, as it has a bearing on how results from each method are interpreted. By default, the normalization method establishes change relative to a normative peer group and therefore also demonstrates response. RCI and/or growth curve estimation establish change if raw or Rasch-scaled scores are used. There are two situations in which these methods can also establish whether response has occurred. First, RCI, growth curve estimation, and WIGROTs can indicate response if standard scores are used. Should positive (and statistically significant, as discussed below) change be demonstrated with standard scores, by definition such change is occurring at a rate that is faster than the similar-aged peers of the normative sample. Second, response can be determined if change is demonstrated with raw or Rasch-scaled scores, if it is also determined that such change is moderated and/or interacts with assignment to intervention versus some form of control group (e.g., alternate treatment, business as usual, or waitlist). In these cases, the interpretation of response is bounded by the quality and nature of the control group. The problems (e.g., the serious confounds between developmental change and other causes of change introduced by using standard scores) and potentials associated with these strategies have been extensively discussed elsewhere (see Yoder & Compton, 2004).

Normalization method

This method identifies participants as having changed, if following intervention they exceed a clinically-relevant criterion (for examples, see Fuchs, 2003; Torgesen et al., 2001). The specific criterion is most typically a standard score greater than or equal to 90 on a nationally-normed measure of reading (i.e., with achievement expected among normally developing children of a similar age set to a standard score of 100 and standard deviation of 15). The two primary advantages of this criterion are clinical relevance and simplicity of calculation. If a child falls within the range clinically recognized as 'Average', having begun intervention below that point, it seems clear that some degree of change has been demonstrated. The third criterion for determining change—that change be meaningful—is defined in this approach as the attainment of a clinically-relevant benchmark. In terms of simplicity, identification of outcome status requires only the calculation of post-intervention standard scores using published normative tables.

These advantages do obscure several notable limitations. First, there is ample evidence that the effect sizes associated with evidence-based intervention vary inversely with age (Swanson, 1999). Children in the early primary grades are much more likely to demonstrate gains approaching normalized levels in reading interventions than adolescents in middle-school (Edmonds et al., 2009), high school (Lovett, Lacerenza, De Palma, & Frijters, 2012) and especially struggling adult readers (Greenberg et al., 2011; Sabatini, Shore, Holtzman, & Scarborough, 2011), limiting usefulness for these populations by identifying very low base rates

of gainers. Second, identification is linked to the severity of initial reading deficits. One commonly-used inclusion criterion for ascertaining struggling readers is reading achievement 1 SD below age-expectations (i.e., <85), which would require a standard score gain of only five points for a participant to be identified as having improved on a specific outcome (i.e., from 85). This degree of change may be less than even the standard error of the reading test, a factor that the next method attempts to take into account.

Psychometric method

A method of identifying reliable gains variously known as the reliable change index (RCI) or the reliable change criterion has been widely used within the literature on evaluation of psychotherapeutic outcomes (McGlinchey, Atkins, & Jacobson, 2002). The formula for calculation of the RCI was originally proposed by Jacobson and Truax (Jacobson, Follette, & Revenstorf, 1984), and was followed by several proposed refinements or modifications that provided alternative estimates of standard error. Subsequent simulation and application studies have been variable in demonstrating advantages in accuracy or replicability of these modifications (Bauer, Lambert, & Nielsen, 2004; Temkin, 2004) and as a result the original formula with the 1986 modification (Christensen & Mendoza, 1986) is currently most widely used. Data requirements are pre- and post-intervention outcomes, along with the sample standard deviation and a high quality estimate of test reliability, preferably derived from an independent normative sample (Maasen, 2004).

Regardless of the formula variant used, the third criterion for identifying gainers—that change be meaningful—is satisfied by establishing whether change was statistically greater than would be expected according to the joint effect of the measure's reliability and sample-specific variability. Though the formula is not part of widely available statistical packages, it requires basic functions that can be executed on various online calculators (Evans, 1988) or manually. Interpretation of the criterion, although not as clinically intuitive as the normalization method, is relatively straightforward: significant change is the degree of gain necessary to exceed the unreliability of the outcome measure. The psychometric method can be used with common pre-post designs, with or without a control group. The reliance on only two assessment points also introduces weaknesses, which will be addressed below. This method has been used very infrequently within the literature on struggling readers, unlike both the normalization and individual growth curve methods. Due in part to the original RCI publication being in the clinical psychology literature, the few examples of its use in studies on reading have emerged from the child psychopathology area. In a rare example, Smart, Sanson, and Prior (1996) calculated reliable changes in subscale scores of the Rutter Child Behavior Questionnaire to characterize reading difficulties, behavior problems, and comorbid groups.

Individual growth curves

Estimating individual growth curves based on multiple repeated assessments within a multilevel model of change is a method that addresses the limitations introduced

by relying on only two outcome observations. Recent research has incorporated this technique into longitudinal research designs to characterize the developmental patterns of normal children and children with disabilities on reading and related tasks (Compton, 2003; Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996). The growth curve approach has also been productive in predicting the rate of reading development and in characterizing variation in growth rates (Aunola, Leskinen, Onatsu-Arvilommi, & Nurmi, 2002; Compton, 2000). This approach begins with individual examination of growth trajectories combined with the fitting of a model for growth, potentially incorporating both linear and curvilinear components, along with a model for the covariance structure of the repeated measurements (Singer & Willet, 2003). The approach uses an iterative framework to estimate individual change over time for each student, balancing the strengths of the between-student (i.e., Level-2, initial reading level), and within-student (i.e., Level-1, growth over hours of intervention) model components (Snijders & Bosker, 1999). Widely available statistical software can fit these models (e.g., PROC MIXED in SAS/STAT, SAS Institute, 2008; MIXED procedure in SPSS, IBM, 2010); along with specialized software packages (e.g., HLM6, Raudenbush, Bryk, & Congdon, 2007; Mplus, Muthén & Muthén, 1998–2010). Each of these packages can generate individual growth estimates, with standard errors and associated significance tests, once a growth model is fit.

Within the growth curve method, the third criterion for identifying change—that change be meaningful—is satisfied by a significance test of each individual's growth rate against zero growth. By implication of using growth *rates*, this approach requires not only the demonstration of positive change (e.g., as with the normalization and RCI approaches), but sustained growth over multiple observations. The primary growth curve advantage in reliability of gainer identification is also the source of its practical disadvantage, with greater reliability of the estimate available only with more costly research designs incorporating more than two outcome assessments. Similarly, the standard error of measurement must be low across all measurement points to generate reliable growth estimates (Bryk & Raudenbush, 1987), restricting growth curve estimation to measures with well-developed psychometric properties. The growth curve method may not require as much absolute change as the normalization method, but the outcome measure must be high quality and must generate reliable growth estimates. A second limitation of this method is that the significance criterion for concluding that a slope is different from zero is somewhat arbitrary (e.g., setting α at .05, .01, .05 as a base rate with correction for multiple comparisons, etc.).

Fixed criterion across multiple correlated outcomes

Each of the preceding methods shares a weakness in that outcome status is determined by performance on a single measure. Current conceptualizations of the development of reading skill are multidimensional (Scarborough, 2001) and best characterized by imperfect but correlated measures that converge on the construct. One method of identifying gainers that attempts to address this is to combine information across multiple measures. A criterion that pre-post change must exceed

is set for individual measures, but the criterion used to define who makes gains combines information from the individual classifications (i.e., gainers are those who exceed cut-points on three of four outcome measures, etc.). Termed WIGROT (see Scarborough et al., this issue) for 'within-individual gains, replicated over tests', this method requires pre-post outcome data across multiple high-quality outcome measures. WIGROT approaches the requirement that change be meaningful through the robust replication of gains across measures. A struggling reader is deemed a gainer if change has been observed on multiple measures.

The interpretation of who is identified as a gainer by the WIGROT method is relatively intuitive. Change to a criterion that has been derived from the measurement characteristics of a well-normed test replicates traditional group-based effect size at an individual level. Swanson (1999) has suggested that individual change corresponding to effect size of .20 could be a minimum criterion for defining improvement. For example, the standard error of measurement of Rasch-scaled W scores on the Woodcock Johnson III Tests of Achievement (WJ-III) Letter-Word Identification subtest is 8.24 for 13 year-old children. With a standard deviation of 21.99 (i.e., the observed standard deviation of the 13 year-olds in the sample reported below), a performance *gain* after intervention equivalent to one standard error would equate to an effect size of .37. Using multiple measures replicates gains across tests, adding evidence that identification as a gainer is valid and generalizable to reading skill more broadly, beyond the reliability limits of any particular test.

At best, WIGROTs may represent a robust individual classification of who improves, under a strong assumption that the suite of outcomes chosen are linked theoretically to the intervention and empirically to each other in previous factor validation studies. This also represents the primary limiting factor to the interpretation of gainers identified by WIGROTs: to the extent that the selection of particular outcome measures is driven by a robust dimensional model of reading skill, WIGROT will produce interpretable results. To the extent that poor underlying models and/or measures of widely disparate reading constructs are used to construct the index, WIGROT will produce uninterpretable results. An additional limitation of this approach is that the number of measures used, along with the criterion for how many measures on which change needs to be demonstrated, is arbitrary. There is currently no basis to conclude that change on four of five outcomes performs more accurately or reliably as a metric of improvement than two of three, or four of four, etc.

Limitations across methods

The first global limitation that clouds interpretation of subsets identified as gainers is regression to the mean. Participants who begin interventions with measured reading skills that are low relative to similar-aged peers—and in many cases not receiving credit for any individual items on commonly administered tests of reading skill and related processes—may be mistakenly identified as having improved due to factors other than skill gain. Each of the four methods is more or less vulnerable to this criticism. With increasing severity of reading difficulties, the normalization method is progressively less likely to identify someone as having changed due to regression

to the mean. The psychometric and WIGROT methods require that each participant exceed a value determined in part by the reliability of the test. For participants performing at floor levels of standardized tests, IRT-based psychometrics (i.e., those commonly used in current tests of reading) will provide wider standard errors, which will make it more difficult for such participants to demonstrate reliable change. Growth curve estimation is inherently less vulnerable to the regression problem because the criterion is statistically reliable growth rates, which are more stable because they depend on more than two measurement occasions. This advantage diminishes the less reliably change is measured (i.e., the fewer the observations over time), since fewer observations per individual will lead to increases in Type-I errors for the statistical test for growth rate. In addition, several studies have shown a pronounced lack of skill growth for children with severe reading difficulties in the absence of intervention (Cirino et al., 2002), further mitigating a regression to the mean interpretation, assuming reliably estimated growth curves. In the case of every method, the inclusion within the study design of an appropriate and randomly assigned control group—at least inactive or business as usual, but ideally an alternate treatment control—would also temper a regression to the mean interpretation of positive results. Demonstrating that the proportion of gainers to non-gainers in the intervention group reliably exceeds that of the control group would indicate net effects beyond regression to the mean.

The second major global limitation emerges from the adequacy of the measurement and developmental model of change. Both the psychometric and WIGROT methods rely on difference scores. The weaknesses of differences scores as a method of assessing change are both known and perhaps overplayed. Rogosa has convincingly demonstrated that the problem with unreliability of the change scores is an overrated concern (Rogosa, Brandt, & Zimowski, 1982; Rogosa, 1995), and that if the reliability of both pre and post-test measures are accounted for, the reliability of the gain scores can be known (Rogosa & Willet, 1983). Regardless of this debate about gain scores themselves, the greater limitation on the viability of all four methods is the adequacy of the psychometric model for, and measurement of, change. Measured outcomes must reflect monotonically increasing development. As well, the construct being measured must remain the same at the first and last assessment over the duration of the intervention. This means that if a test measures a particular aspect of reading for an adolescent in Grade 7, then a qualitatively commensurate skill must be measured for a Grade 8 adolescent 1 year later at intervention post-testing (Willett & Sayer, 1996). Necessary, but not sufficient, evidence of this is the absence of floor and ceiling effects along with standard errors of measurement that remain constant across the developmental period represented by the intervention (Bryk & Raudenbush, 1987). In the case of older struggling readers, this question requires careful consideration with attention to the reading-age relative to the reading assessment and not the chronological age (e.g., for a middle-school child reading at a Grade 3 level, the adequacy of the particular measure of reading should be assessed at that level of performance).

Additional general constraints across methods are that the RCI and WIGROT approaches require norm-referenced tests with a good standardization sample that includes any special population of interest (e.g., children with reading difficulties).

As a result, these methods limit the use of good experimental or population-specific tasks that may have more optimal reliability and/or validity for the sample and intervention under consideration. While experimental measures of reading could be used with both of these approaches, it would require calculation of sample-derived reliabilities (i.e., in the case of RCI) or standard errors of measurement (i.e., in the case of WIGROT), which may not be as robust as those derived from a standardized test that has been developed on a large and broadly representative sample.

The present study, despite a focus on identifying change in the context of an intervention, does not operate with the response to intervention framework (RTI), which has broader systemic and organizational concerns (but see Barth et al., 2008 for a comprehensive comparison of RTI response methods). Neither is the present study meant to provide techniques to compensate for research design faults such as the lack of a high-quality control condition. As just reviewed, each of the techniques reviewed here are vulnerable to interpretive difficulties that random assignment to intervention versus control resolves, or at least mitigates. Similarly, the present study is not proposing that the four techniques reviewed here are comprehensive, nor the best statistical techniques available. With large samples and a comprehensive measurement model, robust approaches such as latent difference score modeling are available for pre-post data (McArdle, 2001). With more detailed characterization of change at the individual level, fine-grained techniques for characterizing single-subject's trajectories are available (Beeson & Robey, 2006; Kromrey & Foster-Johnson, 1996). Rather, the present study assesses four commonly used, methods for identifying subgroups based on their improvement or lack thereof, when provided with a high quality reading intervention for struggling adolescent readers.

Three research questions were considered in the context of a year-long research-based multiple component reading intervention for struggling middle-school readers:

1. How does each method function with respect to identifying base-rates of successful intervention outcomes? What is the ability of each method to replicate traditional group-based effect sizes?
2. Do the methods agree in their identification of those who make gains?
3. The first two questions evaluate how well each method identified *who* benefited from the intervention. To evaluate *why* these participants might benefit, gainers identified by each method would need to differ according to important individual characteristics. Therefore, the final research question evaluates the comparative utility of each method in characterizing the identified subsamples on potential cognitive and language predictors of change.

Method

Participants

From 2006 to 2011, 532 middle-school students were identified as being significantly below age-expected achievement in reading across two sites: Toronto

Table 1 Description of the sample

	M	*SD*
Age	12.42	0.85
WJ-III Word Identification (SS)	76.69	11.36
WJ-III Word Attack (SS)	82.37	8.30
WJ-III Passage Comprehension (SS)	78.10	11.61
WJ-III Fluency (SS)	78.33	8.74
WASI Verbal (IQ)	88.66	10.27
WASI Performance (IQ)	96.20	12.69
Male (proportion)	.715	
Caucasian (proportion)	.711	
African American (proportion)	.152	

WJ-III Woodcock-Johnson III tests of achievement, *WASI* wechsler abbreviated scales of intelligence

and surrounding area in southern Ontario, and the metropolitan Atlanta area. The students were in grades 6, 7, or 8 and were referred by teachers in local schools. To be included in the intervention sample, participants were screened by trained psychometrists to determine that they scored 1 SD or more below age norm expectations (Standard Score <85) on the Broad or Basic Reading Cluster Score of the WJ-III Tests of Achievement (Woodcock, McGrew, & Mather, 2001).

Participants also needed to demonstrate a minimum of low average intellectual functioning (Full Scale IQ SS $\geq$ 80) on the Wechsler Abbreviated Intelligence Scale (WASI; Wechsler, 1999). Students who were English as a Second Language, had histories of significant absenteeism (>15 absences per year), hearing impairment (>25 dB at 500+ Hz bilaterally), uncorrected visual impairment (>20/40), serious emotional/psychiatric disturbance (i.e., major depression, psychotic or pervasive developmental disorder), or chronic medical/neurological condition (i.e., seizure disorder, developmental neurological conditions, acquired brain injuries) were also excluded. The present study reports on the 270 participants who received the full intervention as planned, defined as at least 112 of the full 125 h of intensive remedial intervention described below. Table 1 contains pre-intervention descriptive statistics on the full sample.

Intervention

Participants were matched into small instructional groups of 4-8 students on the basis of raw WJ-III Word Attack and Letter-Word Identification scores. Matching was conducted on raw scores regardless of grade or developmental level in reading to ensure that program scope and sequence could be optimized on a per-group basis. Instructional groups were assigned randomly to one of three conditions, all offering active reading remediation for 125 h, an hour daily, on an equivalent teacher/student ratio. Two intervention conditions offered variations of the same research-based multiple-component reading intervention program (*PHAST Reading*) and a third condition offered a "business-as-usual" special education control intervention,

being the Special Education Language Arts program typically taught in the school. The control programs were locally developed, eclectic, and included varying proportions of decoding, reading comprehension, and/or writing content. Similar to the intervention conditions, students in the control classes received remediation in a "pull-out" model, with the same teacher to pupil ratio as the intervention classes. To qualify as a control group, at least 125 h of small-group instruction must have been delivered at the same pace and intensity as the intervention groups.

The two research interventions included the *PHAST Decoding Program* which integrates phonological and strategy-based reading instruction and helps students with deficient word attack skills become strategic readers who can decode words independently and obtain meaning from print (for a full description, see Lovett, Lacerenza, & Borden, 2000). Five decoding strategies are taught, using a combination of direct instruction and dialogue-based strategy instruction, and daily practice in reading connected text is provided. The program has been described in several publications and its efficacy demonstrated with younger (Morris et al., 2012) and older struggling readers (Lovett et al., 2012). In this study, both *PHAST* interventions were designed to improve decoding, word identification, fluency and reading comprehension skills, with vocabulary development a component of both. The two conditions differed only in their relative emphases on fluency and comprehension instruction. Because the two *PHAST* interventions were characterized by greater similarities than differences and the focus of the present analysis was overall treatment effect sizes, they are treated as one intervention condition in the present analyses.

Measures

Participants completed a large battery of cognitive, language, and reading measures prior to intervention, with outcome measures repeated after 70 h of intervention, and again at the program end-point of 125 h of instruction. Four reading outcomes were chosen for the present study because of their positive psychometric properties and past utility in evaluating intervention outcomes (see Lovett et al., 2000, 2008; Morris et al., 2012).

Reading outcomes

The present study reports on the Woodcock-Johnson-3 (WJ-III) Tests of Achievement (Woodcock et al., 2001) Letter-Word Identification, Word Attack, Passage Comprehension, and Reading Fluency subtests. The Letter-Word Identification is a test of sight word reading, presenting letters and then progressively more difficult words in isolation for students to identify. On the Word Attack subtest, participants decoded a series of progressively harder pronounceable nonsense words. Passage Comprehension assesses comprehension via a cloze procedure. Reading Fluency utilizes a sentence-verification paradigm with the number of sentences correctly judged within 3 min of reading as the outcome. Across the age range of the present sample, lower-bound internal consistency or item-response theory-derived reliabilities for the four subtests are .89, .80, .81, and .90, respectively.

Predictors of improvement status

The final research question compared the utility of each method in characterizing identified gainers to non-gainers. For this step, chronological age at pretesting, gender, intervention/control status, and four individual difference variables were employed from the broader group of study measures, chosen due to past empirical or theoretical relationships with intervention outcomes. The WASI Verbal and Performance IQ scores were included as measures of verbal and nonverbal reasoning skill. The Comprehensive Tests of Phonological Processes subtests Blending Words and Phoneme Reversal were included as measures of phonological awareness and phonological working memory, respectively (Wagner, Torgesen, & Rashotte, 1999).

Identification of gainers

Normalization

To classify outcomes using the normalization approach, each participant's raw score was converted to age-normed standard scores using the procedures described in the WJ-III manual. Participants with standard scores greater than or equal to 90 at the post-intervention assessment were identified as having gained, regardless of pre-intervention starting point. This criterion is widely used in clinical and school psychology and descriptively identifies a child as being within the 'Average' range of ability relative to similar-aged peers in the standardization sample.

Psychometric method

The reliable change index was calculated according to Jacobson/Truax formula (Jacobson et al., 1984), as modified in 1986 (Christensen & Mendoza, 1986) to correct the standard error. The difference between post-intervention performance and pre-intervention performance was divided by the standard error of the pre-post difference (see below[1]). The standard modification was used, rather than later modifications involving the pooled variance of pre- and post-test scores (Maasen, 2004), since very small, and in no cases statistically significant, changes in variance were observed over time. Reliabilities specific to each age year for the WJ-III subtests were drawn from published normative data (Woodcock et al., 2001) and standard deviations were calculated per year on the observed distributions. The distribution of Word Attack W scores for participants aged 12 had one extreme score that inflated the variance, so this standard deviation was calculated on a trimmed distribution. A reliable decrease index (RDI) was also calculated for each outcome, identifying participants with scores that reliably *decreased* from pretest to end of intervention. One instance of negative change on the WJ-III Letter-Word

[1] The standard error of the difference was calculated via $\sqrt{2S_E^2}$, where S_E represents the standard error of measurement of the test score, and was estimated via $\sqrt{S_x^2 r_{xx'}}$, where S_x is the standard deviation of pretest scores and $r_{xx'}$ is the reliability of the test in the standardization sample.

Identification subtest was identified in the intervention group; one instance on the Word Attack subtest was identified in the control group. For each of the WJ-III subtests, Rasch-scaled W scores were used because the interval rescaling better represented scores at both the low and high ends of the distribution (Woodcock et al., 2001).

Growth curves

Individual growth curves were fitted to participant outcome data using outcomes measured at three consecutive times: pre-intervention, after 70 h of instruction, and again post-intervention after at least 112 h of instruction. Visual examination of individual growth curves, along with model fit statistics, were used to determine that a linear model of growth was most parsimonious, with an unstructured covariance matrix for repeated measures. SAS code for this analysis is available from the corresponding author. Gainers were identified as those participants whose individual growth slope exceeded the rate of growth expected by chance. Statistically significant negative growth rates were also identified. Similar to the RCI calculation, W scores were utilized for the WJ-III subtests. Using a base α of .05, a correction to control the false-discovery rate (Benjamini & Hochberg, 1995) and prevent over-identification of gainers was applied to the 270 individual significance tests of growth rates, using SAS PROC MULTTEST. Overall, no negative growth rates and hence no negative gains were observed that met this criterion.

Wigrot

To identify gainers using the WIGROT approach, a two-step process was followed. Participants were first identified as demonstrating univariate change on each individual WJ-III subtest if their increase from pretest to final assessment exceeded the standard error of measurement (SEM) specified in the WJ-III normative reference data. The WJ-III normative data provides SEM by whole-year age group for each measure, and these specific SEMs were used per participant. The second step involved determining outcome status, which was deemed positive for participants who met above criterion on all four of the four reading measures. Four subtests of four, rather than three or two of four were chosen for two reasons. First, using all four would generate a gainer group that was broadly representative of the full range of reading constructs. Second, given that a relatively small amount (i.e., 1 SEM) of change per measure was required per measure, four of four would ensure that the ultimate decision that change had occurred was robust. WIDROTs (within-individual decreases, replicated across measures) were also calculated per participant; however, no negative gainers were identified with this method.

Results

Intervention and control groups were equivalent on the four reading outcomes at pretest (MANOVA multivariate $F(4, 265) = 0.22$, $p = .93$). Basic treatment

Table 2 Intervention outcomes by randomly assigned condition

	Intervention (n = 228)		Control (n = 42)	
	Pretest *M (SD)*	Posttest *M (SD)*	Pretest *M (SD)*	Posttest *M (SD)*
W scores				
WJ-III Letter Word Identification	473.68 (23.27)	496.02 (23.06)	471.43 (24.67)	483.38 (21.39)
WJ-III Word Attack	476.35 (15.54)	492.37 (12.71)	475.90 (14.03)	482.33 (13.58)
WJ-III Passage Comprehension	482.97 (15.25)	496.38 (13.81)	480.71 (16.08)	489.14 (14.90)
WJ-III Reading Fluency	473.50 (20.34)	492.24 (23.02)	471.43 (18.9)	484.98 (20.12)
Standard scores				
WJ-III Letter Word Identification	76.87 (11.18)	84.25 (11.27)	75.69 (12.38)	78.26 (11.46)
WJ-III Word Attack	82.43 (8.38)	88.98 (7.28)	82.10 (8.01)	83.74 (7.55)
WJ-III Passage Comprehension	78.42 (11.42)	85.95 (11.12)	76.38 (12.56)	80.12 (12.79)
WJ-III Reading Fluency	78.48 (8.88)	83.78 (9.43)	77.52 (7.99)	81.31 (8.18)

WJ-III Woodcock-Johnson III tests of achievement

outcomes were thus evaluated with a repeated-measures MANOVA (Overall, 1994), using WJ-III W scores as the outcome metric. After screening for all assumptions, the MANOVA model provided strong evidence for differential change between intervention and control groups over pre- mid- and post-intervention time-points (time x group multivariate $F(8, 261) = 5.32$, $p < .001$, $\eta^2 = .14$). Post hoc univariate time by group orthogonal polynomic interactions were conducted, with a significant linear growth advantage observed for intervention participants on WJ-III Letter-Word Identification ($F(1, 268) = 19.00$, $p < .001$, $\eta^2 = .066$), Word Attack ($F(1, 268) = 24.67$, $p < .001$, $\eta^2 = .084$), Passage Comprehension ($F(1, 268) = 8.52$, $p = .004$, $\eta^2 = .031$), and Reading Fluency subtests ($F(1, 268) = 5.38$, $p = .021$, $\eta^2 = .020$). The results indicated a significant and substantial overall time by intervention effect, extending at the univariate level to each of single-word identification, nonword decoding, passage comprehension, and reading fluency components of reading skill. Table 2 contains pre- and post-intervention W and standard scores for the intervention and control conditions.

Analysis 1: Replicating traditional group-based measures of effect size

The first analysis of the methods of identifying intervention outcomes focused on whether each method replicated traditional group-based effect sizes. Proportions of participants identified as having changed within the intervention and control groups are reported in Table 3. Across outcome measures, the normalization method identified the fewest gainers, with increasing proportions for the RCI and growth curve methods. WIGROTs identified rates of gainers similar to the normalization approach, very close to the average across the four outcomes. Within measures, higher rates of change were observed for the two single-word measures (i.e., WJ-III Letter-Word Identification and Word Attack subtests), with lower rates overall for

Table 3 Performance of each method of determining responder status

	Proportion responders in the Intervention (Control)				Odds-ratio			
	Norm.	RCI	Growth	WIGROT	Norm.	RCI	Growth	WIGROT
WJ-III Letter-Word Identification	.368 (.214)	.539 (.238)	.750 (.476)	.399 (.167)	2.139	3.749	3.300	3.321
WJ-III Word Attack	.443 (.286)	.452 (.214)	.645 (.357)		1.988	3.021	3.267	
WJ-III Passage Comprehension	.417 (.286)	.346 (.190)	.820 (.667)		1.786	2.253	2.281	
WJ-III Fluency	.259 (.167)	.360 (.190)	.868 (.714)		1.746	2.387	2.640	

WJ-III Woodcock-Johnson III tests of achievement

the two passage/sentence-level reading tasks (WJ-III Passage Comprehension and Reading Fluency subtests). Since each of the four methods produced separate proportions of gainers for intervention and control groups, the odds-ratio representing the increase in chance to be identified as having gained, given being an intervention participant, was calculated since it is a commonly used and understood effect size. Since the odds-ratios compared proportion gainers in the treatment compared to control conditions, they can be considered estimates of intervention response. As displayed in Table 3, there were clear strong effects for both single-word measures, and notable effects for passage comprehension and reading fluency. For example, using the RCI method, a participant in the intervention condition was 3.75 times more likely to be identified as a responder on WJ-III Letter-Word Identification. The normalization approach produced the weakest odd-ratios overall, with the growth curve approach producing the greatest discrepancy for word-level versus passage/sentence-level measures. Overall, odds-ratios ranged from 1.7 to 3.7, indicating that intervention participants were 1.7–3.7 times more likely to be identified as responders. Despite identifying the fewest gainers overall, the WIGROT approach was associated with a strong odds-ratio, with a 3.32 times greater chance of being identified as a responder, given random assignment to intervention.

To draw comparisons between each classification method and traditional effect size measures, Cohen's d with Hedges adjustment for sample size was calculated using mean gains and standard deviation of gains for the intervention and control participants. This provided an effect-size statistic that was based on group data. As seen in Table 4, the intervention advantage in change from pretest was associated with strong to moderate effects, with the strongest effect for the WJ-III Letter-Word Identification ($d = .721$) and Word Attack ($d = .792$) subtests, and relatively weaker but still moderate effects for the Passage Comprehension ($d = .453$) and Reading Fluency subtests ($d = .389$).

Using a formula provided in Chinn (2000), the odds-ratios generated by each gainer method were converted to Cohen's d to provide an estimate for how well each identification method replicated group-based effect sizes. This formula functions well as a conversion assuming proportions that range from .03 to .97. With extremely high proportions of gainers or non-gainers, this formula does not accurately convert the odds-ratio to Cohen's d. Not surprisingly, as the RCI

Table 4 Comparison of effect size generated from change scores with d converted from odds-ratios per responder method

	Cohen's *d*	*d* converted from odds-ratios			
		Norm.	RCI	Growth	WIGROT
WJ-III Letter-Word Identification	.721	.419	.729	.658	.662
WJ-III Word Attack	.792	.379	.610	.653	
WJ-III Passage Comprehension	.453	.319	.448	.455	
WJ-III Fluency	.389	.307	.480	.535	

Cohen's *d* with Hedges adjustment for sample size calculated from the mean post–pre difference scores and standard deviation of the difference scores; Cohen's *d* for the four methods of identifying responders generated using the conversion provided by Chinn (2000)

WJ-III Woodcock-Johnson III tests of achievement

approach to identifying gainers relies on difference scores, this method closely replicated Cohen's *d*; similarly, the WIGROT approach produced an effect size that was comparable to the group-based *d*'s blended across outcome measures. The normalization approach produced effect sizes that were substantially lower than the group-based analysis, with the greatest discrepancy for measures at the word level. The growth curve approach, while identifying a much higher *proportion* of gainers across both intervention and control conditions, produced effect sizes similar to Cohen's *d* for all outcomes.

Analysis 2: Agreement among methods in identifying change

For each measure, Kappa was calculated to estimate the degree of agreement among methods of identifying individual intervention outcome (see Table 5). Overall, chance-corrected agreement rates were low, ranging from chance/negative agreement (e.g., −.14, ns) to good agreement (e.g., .60, $p < .001$). Good agreement was observed only between classification by the RCI and growth curve approaches and only when the outcomes were the WJ-III Letter-Word Identification and Word Attack subtests. Moderate agreement was observed between these two approaches on WJ-III Reading Fluency. Agreement between the normalization method and the other three methods was consistently poor (i.e., $\kappa < .20$), excepting fair agreement with RCI and growth curve methods when the outcome was WJ-III Letter-Word Identification. The WIGROT method showed consistently fair agreement with other methods and across measures, excepting the chance agreement observed with the normalization method on WJ-III Word Identification outcomes.

Analysis 3: Capacity to identify individual differences

The classifications created by application of the four methods were next considered as outcomes, using binary logistic regression to establish whether a set of predictors would explain who was identified as a gainer. The goal of this analysis was to evaluate the four methods in their productive capacity to characterize outcome groups. Predictors were intervention assignment as a statistical control, age at

Table 5 Per-measure agreement (Cohen's kappa) among four methods of identifying intervention responders

	Norm.	RCI	Growth	WIGROT
WJ-III Letter-Word Identification (above diagonal)				
WJ-III Word Attack (below diagonal)				
Normalization		.27**	.20**	.09
Reliable change index	.09		.57**	.20**
Growth curve	−.08	.60**		.22**
WIGROT	.12**	.27**	.34**	
WJ-III Passage Comprehension (above diagonal)				
WJ-III Reading Fluency (below diagonal)				
Normalization		.14*	.04	.14**
Reliable change index	−.05		.22**	.24**
Growth curve	.10	.00		.24**
WIGROT	.14**	.20**	.19**	

WJ-III Woodcock-Johnson III tests of achievement

*$.10 > p > .05$; ** $p < .05$

intervention start, gender, along with five cognitive and language variables. These five were chosen because of past theoretical or repeatedly demonstrated empirical relationships with reading, including the following: (a) phonological blending skill (for relationships with reading and/or intervention outcomes, see Wagner & Torgesen, 1987); (b) phonological-loop working memory (Gathercole, Alloway, Willis, & Adams, 2005); (c) rapid automatized naming for letters (Swanson, Trainin, Necoechea, & Hammill, 2003); (c) verbal intelligence (Vellutino, Scanlon, & Lyon, 2000); (d) nonverbal intelligence (Tiu, Thompson, & Lewis, 2003). Measures used to index these constructs were, respectively, Comprehensive Tests of Phonological Processing (CTOPP) Blending Words, CTOPP Phoneme Reversal, Rapid Automatized Naming (RAN) Letters, Wechsler Abbreviated Scale of Intelligence (WASI) VIQ, and WASI PIQ.

After screening for univariate and multivariate outliers, assessing adequacy of proportions for the three categorical predictors, and linearity in the logit for the five continuous predictors, a series of binary logistic models were evaluated. Four models were evaluated for each of the three methods that utilized univariate outcomes (i.e., normalization, RCI, and growth curves). In addition, one model was evaluated using WIGROT gainer classification, since this method combines information from all four outcomes. Table 6 presents the partial results from these models, reporting the odds-ratio for each predictor, along with a pseudo R-squared (Nagelkerke) for each model.

Intervention and categorical predictors

All categorical predictors were dummy-coded, such that the odds-ratio represents an increased (for ratios above 1) or decreased (for ratios below 1) chance of being

Table 6 Logistic regressions predicting responder status from cognitive and language variables

	Normalization	Reliable change index	Growth	WIGROT All outcomes
WJ-III Letter-Word Identification				
Intervention (Int. = 1)	2.46*	4.76*	4.56*	3.54*
Gender (Female = 1)	2.51*		2.21*	
Age				0.66*
WASI Verbal IQ	1.69*	1.58*		
WASI Performance IQ		1.33*		
CTOPP Blending Words				
CTOPP Phoneme Reversal	1.74*			
RAN Letters	2.03*		.730*	
Nagelkerke R^2	.289	.190	.166	.111
WJ-III Word Attack				
Intervention		3.61*	4.54*	
Gender				
Age				
WASI Verbal IQ			.67*	
WASI Performance IQ				
CTOPP Blending Words				
CTOPP Phoneme Reversal	2.22*			
RAN Letters	1.44*	.73*	.71*	
Nagelkerke R^2	.244	.100	.160	
WJ-III Passage Comprehension				
Intervention		2.33*	2.65*	
Gender	1.98*			
Age				
WASI Verbal IQ	2.21*			
WASI Performance IQ		1.37*		
CTOPP Blending Words				
CTOPP Phoneme Reversal	1.35*			
RAN Letters	1.47*		.68*	
Nagelkerke R^2	.266	.087	.110	
WJ-III Reading Fluency				
Intervention		2.44*	2.59*	
Gender	2.31*			
Age	0.62*			
WASI Verbal IQ				
WASI Performance IQ		1.34*	1.50*	
CTOPP Blending Words		1.32*		
CTOPP Phoneme Reversal				
RAN Letters	2.52*		1.91*	
Nagelkerke R^2	.22	.126	.164	

WJ-III Woodcock-Johnson III tests of achievement, *WASI* Wechsler abbreviated scale of intelligence, *CTOPP* comprehensive tests of phonological processing, *RAN* rapid automatized naming

* $p < .05$

identified as a gainer for the group coded one (see Table 6 for codes). Intervention assignment remained significantly associated with outcomes in ten of the 13 models, indicating that intervention effects were robust to the inclusion of phonological and cognitive predictors. When classifications were generated by the RCI or growth curve methods, intervention assignment remained a significant predictor for all outcomes. For example, after controlling for cognitive and language skills, being assigned to the intervention remained associated with a 4.76 times greater chance of being identified as a gainer by the RCI method when the outcome was WJ-III Word Identification, and 4.56 when the classification method was growth curve. Intervention assignment was *not* predictive of improvement for all but one (i.e., WJ-III Word Identification) of four models in which the normalization approach was used to generate classifications. The WIGROT approach was robust to the inclusion of other predictors, with a 3.54 times greater chance of identification as a gainer having been assigned to intervention. Being female was associated with a greater chance of being identified as having made gains, but *only* when the method was normalization, for three of four outcomes.

Continuous predictors

To simplify interpretation all continuous predictors were converted to z-scores prior to model entry. As a result, the odds-ratios reported in Table 6 can be interpreted as the increase (for ratios above 1) or decrease (for ratios below 1) in the chance to be a gainer, given a one standard deviation higher (or lower) score on that variable relative to the whole sample. For example, being 1 SD above the sample mean on WASI Verbal IQ was associated with a 1.58 times greater probability of being identified as a gainer on WJ-III Letter-Word Identification, when the method of identifying positive outcomes was RCI. This effect for verbal intelligence was also observed for the normalization method on this outcome (OR = 1.69) and on WJ-III Passage Comprehension (OR = 2.21). Rapid automatized naming speed was associated with an increased chance of being identified as a gainer across all outcomes, for the normalization method. Naming speed produced varying results when applied to classifications generated by the growth curve method. Fast naming speed was associated with increased chance of being a gainer when the outcome was WJ-III Reading Fluency (OR = 1.91). For WJ-Letter Word Identification, Word Attack, and Passage Comprehension, *slower* initial naming speed was associated with a greater chance to be identified as a gainer (ORs = .73 .71, and .68, respectively).

Across methods some inconsistencies were also noted. For example, faster performance on the RAN Letters was associated with greater odds of being identified as a gainer across all outcomes, but only when the method was normalization. This effect was inverted when the method of classification was growth curve, with slower performance on the task associated with greater odds of being identified as a gainer on most outcomes. Post hoc examination of this effect for the WJ-III Passage Comprehension outcome revealed that for the normalization method, gainers completed the RAN Letters with a median time of 28.0 s (IQR = 8.6), approximately two s faster than those identified as non-gainers

(median = 30.4; IQR = 10.7). For the same outcome, this effect was reversed under the growth curve method, with gainers completing the RAN Letters with a median time of 30.2 s (IQR = 10.2), slower than those identified as non-gainers (median = 28.1; IQR = 9.0). Recall from Table 5 that the agreement between these two measures, when WJ-III Passage Comprehension was the outcome, was at chance (i.e., $\kappa = .04$, ns) indicating that the two methods identified almost completely different subsamples as having made significant change on those measures. Similar reversals were observed across other models, with the growth curve approach tending to indicate that *lower* cognitive and language skills were associated with increased odds of being classified as a gainer. The normalization method tended to indicate that *higher* cognitive and phonological skills (along with being female) were associated with increased odds of being classified as a gainer.

Discussion

The present study had three goals in assessing the performance of three common, and one recently proposed, method for identifying who benefits from reading intervention. First, how does each method function with respect to identifying base-rates of improvement for intervention participants? Rates ranged from a low of 25.9 % to a high of 86.8 %, across outcomes and identification method. Taking the results of the first analysis in isolation, it is clear that the four methods of identifying a subsample of intervention gainers cannot substitute for comparisons to a control group, and according to the present results are likely to dramatically under *and* over-estimate intervention effects. The normalization method for fluency outcomes identified only 25.9 % as having changed within the intervention condition. This is a rate low enough to conclude no fluency effect after 125 h of instruction and would have limited utility in identifying a sizeable enough pool of individuals showing positive change for further analysis in all but the largest samples. As well, this rate belies the strong results from the standard repeated measures analysis of fluency outcomes, which showed clear intervention effects on fluency over the course of intervention. Similarly, should the proportion of fluency gainers identified by the growth curve method be interpreted in the absence of control results, the result would be an extremely encouraging result of 86.8 % of intervention students showing improvement. In both cases, these results are properly contextualized after accounting for control gains, with fluency effect sizes converted from these proportions of .31 and .54 for the normalization and growth curve method, respectively.

The first analysis goal also investigated the ability of each method to replicate traditional group-based effect sizes. Across the four outcomes, the normalization method underestimated traditional group-based effect sizes. The RCI and growth curve approaches slightly underestimated effect sizes for word-level outcomes, and in all but one case overestimated effect sizes on outcomes that involved connected text. On a substantive level, basic group-based effect sizes (e.g., Cohen's *d*) ranged from moderate to large for this intervention with middle-school adolescents,

comparable to the higher end of the effect sizes reported by Edmonds and colleagues in their meta-analysis (Edmonds et al., 2009).

Second, do the methods agree in their identification of gainers? There was profoundly low and inconsistent agreement across some methods of identifying successful outcomes. Other methods produce higher gain rates, but agree moderately with each other (e.g., especially RCI and growth curve approaches). Agreement was highest on reading skill outcomes that involved single words or nonwords, and approached acceptable levels between the RCI and growth curve approaches. This level of agreement provides converging evidence of the validity of these approaches, since it was attained despite the differences in statistical method and structure of outcome data used by each. The growth curve approach used all three assessments for generating individual slopes, whereas the RCI approach used only the first and last assessments in calculation of gains. This is evidence to suggest that in the absence of data from multiple (i.e., >2) assessments, the RCI approach can yield consistent and similar estimates of outcome success. In contrast, when information from the proportion of gainers identified and agreement among methods is combined, the normalization approach not only identifies lower rates, but also does not agree with other methods.

Third, what is the comparative utility of each method in characterizing the identified subsamples on potential predictors of change? This question is critical since the motivation to identify gainers is often planned as a step toward identifying factors that may moderate intervention outcome—individual differences potentially relevant to the *who* and *why* of intervention response. To address this question, the change/no-change classification was investigated as an outcome, with six demographic, cognitive, and phonological factors as predictors. Overall, assignment to intervention or control was robust, with clear and statistically significant differences in rates of responders even after accounting for all predictors. The RCI and growth curve classification performed best in this regard. In contrast, intervention condition did *not* remain predictive of change for the normalization method, which also appeared to identify subsamples by overall lack of severity in the cognitive and phonological profile. Higher IQ and phonologically capable children, along with girls, were identified as having positive outcomes by this method. Since the normalization method relies on only one score (i.e., post-test) and since classification begins with an index of severity (i.e., age-normed standard scores), this is not surprising. These issues limit the utility of the normalization method in identifying the *who* and *why* of intervention response.

It is also important to note that each of the significant predictors of who was classified as a gainer was entered into the model only as a main effect. As noted elsewhere (Yoder & Compton, 2004), this is insufficient evidence to support these predictors as predictors of treatment *response*. For such a determination a predictor would have to interact with treatment status in predicting those identified as gainers. While exploratory analyses were conducted, such interactions were not reported since post hoc investigations of the significant interactions suggested they were being driven by a small group of controls who had been classified as gainers. Once participants were subdivided by treatment/control, and then gainer/nongainer relatively few (e.g., 7–8 depending on method) gainer controls were identified. This

underlines the fact that the search for factors that characterize outcome groups is a large-sample issue, requiring many participants, good base-rates identified with positive outcomes, and adequately-sized control samples to be successful.

In the single model evaluating WIGROT classifications, intervention condition was robust to the inclusion of cognitive and phonological predictors. Despite this robust effect, none of the individual difference predictors accounted for gainer categorization. There are two possible explanations for this result. On the one hand, the WIGROT approach utilizes information from all reading outcomes. Recall that estimates of the proportion of gainers for this approach were similar to those observed with the other methods, averaging across the four reading outcomes. As such, robust determination of who made positive outcomes may be at a cost of the loss of differential predictive power across outcomes. On the other hand, it is possible that the predictors identified by the other approaches are spurious, resulting from gainer classification being determined by one reading measure. This interpretation is less likely, since the effects observed for the cognitive and phonological predictors were substantial and significant at a corrected alpha level.

The RCI and growth curve approaches appeared to function well in the task of identifying factors that predict gainer categorization. Intervention condition was robust to the inclusion of these predictors, which themselves continued to predict response category. The particular factors confirm and add to the growing literature that identifies correlates of intervention response (see Fletcher et al., 2011; Frijters et al., 2011). A minor limitation should be noted in that the intervention group combined two very similar versions of the PHAST reading intervention, one with a fluency focus and one with a comprehension focus. While separating effects by PHAST variant was beyond the scope of this paper, future analyses could make productive use of the methods presented here to unpack these dynamics.

To summarize the results that have the greatest implication for their use in further studies, a pattern of strengths and weaknesses emerged when each was applied to data on adolescent struggling readers participating in an intensive multi-component small-group intervention. The normalization method, as expected based on the age of these struggling readers, identified low rates of responders in both intervention and control groups. This method also consistently *underestimated* traditional group-based effect size measures. The normalization method produced the greatest variety of predictors; however, the predictors identified appeared to tap into a global cognitive and language severity trend—those with better cognitive and language skills were consistently more likely to be identified as having made gains. The RCI and growth curve approaches provided the most accurate and consistent replication of traditional group-based effect sizes across all four outcomes, while also providing the clearest separation between responder rates across the intervention and control conditions. The growth curve approach also identified the highest rate of gainer classifications of all four methods, though care needs to be taken in interpreting this as a strength given the somewhat arbitrary setting of significance levels. The present study also measured outcomes on only three occasions. This frequency is the minimum for establishing and testing linear growth rates using hierarchical linear models, and underlines the notion that the utility of the growth rate approach depends directly on the model and measurement of change. Despite this, both the

RCI and growth curve approaches identified several potential predictors of successful outcomes. This was consistent with the very high levels of classification agreement across these two measures. The WIGROT approach only identified one predictor, with younger adolescents being identified as those who showed robust gains across all four outcomes. This approach also identified the lowest proportion of gainers overall, though this was due in part to change on four out of four measures being required in the present analysis. Despite this, the WIGROT approach demonstrated excellent sensitivity to detecting intervention effects, similar to the RCI and growth curve approaches.

There has been increasing attention to the cost of averaging outcomes across members of an intervention group, when potentially important information about outcome is contained within the heterogeneity of the group. Whether from the covariance structure approach (e.g., latent class analysis, latent transition analysis, latent class growth analysis, growth mixture modeling, and general growth mixture modeling; Muthén & Muthén, 2000) or from the movement away from variable-centric views (e.g., to person-centric analytic methods; Magnusson & Allen, 1983; Magnusson & Bergman, 1988), there is a growing emphasis on making use of such heterogeneity. Within these views, person-to-person variation has substantive interest beyond the functional role of calculating error terms prescribed by traditional parametric group comparisons. The methods reviewed in this paper provide accessible, and in some cases robust and consistent, starting points for quantifying and using person-to-person variability in outcomes.

Acknowledgments Supported by a grant from the Institute for Educational Sciences, Georgia State University (#R324G06005).

References

Al Otaiba, S., & Fuchs, D. (2002). Characteristics of children who are unresponsive to early literacy intervention: A review of the literature. *Remedial and Special Education, 23*, 300–316.

Aunola, K., Leskinen, E., Onatsu-Arvilommi, T., & Nurmi, J.-E. (2002). Three methods for studying developmental change: A case of reading skills and self-concept. *British Journal of Educational Psychology, 72*(3), 343–364.

Barth, A. E., Stuebing, K. K., Anthony, J. L., Denton, C. A., Mathes, P. G., Fletcher, J. M., et al. (2008). Agreement among response to intervention criteria for identifying responder status. *Learning and Individual Differences, 18*(3), 296–307.

Bauer, S. B., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment, 82*(1), 60–70.

Beeson, P. I. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology Review, 16*(4), 161–169. doi:10.1007/s11065-006-9013-7.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, B, 57*, 289–300.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*(1), 147–158.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine, 19*, 3127–3131.

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy, 17*, 305–308.

Cirino, P. T., Rashid, F. L., Sevcik, R. A., Lovett, M. W., Frijters, J. C., Wolf, M., et al. (2002). Psychometric stability of nationally normed and experimental decoding and related measures in children with reading disability. *Journal of Learning Disabilities, 35*(6), 526–539.

Compton, D. L. (2000). Modeling the response of normally achieving and at-risk first grade children to word reading instruction. *Annals of Dyslexia, 50*, 53–84.

Compton, D. L. (2003). Modeling the relationship between growth in rapid naming speed and growth in decoding skill in first-grade children. *Journal of Educational Psychology, 95*(2), 225–239.

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Klinger-Tackett, K., et al. (2009). A Synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of Educational Research, 79*(1), 262–300.

Evans, C. (1988). *Reliable change criterion calculator*. Retrieved from http://www.psyctc.org/stats/rcsc1.htm.

Fletcher, J. M., Stuebing, K. K., & Barth, A. E. (2011). Cognitive correlates of inadequate response to reading intervention. *School Psychology Review, 40*(1), 3–22.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90*(1), 37–55.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88*(1), 3–17.

Frijters, J. C., Lovett, M. W., Steinbach, K. A., Wolf, M., Sevcik, R. A., & Morris, R. (2011). Neurocognitive predictors of reading outcomes for children with reading disabilities. *Journal of Learning Disabilities, 44*(2), 150–166.

Fuchs, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities Research and Practice, 18*(3), 172–186.

Gathercole, S. E., Alloway, T. P., Willis, C., & Adams, A.-M. (2005). Working memory in children with reading disabilities. *Journal of Experimental Child Psychology, 93*, 265–281.

Greenberg, D., Wise, J. C., Morris, R., Fredrick, L. D., Rodrigo, V., Nanda, A. O., et al. (2011). A randomized control study of instructional approaches for struggling adult readers. *Journal of Research on Educational Effectiveness, 4*(2), 101–117.

IBM. (2010). IBM SPSS statistics (version 19.0). Chicago, IL.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15*, 336–352.

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *Journal of Experimental Education, 65*(1), 73–93.

Lovett, M. W., De Palma, M., Frijters, J. C., Steinbach, K. A., Temple, M., Benson, N. J., et al. (2008). Interventions for reading difficulties: A comparison of response to intervention by ELL and EFL struggling readers. *Journal of Learning Disabilities, 41*(4), 333–352.

Lovett, M. W., Lacerenza, L., & Borden, S. L. (2000). Putting struggling readers on the PHAST track: A program to integrate phonological and strategy-based remedial reading instruction and maximize outcomes. *Journal of Learning Disabilities, 33*(5), 458–476.

Lovett, M. W., Lacerenza, L., Borden, S. L., Frijters, J. C., Steinbach, K. A., & De Palma, M. (2000). Components of effective remediation for developmental reading disabilities: Combining phonological and strategy-based instruction to improve outcomes. *Journal of Educational Psychology, 92*(2), 263–283.

Lovett, M. W., Lacerenza, L., De Palma, M., & Frijters, J. C. (2012). Evaluating the efficacy of remediation for struggling readers in high school. *Journal of Learning Disabilities, 45*(2), 151–169.

Maasen, G. H. (2004). The standard error in the Jacobson and Truax reliable change index: The classical approach to the assessment of reliable change. *Journal of the International Neuropsychological Society, 10*, 888–893.

Magnusson, D., & Allen, V. L. (1983). An interactional perspective for human development. In D. Magnusson & V. L. Allen (Eds.), *Human development: An interactional perspective* (pp. 369–387). New York, NY: Academic Press.

Magnusson, D., & Bergman, L. R. (1988). Individual and variable-based approaches to longitudinal research on early risk factors. In M. Rutter (Ed.), *Studies of psychosocial risk: The power of longitudinal data* (pp. 45–61). Cambridge: Cambridge University Press.

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analysis. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 342–380). Lincolnwood, IL: Scientific Software International.

McGlinchey, J. B., Atkins, D. C., & Jacobson, N. S. (2002). Clinical significance methods: Which one to use and how useful are they? *Behavior Therapy,* 33, 529–550.

Morris, R. D., Lovett, M. W., Wolf, M., Sevcik, R. A., Steinbach, K. A., Frijters, J. C., et al. (2012). Multiple-component remediation for developmental reading disabilities: IQ, SES, and race as factors on remedial outcome. *Journal of Learning Disabilities, 45*(2), 99–127.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24*(6), 882–891.

Nelson, J. R., Benner, G. J., & Gonzalez, J. (2003). Learner characteristics that influence the treatment effectiveness of early literacy interventions: A meta-analytic review. *Learning Disabilities Research and Practice, 18*, 255–267.

Olson, R. K., Wise, B., Ring, J., & Johnson, M. (1997). Computer-based remedial training in phoneme awareness and phonological decoding: Effects on the posttraining development of word recognition. *Scientific Studies of Reading, 1*(3), 235–254.

Overall, J. E. (1994). Issues in the design and analysis of controlled clinical trials. *Journal of Clinical Psychology, 51*(1), 95–102.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. T. (2007). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Reed, D., & Vaughn, S. (2010). Reading interventions for older students. In T. A. Glover & S. Vaughn (Eds.), *The promise of response to intervention: Evaluating current science and practice* (pp. 143–186). New York, NY: The Guilford Press.

Rogosa, D. R. (1995). Myths and methods: "Myths About Longitudinal Research" plus supplemental questions. In J. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah, NJ: Lawrence Erlbaum Associates.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*(3), 726–748.

Rogosa, D. R., & Willet, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20*(4), 335–343.

Sabatini, J. P., Shore, J., Holtzman, S., & Scarborough, H. S. (2011). Relative effectiveness of reading intervention programs for adults with low literacy. *Journal of Research on Educational Effectiveness, 4*(2), 118–133.

SAS Institute, I. (2008). SAS/STAT system (Version 9.2). Cary, NC.

Scarborough, H. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. Neuman & D. Dickinson (Eds.), *Handbook for research in early literacy* (pp. 97–110). New York: Guilford Press.

Singer, J. D., & Willet, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Smart, D., Sanson, A., & Prior, M. (1996). Connections between reading disability and behavior problems: Testing temporal and causal hypotheses. *Journal of Abnormal Child Psychology, 24*(3), 363–383.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Swanson, H. L. (1999). Reading research for students with learning disabilities: A meta-analysis of intervention outcomes. *Journal of Learning Disabilities, 32*(6), 504–532.

Swanson, H., Trainin, G., Necoechea, D. M., & Hammill, D. D. (2003). Rapid naming, phonological awareness, and reading: A meta-analysis of the correlation evidence. *Review of Educational Research, 73*(4), 407–440.

Temkin, N. R. (2004). Standard error in the Jacobson and Truax Reliable Change Index: The "classical approach" leads to poor estimates. *Journal of the International Neuropsychological Society, 2004*(10), 899–901.

Tiu, R. D., Thompson, L. A., & Lewis, B. A. (2003). The role of IQ in a component model of reading. *Journal of Learning Disabilities, 36*(5), 424–436.

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34*(1), 33–58.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Alexander, A. W., & Conway, T. (1997). Preventive and remedial interventions for children with severe reading disabilities. *Learning Disabilities: A Multidisciplinary Journal, 8*, 51–62.

Vellutino, F. R., Scanlon, D. M., & Lyon, G. R. (2000). Differentiating between difficult-to-remediate and readily remediated poor readers. *Journal of Learning Disabilities, 33*(3), 223–238.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R., et al. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88*(4), 601–638.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin, 101*(2), 192–212.

Wagner, R., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive tests of phonological processing (CTOPP)*. Austin, TX: Pro-Ed.

Wechsler, D. (1999). *Wechsler abbreviated scale of intelligence (WASI)*. New York: The Psychological Corporation.

Willett, J. B., & Sayer, A. G. (1996). Cross-domain analysis of change over time: Combining growth modeling and covariance structure analysis. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling. Issues and techniques*. Mahweh, NJ: Lawrence Erlbaum.

Wise, B. W., Ring, J., Sessions, L., & Olson, R. K. (1997). Phonological awareness with and without articulation: A preliminary study. *Learning Disability Quarterly, 20*(3), 211–225.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of achievement (WJ-III)*. Itasca, IL: Riverside Publishing.

Yoder, P., & Compton, D. (2004). Identifying predictors of treatment response. *Mental Retardation and Developmental Disabilities Research Reviews, 10*, 162–168.