Published in final edited form as:

J Learn Disabil. 2012 November ; 45(6): 565–569. doi:10.1177/0022219412451999.

Meta-Analysis and Inadequate Responders to Intervention: A Response

Karla K. Stuebing, PhD¹, Jack M. Fletcher, MD¹, and Lisa C. Hughes, MA¹

¹University of Houston, Houston, TX, USA

In a recently published meta-analysis, Tran, Sanchez, Arellano, and Swanson (2011) synthesized 13 studies that permitted assessment of characteristics of children who were adequate and inadequate responders to instruction. The authors indicated that “[t]he central question addressed in this review is whether individual differences in reading-related skills at pretest predict responders at posttest across a variety of interventions and sets of criteria for determining responding and low responding” (p. 283). Of specific interest was whether the performance on specific individual difference variables at pretest, such as phonological awareness, better predicted posttest performance than other measures (e.g., rapid automatic naming) (p. 284). It is surprising, given the focus on individual differences, Tran et al. (2011) operationalized their research questions with a between-groups model by calculating effect size (ES) differences between groups of responders and nonresponders at pretest and posttest and then predicting the posttest ES from the pretest ES. HLM analyses were performed to investigate whether the class of variable (e.g., reading comprehension vs. rapid naming) or methodological variables explained further variance in the posttest effects. Their analyses lead them to conclude,

The validity of RTI procedures, particularly in comparison to other assessment approaches, has not been adequately established in the present synthesis of the literature. The current synthesis suggests that RTI procedures do improve performance as reflected in gain scores, however, the differences (achievement gap) between responding and low responding children were maintained across pretest and posttest conditions. (p. 293)

In terms of individual predictors, they further concluded that “we did not find that phonological awareness was a significant moderator of ES at posttest when all other measures were entered into the hierarchical linear model” (p. 292). Al Otaiba and Fuchs (2002) and others (e.g., Fletcher et al., 2011) found that phonological awareness (PA) was a good predictor of inadequate response.

In this article we consider the following questions: (a) Is the analytic framework in Tran et al. (2011) using pretest ESs to predict posttest ESs appropriate for their research question? (b) Are the Cohen guidelines for the use and interpretation of ES d meaningful in the context of this meta-analysis, that is, should d values from studies where groups are formed by dichotomization be compared to the Cohen benchmarks or to d values from experimental studies? (c) Have Tran et al. demonstrated that PA is not a good predictor of response? and

© Hammill Institute on Disabilities 2012

Corresponding Author: Karla K. Stuebing, Texas Institute for Measurement, Evaluation and Statistics, University of Houston, 2151 W. Holcombe, 222 Texas Medical Center Annex, Houston, TX 77204, USA, Karla.stuebing@times.uh.edu.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

(d) Does their analysis and interpretation of these results support the conclusion that RTI does not reduce the postintervention achievement gap?

Is the pre-post effect size analytic framework appropriate for a metaanalysis of intervention response?

Tran et al. (2011) used hierarchical linear modeling (HLM) to predict the posttest ES from the pretest ES, finding that “the magnitude of the ESs increased in some cases from pretest to posttest. Thus the data do not support the notion that posttest scores as a function of RTI provide outcomes independent of pretest scores” (p. 292). It is difficult to see why effective interventions in any service delivery model, including RTI, would lead to predictions of ESs that are not larger at posttest than at pretest. In fact, the larger difference between adequate and inadequate responders at posttest seems consistent with typical RTI studies where students are rather homogenous at pretest (i.e., meet criteria for risk) but heterogeneous after intervention, with some responding adequately and others inadequately. By definition, those designated responders will have higher scores than those designated nonresponders on the measures used to define the groups.

The generally larger posttest ESs can also be explained methodologically. Tran et al. (2011) estimated ES on a variety of measures and time points. In most (but not all) of the studies in this meta-analysis, the groups were defined by dichotomizing an outcome (not necessarily the planned posttest) used to indicate instructional response. This outcome is correlated to the posttest both because of construct overlap but also because the outcome and posttest were measured close to the same point in time. Because of time-specific measurement error and differential growth, the correlation of the dichotomized outcome with the pretest will be lower than the correlation of the outcome with the posttest. The result of this pattern of intercorrelations is that the ES will be largest on the variable that is dichotomized (the outcome variable), smaller in the posttest, and smaller still in the pretest. For example, in a study where the posttest was performance on a measure of word identification and dichotomization was done on a measure of reading comprehension, the posttest effects will be smaller on word identification than reading comprehension because they are measuring different but correlated constructs at the same point in time. In contrast, we would expect the pretest word identification ESs to be smaller than the posttest effects on the word identification because they differ from the dichotomizing variable in both construct measured and time. Groups that have no overlap on the dichotomized variable will overlap on other variables as a function of the degree of correlation. We would expect total group overlap, equal means, and an ES of 0 for the two groups on a variable with a correlation of 0 with the dichotomized variable.

For some variables in this synthesis, such as rapid automatized naming (RAN) and PA, the average ES is larger at pretest than it is at posttest. This is an unexpected pattern if responders are defined after an intervention. We located the 13 studies included in this synthesis and coded each one for the timing of the group designation. In three of the studies, each of which could contribute multiple ESs to the meta-analysis (Allor, Fuchs, & Mathes, 2001; Menzies, Mahdavi, & Lewis, 2008; Ukrainetz, Ross, & Harm, 2009), the “responder” group designation was made on the basis of pretest variables. Tran et al. (2011) specifically cited Allor et al. (2001) as a study where the responder groups were defined on the basis of the alternating stimulus RAN, a test given at pretest to assess the risk status of the student. In these situations, we would expect a larger ES at pretest and a smaller effect at posttest. The resulting data set is a heterogeneous mix of ESs where, for the most part, groups are formed at posttest and we would expect larger posttest differences. However, in a substantial number of cases, the designation of “responder” was made on pretest variables. In these studies, we would expect larger pretest effects. This latter approach is on its face

inconsistent with the notion of response occurring after the intervention and is logically troublesome. What is more difficult is that the analysis of the relation between the pre- and posttest ESs is highly influenced by the mix, or the proportion of the ESs representing each of these two approaches. If there were mostly posttest designations, we would expect a positive intercept in the HLM model. If there were mostly pretest designations, we might expect a negative intercept. The slope in the analysis tells us whether the pretest difference is informative about the posttest difference. When pretest and posttest correlations are fairly homogeneous across all of the measures being considered, this must be the case. Thus, the analysis and its results confirm what we know about the mix of different approaches to designation in the set of studies and about pre- and posttest correlations and tell us little about which individual characteristics in reading-related skills at pretest allow prediction of response in a group of students who are nearly uniformly at risk prior to an intervention. In individual studies, this question is operationalized via the correlates of slope in growth curve analysis, via the correlates of the gain score in two-wave data collections, and by looking at the effect of baseline characteristics on posttest while controlling for pretest in other studies. We believe that a more comprehensible synthesis of these studies would aggregate these within-person effects.

Because the analytic framework used in this synthesis is novel and complicated, it should be accompanied by a stringent mapping of the analysis onto the research question and should possibly also be accompanied by a simulation demonstrating what the effects would be expected to look like in the presence and absence of effective interventions and individual characteristics that predict response (see Note 1).

Can Cohen's heuristic interpretation of d be used for dichotomized outcomes?

ESs were originally developed to create a common metric across studies that used different measures and scales. ES d , which was created to represent the magnitude of the difference between the means of an experimental group and a control group in randomized experiments, assumes that both groups are selected from the same population of individuals and that prior to intervention, these groups have the same population mean and standard deviation (SD). With randomization, the expected difference between the two means is 0 in the absence of a treatment effect. To the extent that the treatment separates the groups and reduces group overlap, d will deviate from 0. Cohen (1988) proposed that in the absence of prior data, when designing experiments for adequate power, d of 0.2, 0.6, and 0.8 might represent small, medium, and large effects of treatment. Tran et al. (2011) used Cohen's (1988) heuristic for interpreting the meaning of all pre- and posttest ESs and found many in the large range ($> .80$). In observational studies where groups are created by dividing a distribution, there is no differential treatment for the groups, so it is difficult to understand what the "treatment" ES represents. Also, the expected difference between the means is not 0. In fact, it can be shown analytically and via simulation that the ES for a variable being dichotomized will be about $d = 2.5$ if the cut point is between the 5th and 95th percentiles. To understand why, consider Figure 1. The cut point at the 16th percentile divides the normal distribution into two nonoverlapping groups. The smaller group's mean is $z = -1.52$, and the larger group's mean is $z = 0.29$. The SD s for these two groups are 0.46 and 0.79, respectively. When these two SD s are pooled and the 1.81 difference in z s is divided by the pooled SD , the ES is 2.4. Smaller values, such as the benchmarks suggested by Cohen, are possible when there is overlap between the two groups, which is what happens when two

¹Note

We would like to acknowledge an anonymous reviewer for this suggestion.

groups from the same population are created through randomization. This latter scenario is possible also if the variable in question has a very small correlation with the variable being dichotomized but is not possible when dichotomizing a distribution because the means of the observations above and below the cut point are of necessity quite different from each other: Substantial overlap between the two groups is required to move the means closer together. The ES is also larger because the variance within the two groups is substantially reduced through the dichotomization. ESs calculated in this way cannot be compared to the Cohen benchmarks because they are not on the same scale, that is, they do not have the same expected value and do not represent the same concept (the effect of a treatment versus the effect of being dichotomized).

In addition, and possibly because of conflating the d s from the two types of studies, Tran et al. (2011) argued that RTI does not close the gap between students at risk and typical students. Tran et al. reported treatment versus control ESs of 0.45 to 0.79 from randomized studies of children with learning disabilities (LD; Swanson & Hoskyn, 1998; Swanson & Sachse-Lee, 2000). They then compared these to the ESs representing the difference between children with LD in treatment groups with nondisabled children of the same grade or age. These groups are formed by dividing a continuum not by random assignment. The average ES in this comparison was $d = 0.97$. Tran et al. interpreted this larger average difference to mean that the intervention is not closing the gap. The problem is that they are comparing the average ESs over some set of variables in treatment-control studies where all ESs are expected to be 0 in the absence of a treatment effect with the average ES over some set of variables in studies where groups are formed by dichotomizing. In these latter studies, the effects are expected to be 0 if the variable of interest is uncorrelated with the dichotomizing variable but approximately 2.5 if the correlation is 1 and between 0 and 2.5 for other correlations. The expected size of the effect for any given variable depends on the correlation of that variable with the dichotomizing variable. The average effect overall then depends on the mix of variables included in the average and the strength of their correlations with the dichotomizing variable. If we assume that most variables reported in studies comparing LD and non-LD students are moderately to highly correlated with the dichotomizing variable, it is not surprising to find an average ES of 0.97. It is also not clear that comparing it to the average ES obtained from randomized trials has any meaningful interpretation, including that made by Tran et al., that RTI is not closing the gap.

Is phonological awareness a good predictor of response to intervention?

In a model where posttest ESs were predicted from pretest ESs as well as dummy-coded vectors representing both the category of posttest variable and methodological variables (such as method used to determine intervention response), Tran et al. (2011) reported that the beta weight for the PA coded vector was close to 0 and was nonsignificant. Does this mean that PA is not important for reading acquisition?

The issue is how to interpret the slopes of the centered, dummy coded variables in the context of the whole model. Although HLM was used for this analysis to account for the dependencies of multiple effects within the same studies, the issues in interpretation are the same as in any regression analysis. The magnitudes of the slopes of these variables are affected by issues of centering and the number of pre- and posttests within each category, as well as the number of studies where the groups were created at pretest versus posttest, which makes it difficult to precisely interpret the beta weights. However, we can make an attempt to understand the pattern of results by plotting the regression line from the HLM in the space of the mean pre- and posttest ESs (see Figure 2). The slope and intercept are taken from the fullest model (Tran et al., 2011, Table 4), which included the PA vector and a PA value for slope. The results from this model indicate that the predicted value of the posttest is 0.55

(intercept) when the pretest is 0. This finding is consistent with there being more studies in the mix where the designation of responder was made on the basis of a posttest variable so that the posttest effect tends to be larger than the pretest effect when the pretest is small. For each gain of 1 on the pretest ES, an increase of 0.26 on the posttest ES over the intercept is expected. This result is consistent with the pretest and posttest measures having positive correlations, which will result in the pre- and posttest ESs being positively related. On average then, and contingent on all of the other predictors in the model holding constant at their means, if the pretest ES = 1.0, our best prediction of the size of the posttest effect would be the intercept plus 0.26 times 1.0, or 0.81.

In Figure 2, the regression line implied by the HLM model is plotted, and the mean pre- and posttest values from Tables 2 and 3 (Tran et al., 2011) were overlaid on the plot of the regression line to elucidate the meaning of the beta weights presented in the HLM results. Their analysis indicates that pretest accounts for significant variation in the posttest scores but that there is significant remaining variance. The centered dummy variables represent a comparison of each category of posttest ES with all other categories, and they are included in the model to determine if they account for additional variance in the posttest ESs. To illustrate, the vector coded “reading comprehension” compares the posttest ESs on reading comprehension measures with the combined pool of RAN, PA, word identification, word attack, behavior, and so on. What the HLM weight of 0.80 for reading comprehension indicates is that the average posttest effect was larger for this category than would be predicted by the overall pretest-only model. As the plot shows, the negative weight for RAN indicates that its average posttest ES is smaller than would be predicted based on the entire set of pre- and posttest ESs. Categories with positive weights are above the regression line, indicating that their posttest effects were underpredicted by the pretest, and categories with negative weights are below the regression line and were overpredicted by the pretest. For PA, the beta weight is $-.04$. Thus, the regression line *slightly* overpredicts the average posttest ES for PA, but this difference was not large enough to meet the level of alpha ($p < .05$) specified for the test of significance. In fact, the predicted posttest ES for PA is 0.85 (or $0.55 + 0.26 \times 1.15$) when the actual posttest mean effect is 0.81. The regression weight of $b = -.04$ tells us that the predicted value is higher than the observed value by 0.04. It is coincidental that this beta weight exactly equals the difference between the predicted mean value and the observed mean value. In sum, the nonsignificant finding for PA tells us only that its effect is consistent with the overall set of effects and that allowing it a separate regression line with a different intercept (but the same slope) does not add substantially to the explanation of posttest effects. This result and the results from the other dummy codes entered into the model tell us only whether the effects for variable categories differ in level when the pretest is controlled, not which ones are important in predicting response to intervention.

Does this meta-analysis support the conclusion that RTI is not effective?

Tran et al. (2011) conclude that “response to intervention (RTI) conditions were not effective at mitigating learner characteristics related to pre-test conditions” (p. 283). The evidence presented in support of this assertion was that preand posttest ESs were substantially correlated. They also conclude that “unfortunately, the validity of RTI procedures, particularly in comparisons to other assessment approaches, has not been adequately established in the present synthesis of the literature” (p. 293). However, there was no place where the researchers established what the results should look like if RTI is “valid.” The synthesis of reading interventions by Scammacca et al. (2007) showed a narrowing of the gap between responders and *typically developing* children. Most RTI studies compare adequate responders with nonresponders, where *by definition* the gap must increase between these groups if they are defined at posttest, unless a ceiling or floor effect

is reached. Thus, these methods are not appropriate for determining the effectiveness of interventions based on RTI approaches, which is more appropriately determined via randomized control trials and syntheses of these studies. This strong conclusion strays beyond the data and the research questions, which involved predictors of intervention response. What is not clear is why Tran et al. (2011) did not conduct a much simpler meta-analysis of the correlations among pretests, posttests, and other individual characteristics measured at baseline. Moreover, it is not clear what advantage is gained by using this complicated ES model when a more simple approach would actually allow investigation of the research questions that seemed to motivate the study, including whether individual differences in pretest predicted response to intervention. There are substantive reasons to expect greater differences in adequate and inadequate responders after intervention than before, and larger posttest effects could easily be the result of the correlation of posttest variables with the variables used to dichotomize students into responders and nonresponders.

Acknowledgments

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by Grant P50 HD052117, Texas Center for Learning Disabilities, from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NICHD or the National Institutes of Health.

References

- Al Otaiba S, Fuchs D. Characteristics of children who are unresponsive to early literacy intervention: A review of the literature. *Remedial and Special Education*. 2002; 23:300–316.
- Allor JH, Fuchs D, Mathes P. Do children with or without lexical retrieval difficulties respond differently to instruction. *Journal of Learning Disabilities*. 2001; 34:264–275. [PubMed: 15499880]
- Cohen, J. *Statistical power analysis for the behavioral sciences*. (2nd ed.). San Diego, CA: Academic Press; 1988.
- Fletcher JM, Stuebing KK, Barth AE, Denton CA, Cirino PT, Francis DJ, Vaughn S. Cognitive correlates of inadequate response to intervention. *School Psychology Review*. 2011; 40:2–22.
- Menzies HM, Mahdavi JN, Lewis JL. Early intervention in reading: From research to practice. *Remedial and Special Education*. 2008; 29:67–77.
- Scammacca, N.; Roberts, G.; Vaughn, S.; Edmonds, M.; Wexler, J.; Reutebuch, CK.; Torgesen, JK. *Reading interventions for adolescent struggling readers: A meta-analysis with implications for practice*. Portsmouth, NH: RMC Research Corporation, Center on Instruction; 2007.
- Swanson HL, Hoskyn M. Experimental intervention research on students with learning disabilities: A meta-analysis of treatment outcomes. *Review of Educational Research*. 1998; 68:277–321.
- Swanson HL, Sachse-Lee C. A meta-analysis of single subject design intervention research for students with LD. *Journal of Learning Disabilities*. 2000; 33(2):114–136. [PubMed: 15505942]
- Tran L, Sanchez T, Arellano B, Swanson HL. A meta-analysis of the RTI literature for children at risk for reading disabilities. *Journal of Learning Disabilities*. 2011; 44(3):283–295. [PubMed: 21521870]
- Ukrainetz TA, Ross CL, Harm HM. An investigation of treatment scheduling for phonemic awareness with kindergartners who are at risk for reading difficulties. *Language, Speech, and Hearing Services in Schools*. 2009; 40:86–100.

Figure 1.
Demonstration of effect of forming groups by dichotomizing a normal distribution

Figure 2.
Mean effect sizes at pre and post time points. Regression line from HLM superimposed.
Size of bubble proportional to number of posttest ESs.