What counts as evidence for working memory training? Problems with correlated gains and dichotomization

Joe W. Tidwell, Michael R. Dougherty, Jeffrey R. Chrabaszcz, Rick P. Thomas & Jorge L. Mendoza

Psychonomic Bulletin & Review

ISSN 1069-9384

Psychon Bull Rev
DOI 10.3758/s13423-013-0560-7

Springer

Psychon Bull Rev
DOI 10.3758/s13423-013-0560-7

THEORETICAL REVIEW

What counts as evidence for working memory training? Problems with correlated gains and dichotomization

Joe W. Tidwell · Michael R. Dougherty · Jeffrey R. Chrabaszcz · Rick P. Thomas · Jorge L. Mendoza

Abstract The question of whether computerized cognitive training leads to generalized improvements of intellectual abilities has been a popular, yet contentious, topic within both the psychological and neurocognitive literatures. Evidence for the effective transfer of cognitive training to nontrained measures of cognitive abilities is mixed, with some studies showing apparent successful transfer, while others have failed to obtain this effect. At the same time, several authors have made claims about both successful and unsuccessful transfer effects on the basis of a form of responder analysis, an analysis technique that shows that those who gain the most on training show the greatest gains on transfer tasks. Through a series of Monte Carlo experiments and mathematical analyses, we demonstrate that the apparent transfer effects observed through responder analysis are illusory and are independent of the effectiveness of cognitive training. We argue that responder analysis can be used neither to support nor to refute hypotheses related to whether cognitive training is a useful intervention to obtain generalized cognitive benefits. We end by discussing several proposed alternative analysis techniques that incorporate training gain scores and argue that none of these methods are appropriate for testing hypotheses regarding the effectiveness of cognitive training.

Electronic supplementary material The online version of this article (doi:10.3758/s13423-013-0560-7) contains supplementary material, which is available to authorized users.

J. W. Tidwell · M. R. Dougherty · J. R. Chrabaszcz
University of Maryland, College Park, MD, USA

R. P. Thomas · J. L. Mendoza
University of Oklahoma, Norman, OK, USA

J. W. Tidwell (✉) · M. R. Dougherty (✉)
Department of Psychology, University of Maryland,
1147 Biology/Psychology Building, College Park, MD 20742, USA
e-mail: jtidwell@umd.edu
e-mail: mdougher@umd.edu

Published online: 05 December 2013

Keywords Cognitive training · Human memory and learning · Individual differences · Memory capacity · Statistical inference

A much-debated question within psychology is whether general intellectual abilities can be improved through cognitive training (Chein and Morrison 2010; Jaeggi, Buschkuehl, Jonides, and Perrig 2008; Jaeggi, Buschkuehl, Jonides, and Shah 2011; Shipstead, Redick, and Engle 2012). Similar to the way that fitness training leads to generalized health benefits, some have hypothesized that cognitive training can lead to generalized cognitive benefits, such as increased general fluid intelligence and working memory. If core cognitive abilities like working memory can be improved, then any task that draws on these core cognitive abilities should also show a performance boost. Because working memory processes are instrumental to a wide range of real-world activities (Engle 2002), improving core working memory abilities through training could be enormously beneficial.

The usefulness of cognitive training depends on the extent and nature of transfer of improvement on a trained ability to tasks and abilities that are not directly trained. Researchers have made the distinction between two types of transfer: near and far. Near transfer occurs when the transfer task is structurally similar to the training task. Far transfer occurs when the transfer task is structurally different than the training task (Hussey and Novick 2012; Morrison and Chein 2011). The potential for far transfer has garnered much attention recently in both the scientific literature and the general media (Hurley 2012; Reddy 2013). Given the potentially enormous societal benefit of cognitive training and the potential marketplace for effective products, estimated to be $6 billion by 2020 (Sukel 2013), it is important that the evidence be accurately portrayed to the public.

Data regarding the effectiveness of working memory training are mixed. Several studies have reported statistically

significant effects of training on a variety of transfer measures (Jaeggi et al. 2008; Klingberg 2010; Schmiedek, Lövdén, and Lindenberger 2010), whereas others have failed to find evidence of transfer (Owen et al. 2010; Redick et al. 2013; Sprenger et al. 2013). Critical reviews of the literature have suggested that some studies reporting successful transfer have suffered from methodological flaws (Shipstead et al. 2012), whereas reviews of experimental results have been both pessimistic (Melby-Lervåg and Hulme 2013) and optimistic (Morrison and Chein 2011) with regard to successful transfer. The debate over the degree to which cognitive training is effective has even landed the topic its own webpage, at www.psychfiledrawer.org. The purpose of the present article is to examine claims regarding the effectiveness of working memory training that are based on a specific type of statistical analysis—a form of the responder analysis. Setting aside the question of whether working memory training is effective, we argue that statistical inferences based on responder analysis, as it has been used in the working memory training literature, are not appropriate tests for hypotheses about working memory training.

Types of responder analysis

The gold standard for assessing transferability is to illustrate, within a controlled trial, that cognitive training leads to improvements on a transfer task relative to a placebo control-training task. In the absence of finding group-level differences between training and control, however, some researchers have turned to a particular form of the responder analysis. For illustrative purposes, we use *N*-back and Raven's Progressive Matrices (henceforth denoted as *RPM*) as canonical examples of training and transfer tasks, though the critique generalizes to any training and transfer task, or for that matter to any experiment in which an intervention is assessed over time. *N*-back is a working memory task wherein participants are presented a series of stimuli and must decide whether the current stimulus matches the stimulus presented *n* positions back in the series. Working memory training researchers often use an adaptive form of *N*-back, in which the *n*-level is continuously matched to the participant's performance (Jaeggi et al. 2008). RPM is a widely used nonverbal measure of general intelligence (Raven 2000).

Two types of responder analyses have appeared in the broader literature. In the first, the researcher defines a threshold on an outcome variable and compares proportional changes of subjects above or below that threshold between treatment and control conditions following an intervention (Senn and Julious 2009). In the second, the researcher defines performance groups on one variable, for example by dichotomizing difference scores between a first and last *N*-back training session with a median split, and then compares performance between these groups on a second set of correlated difference scores, for example a pre/post measure of intelligence. Variants of the second form, which we will demonstrate includes a test of the correlation between training and transfer difference scores, will simply be referred to as "responder analysis" for the remainder of this article. This form of responder analysis has been used on numerous occasions within the working memory training literature to test for the existence of transfer from working memory training (Jaeggi et al. 2011; Kundu, Sutterer, Emrich, and Postle 2013; Novick, Hussey, Teubner-Rhodes, Harbison, and Bunting 2013; Redick et al. 2013; Rudebeck, Bor, Ormond, O'Reilly, and Lee 2012; Thompson et al. 2013; Zinke et al. 2013), and in other recent studies outside this literature (Miaskowski et al. 2007).

As we demonstrate below, responder analysis is a proxy for testing the correlation between training and transfer difference scores. However, correlated difference scores can arise for many reasons, including simply as a function of assessing correlated measures across time. Because responder analysis cannot distinguish amongst the various factors that might lead to correlated difference scores, it cannot be used to test hypotheses about training effectiveness.

Theoretical rationale for responder analysis

The rationale for using responder analysis in the working memory training literature is based on the hypothesis that the effects of transfer might be moderated by individual differences in how much participants have gained on the training task (Jaeggi et al. 2011): Whereas some participants improve on a training task, others may not improve at all. Only those who benefited from training (responders) are theoretically expected to exhibit gains on the transfer task. Arguably, participants who fail to "respond" to training (nonresponders) may mask the true transfer effects obtained by responders.

The idea that transfer might be moderated by training gains is consistent with the hypothesis that people differ in their capacities for plasticity. Some have speculated that individual differences in cognitive training effectiveness may arise from neurological mechanisms or genetic polymorphisms that enable neural plasticity (McNab et al. 2009). The hypothesis that individuals differ in their capacities for plasticity suggests that training gains should be correlated with gains on the transfer task, a finding that has been reported on multiple occasions (Chein and Morrison 2010; Jaeggi et al. 2011; Schweizer, Grahn, Hampshire, Mobbs, and Dalgleish 2013; Zinke et al. 2013). For example, in a recent article in *Proceedings of the National Academy of the Sciences*, Jaeggi et al. (2011) remarked that the extent of transfer was "critically dependent on the amount of the participants' improvement on the WM task" (p. 10083). Using responder analysis, they showed that responders and nonresponders differed significantly in the

amounts of transfer: Those who obtained the large training (*N*-back) gains showed significantly larger transfer (RPM) gains than did those who had smaller training gains. Although Jaeggi et al. (2011) failed to find a significant difference between treatment and control performance on the transfer task in an ANOVA framework, they concluded that the significant responder analysis result provided evidence that trainability moderated the degree to which transfer occurred. Other authors have drawn similar conclusions on the basis of correlated gain scores (Jaeggi et al. 2011; Kundu et al. 2013; Novick et al. 2013; Rudebeck et al. 2012), whereas Thompson et al. (2013) and Redick et al. (2013) used a nonsignificant correlation between gain scores to support their conclusion that training failed to transfer.

What does responder analysis really tell us?

The goal of many working memory training studies is to address the question of whether training on one cognitive task generalizes to improved performance on other tasks. Responder analysis is sometimes used to test whether how much training occurs moderates this effect. It seems reasonable to assume that the degree of transfer would be dependent on how much improvement a participant showed on the training task. However, testing this assumption statistically with responder analysis is problematic: Responder analysis tells one only whether selecting a subgroup from the training task obtains a similar subgroup on the transfer task, independent of the group performance on either the training or the transfer tasks. This is equivalent to determining whether training and transfer difference scores are correlated. Since it lacks an appropriate control comparison and is independent of subjects' performance, responder analysis cannot provide any evidence that the training task drives this correlation. Although the reason that responder analysis cannot be used in this context is straightforward and follows from the definition of correlation, it is nonintuitive, especially in light of the (sensible) theory that training gains should moderate transfer gains. Thus, in the rest of the article, we will describe responder analysis in detail, further elucidate its problems, and evaluate the viability of possible solutions and alternative methods to assess the effectiveness of working memory training.

A common form of responder analysis is a 2 × 2 repeated measures analysis of variance (ANOVA) that defines performance on the transfer task as a continuous dependent variable, Mean Performance on the first and last, or first two and last two, training sessions as a dichotomous within-subjects factor, and Responder Group (responder, nonresponder) as a between-subjects factor. Typically, researchers interpret a significant interaction between session and group as evidence that the degree of transfer is moderated by the degree to which one responds to the training task. However, as we demonstrate below, this analysis is solely a function of the correlation between training and transfer task differences and is independent of the actual training effectiveness. Therefore, no particular pattern of means for the training and transfer tasks is either necessary or sufficient to obtain a significant difference between the responder and nonresponder means on the transfer task.

One of the fundamental problems with responder analysis involves selection. When the transfer task is correlated with the training task, selecting a subset of participants who perform highly on the training task entails selection on the transfer task. In the case in which difference scores are positively correlated, this entails that responders on *N*-back will necessarily perform higher on RPM than will nonresponders. This selection process is similar to the researcher eliminating "bad" subjects for failing to complete the experiment as expected after looking at the dependent variable, except that responder analysis goes one step further: Rather than comparing the remaining "good" subjects to an appropriate baseline, responder analysis compares performance between the "good" and "bad" subgroups. This is shown in Fig. 1, which plots an ellipsoid representing a case in which gains on a training task (*x*-axis) are dichotomized by a median split. Projecting the mean of the transfer task differences on the *y*-axis yields a similar split on the transfer task: Responders on training show larger gains on transfer. In order to make claims that any differences between responders and nonresponders (or control participants) are due to training, one must establish that the correlation between the difference scores is due to training.

Properties of responder analysis

In responder analysis, one first computes difference scores on both the training and transfer tasks and then makes a comparison between these scores, most often after dichotomizing training

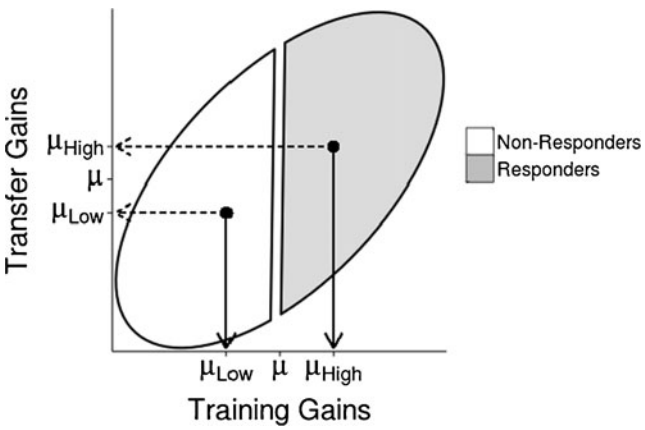

Fig. 1 Ellipse corresponding to a 95%-probability region for the bivariate normal distribution $N(\mu, \Sigma)$, where $\mu = \binom{0}{0}$ and $\Sigma = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$ dichotomized with a median split. Since the distribution is symmetric, the mean and median are identical

task difference scores to obtain large (responder) and small (nonresponder) training gain subgroups. We will demonstrate below that although there are seemingly many different ways to analyze these difference scores—for example, through t tests, regression, ANOVA, and repeated measures ANOVA—all of these methods evaluate the same null hypothesis: whether the correlation between the difference scores is zero. Since this correlation is independent of any mean shift in performance for either task, it is uninformative with respect to whether training leads to transfer gains.

Define X_1, X_2 as the first and last training session scores, and Y_1, Y_2 as the first and last transfer task scores. Let the difference between the pre- and posttest transfer task scores (d_Y), the difference between the first and last training task scores (d_X), and the distribution of training difference scores dichotomized at some value c ($d_{X\text{split}}$) be defined as follows:

$$d_Y = Y_2 - Y_1, \tag{1}$$

$$d_X = X_2 - X_1, \tag{2}$$

$$d_{X\text{split}} = \begin{cases} E(d_X > c) \text{ for } d_X > c. \\ E(d_X \le c) \text{ for } d_X \le c. \end{cases} \tag{3}$$

Let the mean of responder (μ_{resp}) and nonresponder ($\mu_{\sim\text{resp}}$) performance on the transfer task be defined as

$$\mu_{\text{resp}} = E\left(d_Y \,\middle|\, d_{X\text{split}} > c\right), \tag{4}$$

$$\mu_{\sim\text{resp}} = E\left(d_Y \,\middle|\, d_{X\text{split}} \le c\right). \tag{5}$$

Responder analysis tests whether responders exhibit different transfer task gains than do nonresponders, which can be represented with the null hypothesis $\mu_{\text{resp}} - \mu_{\sim\text{resp}} = 0$. This is equivalent to testing whether $d_{X\text{split}}$ is a significant predictor of d_Y in the following linear model and corresponding null hypothesis:

$$d_Y = b_0 + b_1 d_{X\text{split}} + e, \tag{6}$$

$$H_0 : b_1 = 0. \tag{7}$$

H_0 does test for a difference in means; however, this difference is a necessary result of the correlation between d_Y and d_X. This may be more easily seen by looking at the F statistic to test H_0:

$$F_{1,\,d.f.} = \frac{R^2}{1-R^2} d.f. = \frac{r^2_{d_X\ split\ d_Y}}{1-r^2_{d_X\ split\ d_Y}} d.f. \tag{8}$$

Since $d_{X\text{split}}$ has only two values, the F statistic in Eq. 8 is identical to the F statistic for a one-way ANOVA using $d_{X\text{split}}$ as a binary Grouping factor (responder, nonresponder) and d_Y as the dependent variable, and will yield the same p value as a t test between responders and nonresponders on d_Y. Furthermore, Eq. 8 is also identical to the F statistic for the interaction between the Grouping factor and a two-period session variable (Time1, Time2) in a repeated measures ANOVA (Dimitrov and Rumrill 2003; Huck and McLean 1975).

Calculating a one-way ANOVA or t test of responder group on transfer difference scores, testing the interaction of session and responder group in a repeated measures ANOVA, and conducting a regression of the mean responder group training difference scores on transfer difference scores all reduce to a test of the correlation between the dichotomized difference scores on the training task ($d_{X\text{split}}$) and difference scores for the transfer task (d_Y). Furthermore, since dichotomizing one of two continuous correlated variables to yield a point-biserial correlation generally attenuates the original Pearson correlation (MacCallum, Zhang, Preacher, and Rucker 2002; McClelland and Irwin 2003), this is a less efficient method to discern the relationship between the original training task difference scores (d_X) and d_Y than would be simply regressing d_Y on d_X versus $d_{X\text{split}}$.

By definition, the correlation between the difference scores is a function of the covariance between the difference scores and the variances of the difference scores (Eq. 9). The covariance between the differences can be expanded as a function of the cross-session covariances, $\text{Cov}(X_1Y_2)$ and $\text{Cov}(X_2Y_1)$, and the within-session covariances, $\text{Cov}(X_1Y_1)$ and $\text{Cov}(X_2Y_2)$ (Eq. 10). Accordingly, one will obtain a positive correlation between the difference scores—that is, a "responder" effect—in any instance where the sum of the cross-session covariances is less than the within-session covariances

$$r_{d_X d_Y} = \frac{\text{Cov}(d_X d_Y)}{\sqrt{\text{Var}(d_X)\text{Var}(d_Y)}}, \tag{9}$$

$$\begin{aligned} \text{Cov}(d_X d_Y) &= \text{Cov}(X_2 - X_1, Y_2 - Y_1) \\ &= \text{Cov}(X_1Y_1) + \text{Cov}(X_2Y_2) - \text{Cov}(X_1Y_2) - \text{Cov}(X_2Y_1). \end{aligned} \tag{10}$$

Since the magnitude of a correlation is invariant to a linear transformation of either variable, a significant responder analysis provides no support for any particular pattern of means for d_X and d_Y. One could obtain the same correlation, and therefore the same F statistic, independent of whether gains occur in either the training or transfer task. This would be true even in the case in which losses occur across sessions for both the training and transfer tasks. Again, the crucial insight from

Eq. 10 is that since covariances (and therefore correlations) are devoid of means, a significant responder effect does not address the question of whether training led to the transfer task gains, or whether transfer gains were moderated by training effectiveness.

Monte Carlo experiments

To illustrate that the results obtained through responder analysis are invariant to subjects' performance on training and transfer tasks, five Monte Carlo experiments representing potential working memory training study outcomes were conducted. All experiments were programmed using R version 2.15.3 and the following R packages: data.table, plyr, doMC, MASS, and reshape2 (Dowle, Short, and Lianoglou 2013; R Development Core Team 2013; Revolution Analytics 2013; Venables and Ripley 2002; Wickham 2007). The source code used to generate the data is included in the supplemental materials.

Four correlated random variables were sampled from a multivariate normal distribution with the covariance matrix defined in Table 1, but with unique mean vectors for each simulation. For illustrative purposes, we used N-back training and transfer to Raven's Progressive Matrices (RPM) to represent the canonical working memory training experiment, although our analysis generalizes to any other set of training and transfer tasks. Two of the variables represent scores for a first and a last N-back training task (N-Back$_1$, N-Back$_2$), and the remaining two variables represent scores for a pre- and a posttraining RPM transfer task (RPM$_{Pre}$, RPM$_{Post}$). Each mean vector represents a potential pattern of training and transfer effects: (1) improvement on both N-back and RPM, (2) improvement on N-back but not on RPM, (3) no improvement on N-back but improvement on RPM, (4) no improvement on either task, and (5) decline on both tasks (Table 2). The sample size was set at $N = 50$ for all experiments, and each experiment was replicated 100,000 times.

Our dependent variables for these experiments were the probability of obtaining a significant responder analysis, the correlation between responder group and transfer task difference scores, and the F statistic for the linear regression predicting transfer task difference scores from responder groups (Eq. 11). Given our preceding analysis, we predicted that these dependent measures would be invariant to the pattern of means for both the training and transfer tasks. Using long-established properties of sample correlations (Fisher 1915) and the values from the variance–covariance matrix in Table 1, we also calculated analytic predictions for these measures, included as the last row in Table 3.

Table 1 Covariance matrix for data sampled from a multivariate normal distribution

Measure	1	2	3	4
1. N-Back$_1$	1			
2. N-Back$_2$	.6	1		
3. RPM$_{Pre}$	.3	.15	1	
4. RPM$_{Post}$	.15	.3	.6	1

RPM = Raven's Progressive Matrices

For the experimental data, the difference scores for N-back and RPM were calculated following data generation for each replication. "Subjects" who obtained higher difference scores than the median N-back difference score were classified as responders, all others were classified as nonresponders, and an F statistic was then calculated for the regression:

$$\text{RPM}_{\text{Gains}} = b_0 + b_1 \text{N-Back}_{\text{Split}} + e, \tag{11}$$

where RPM$_{Gains}$ is the difference between the pre and post RPM scores, and N-Back$_{Split}$ is a median split dichotomization of N-Back$_{Gains}$, the difference between the first and last N-back sessions. The mean probability of rejecting the null hypothesis H_0: $b_1 = 0$, as well as the median F value, correlation between N-Back$_{Gains}$ and RPM$_{Gains}$, and correlation between N-Back$_{Split}$ and RPM$_{Gains}$ ($r_{N_{split}R_d}$) are included in Table 3.

The probabilities of rejecting the null hypothesis, the correlations between responder group and transfer task difference scores, and the F statistics for the regression in Eq. 11 are nearly identical across the five different scenarios, independent of whether any gains occurred across N-back and/or RPM sessions (Simulations 1–4), and even when scores for both measures declined (Simulation 5). Furthermore, the mean of the dependent variables across scenarios matches our analytic predictions, which were dependent only on the variance–covariance matrix in Table 1. It is clear that finding a statistically significant difference between responders and nonresponders on the transfer task tells one nothing regarding the effectiveness of working memory training. Furthermore, for Experiments 1–4, responders always showed the largest gain on the transfer task, and in Experiment 5, in which a mean decrease on transfer occurred from pre to post, they showed the smallest losses, with identical effect sizes across all five experiments. Thus, independent of whether or not true gains occurred, responders always came out looking better than nonresponders. Again, this result is a product of the covariance matrix, and does not count as evidence for successful transfer.

Three related objections to this critique

Some have argued that whereas responder analysis may be a test of the correlation between training and transfer task difference scores, the conjunction of this correlation with a pattern of difference scores wherein responders demonstrate

Table 2 Mean vectors for data sampled from multivariate normal distribution and difference scores

		Mean Vector				Difference Scores	
Exp. #	Scenario	$N\text{-Back}_1$	$N\text{-Back}_2$	RPM_{Pre}	RPM_{Post}	$N\text{-Back}_{Gains}$	RPM_{Gains}
1	Training & transfer gains	1	5	1	2	4	1
2	Training gains only	1	5	1	1	4	0
3	Transfer gains only	1	1	1	2	0	1
4	No training or transfer gains	1	1	1	1	0	0
5	Training and transfer losses	5	1	2	1	−4	−1

RPM = Raven's Progressive Matrices

greater gains on the transfer task than nonresponders provides evidence that training effectively increased transfer task performance for those who responded most to training. However, as demonstrated previously, this difference in means is a function of dichotomization and the form of the variance–covariance matrix for the variables. That is, the responder groups do not actually exist; they are created by the researcher post hoc, and are based on the correlation between difference scores. This can be easily discerned from Fig. 1, in which the difference scores are positively correlated; dichotomizing transfer gains by performance on the training task necessarily yields a responder group with greater transfer task differences than are present in the nonresponder group. This is true, independent of any linear transformation applied to either training or transfer gains. This was also demonstrated by the Monte Carlo experiments: Responder analysis is independent of mean performance on either the training or transfer tasks.

Some have also suggested to the authors that even if responder analysis is a proxy for testing the correlation between training and transfer difference scores, a positive correlation between these scores is itself evidence that gains on the training task bring about gains on the transfer task. In the case of working memory training, this means that subjects who increased training task performance across training sessions improved some cognitive ability that also contributes to general fluid intelligence, which then causes their transfer task scores to increase between the pre- and posttest administrations. As is clear from Eqs. 9 and 10, a positive correlation between the training and transfer difference scores is obtained when the sum of the cross-session covariances is less than the sum of the within-session covariances. However, one should expect this pattern in any experimental design that includes correlated tasks measured across time. Since correlations between tasks—for example, test–retest reliability—tend to decrease as the time between measurements increases, one should expect more proximal tasks to be more highly correlated than more distal tasks. In the case of working memory training experimental designs, this implies that even under the null hypothesis one should expect a priori for the cross-session covariances to be less than the within-session covariances. This does not, however, imply that within this paradigm one will always observe correlated difference scores, only that a training effect is unnecessary to obtain a correlation between the differences.

A related objection that has been raised is that, since difference scores will not necessarily be correlated—for example, as was found by Thompson et al. (2013)—the most likely explanation for the existence of a correlation between difference scores is an effect of the training task. This objection, however, ignores the fact that the model tested with responder analysis is absent a main-effect term, and that the means that are being compared are created, *not observed*, by the researcher. Thus, any correlation between difference scores, be it positive or negative, and any pattern of means within the responder analysis can be obtained independent of the effectiveness or ineffectiveness of training. With regard to whether training leads to transfer, a positive correlation

Table 3 Results of Monte Carlo simulations

Exp. #	Scenario	$P(p < .05)$	$r_{N_{split}R_d}$	F
1	Training and transfer gains	69%	.30	4.81
2	Training gains only	69%	.30	4.86
3	Transfer gains only	70%	.30	4.82
4	No training or transfer gains	69%	.30	4.85
5	Training and transfer losses	69%	.30	4.78
Mean of Experiments 1–5		69%	.30	4.82
Analytically derived prediction		69%	.30	4.82

$r_{N_{split}R_d}$ = correlation between *N*-back and Raven's difference scores, with the *N*-back difference scores median-split into responders and nonresponders

between difference scores, combined with a pattern of means wherein responders show greater gains on the transfer task than do nonresponders, is just as uninformative as a negative correlation, wherein transfer for nonresponders will be greater than that of responders. In both cases—and, indeed, any case—the means of the responder and nonresponder groups are a necessary consequence of dichotomizing correlated variables (Fig. 1). One could obtain these correlations whether the training improved, attenuated, or had no effect whatsoever on transfer task performance. Simply put, the pattern of means on the transfer task based on the dichotomization of training gains is theoretically meaningless, because they do nothing more than redescribe the correlation between gain scores.

Proposed solutions that fail to address the problem

The most straightforward approach to studying the effects of cognitive training is to rely on experimental methods and random assignment to groups. However, two modifications to responder analysis have been suggested to the authors as potential ways to exploit training gains to identify responders and nonresponders: (1) compare responders and nonresponders with a control group, rather than with each other, and (2) dichotomize participants in the control group using gains on a placebo-training task. Neither of these options is appropriate. The former is uninformative because it compares the performance of extreme groups from the training condition to average performance from the control condition. Even if the null were true, comparing extreme groups would yield variation between the conditional means and the control mean.

Comparing training condition responders and nonresponders with a similarly dichotomized—for example, median-split—placebo control group provides no better solution. Although this allows one to compare two sets of conditional means, the training and control groups would be conditioned on different tasks. For example, if one were to compare mean change on RPM for responders on the N-back task with responders on a visual search task, this would amount to comparing $\mu(d_{\text{RPM}} \mid d_{\text{N-Back}} > c_1)$ with $\mu(d_{\text{RPM}} \mid d_{\text{search}} > c_2)$, where c_1 is the median N-back difference score and c_2 is the median visual search task difference score. Because the conditioning events are different between the two sets of conditional means, two different population parameters are being estimated, and they cannot be directly compared.

These concerns aside, we suggest that the use of training data in the analysis of transfer effects should be avoided altogether. First, it is unclear what the measurement properties are of an adaptive training task and whether these measurement properties change as a function of increasing task difficulty. For example, does N-back measure the same thing in the first training session as it does in the last training session? We suspect not, given that people routinely perform at low levels of N in the first few sessions, and at higher levels of N in subsequent sessions. Also, if people learn strategies over time, later trials will be more reflective of strategy use, as opposed to processing capabilities. Alternatively, it is theoretically possible that an individual's asymptotic level of performance reached in later training sessions would reflect his or her latent ability, whereas the apparently poorer performance in initial training sessions simply indicates that the task has yet to adapt to the individual's latent ability level.

Second, because the training task is adaptive within the training session for each individual participant, any derived scoring function will be contaminated by task variability. That is, not only do latent abilities vary across people, but so also does the difficulty and structure of the task. This is similar to administering a psychometric test in which items and item difficulty are random variables that differ across both individuals and time points, and in which the items have not been normed for difficulty or validity. Both changes in measurement properties and variability across individuals in terms of the level achieved in the adaptive task would likely drive down the cross-session validities and, as a consequence, set up the conditions needed to obtain a significant responder effect, as is suggested by Eq. 10.

The discussion above suggests that one potential methodological approach to examining the moderating role of training effectiveness would be to use standardized assessments of the training tasks at pre and post for both a training and control condition, and to use gains on the standardized assessments as a proxy for training gain. However, one would still have to establish that the construct validity of the assessment did not change as a result of training. For example, if participants develop strategies as a result of practicing N-back, then it is likely that performance on the pretest version of N-back would be driven by different processes than performance at posttest.

What counts as evidence?

Returning to the question of what counts as evidence in working memory training, we argue that evidence for successful transfer requires that researchers address at least three main challenges. First, training effects must be evaluated relative to a proper control condition (Shipstead et al. 2012). Although there is much debate over what constitutes a proper control, the success of any intervention can only be gauged by comparison to a control condition (for a discussion of the necessary properties of a proper placebo control group that is particularly relevant to working memory training, see Boot, Simons, Stothart, and Stutts 2013).

Second, the study must be able to differentiate between transfer due to a change in the underlying processes as opposed to effects due to shared stimulus or task characteristics, as well as effects due to strategy shifts versus changes in cognitive ability (Gibson, Gondoli, Johnson, and Robison 2013). What

constitutes a shared stimulus or task characteristic is defined by the subject, not the researcher. Imagine a subject who is trained on a version of *N*-back that uses letters as the stimuli, but is administered a working memory span task that uses words. These two classes of stimuli are only different in as much as the subject treats the stimuli differently. If the participant treated the letters as words (e.g., by assigning a unique word to each letter to improve memorability), then the stimuli would be functionally identical. Short of measuring strategy, there is no obvious way to discern transfer effects that are due to the mnemonic strategy versus those due to improvement of a core process (Sprenger et al. 2013).

Third, the evidence should be evaluated within a Bayesian framework, so that it can be assessed relative to both the alternative hypothesis that working memory training works and the null hypothesis that fluid abilities are unchangeable. This is more than a generic endorsement of Bayesian methods. In working memory training, what researchers define as the null hypothesis is not simply the absence of a training effect, but is a competing hypothesis. Not only is the invariance hypothesis a valid and plausible possibility, historically it has been assumed that fluid abilities such as intelligence and working memory were immutable in adulthood. Thus, the domain of working memory training is a perfect example of when Bayesian methods are most useful: when endorsement of the null model is of both theoretical and practical importance (Gallistel 2009; Rouder, Speckman, Sun, Morey, and Iverson 2009; Wetzels et al. 2011). Although a variety of statistically significant effects have been identified in the literature, not all significant effects are created equal when evaluated within a Bayesian framework. For example, Sprenger et al. (2013) used Bayes-factor analysis to analyze the training effects from three studies and showed that the evidence overwhelmingly supported the null hypothesis at a ratio of over 22:1, even after including studies that reported significant effects. This finding is consistent with recent meta-analyses indicating that transfer effects are weak at best, and potentially limited to special populations (Melby-Lervåg and Hulme 2013; Protzko, Aronson, and Blair 2013).

Conclusions

The idea that cognitive abilities can be trained in a manner similar to physical strength is quite compelling, and the potential benefits of such training effects would be far-reaching and transformative. Although responder analysis has been used relatively infrequently in the literature, its use is becoming increasingly common. Studies using this method have appeared in highly reputable journals and have already been cited over 170 times.[1] Nevertheless, responder analysis cannot be used to argue either whether cognitive training leads to transfer (Chein and Morrison 2010; Jaeggi et al. 2011; Novick et al. 2013; Rudebeck et al. 2012) or whether it does not (Redick et al. 2013; Thompson et al. 2013). Although the question of whether working memory training is effective at improving intelligence remains debated, statistical conclusions based on responder analysis or correlated gain scores should not be considered as part of this debate.

[1] Google Scholar citations as of 25 October, 2013.

Author Note This work was supported by the Office of Naval Research, Grant No. N000141010605, awarded to M.R.D. The authors thank Thomas Wallsten for comments on a prior version of the manuscript.

References

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science, 8,* 445–454.

Chein, J. M., & Morrison, A. B. (2010). Expanding the mind's workspace: training and transfer effects with a complex working memory span task. *Psychonomic Bulletin & Review, 17,* 193–199. doi:10.3758/PBR.17.2.193

Dimitrov, D. M., & Rumrill, P. D. (2003). Pretest-posttest designs and measurement of change. *Work, 20,* 159–165.

Dowle, M., Short, T., & Lianoglou, S. (2013). data.table: Extension of data.frame for fast indexing, fast ordered joins, fast assignment, fast grouping and list columns (R package version 1.8.8) [Computer software]. Retrieved from http://cran.*r*-project.org/web/packages/data.table/index.html

Engle, R. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11,* 19–23. doi:10.1111/1467-8721.00160

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10,* 507–521.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116,* 439–453. doi:10.1037/a0015251

Gibson, B. S., Gondoli, D. M., Johnson, A. C., & Robison, M. K. (2013). Recall initiation strategies must be controlled in training studies that use immediate free recall tasks to measure the components of working memory capacity across time. *Child Neuropsychology*. doi:10.1080/09297049.2013.826185

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest–posttest design: A potentially confusing task. *Psychological Bulletin, 82,* 511–518. doi:10.1037/h0076767

Hurley, D. (2012, Nov 4). The brain trainers. *New York Times*, p. ED18

Hussey, E. K., & Novick, J. M. (2012). The benefits of executive control training and the implications for language processing. *Frontiers in Psychology, 3,* 158. doi:10.3389/fpsyg.2012.00158

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences, 105,* 6829–6833. doi:10.1073/pnas.0801268105

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences, 108,* 10081–10086. doi:10.1073/pnas.1103228108

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences, 14,* 317–24. doi:10.1016/j.tics.2010.05.002

Kundu, B., Sutterer, D. W., Emrich, S. M., & Postle, B. R. (2013). Strengthened effective connectivity underlies transfer of working memory training to tests of short-term memory and attention. *Journal of Neuroscience, 33,* 8705–8715. doi:10.1523/JNEUROSCI.5565-12.2013

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40. doi:10.1037/1082-989X.7.1.19

McClelland, G., & Irwin, J. (2003). Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research, XL,* 366–371.

McNab, F., Varrone, A., Farde, L., Jucaite, A., Bystritsky, P., Forssberg, H., & Klingberg, T. (2009). Changes in cortical dopamine D1 receptor binding associated with cognitive training. *Science, 323,* 800–802. doi:10.1126/science.1166102

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology, 49,* 270–291. doi:10.1037/a0028228

Miaskowski, C., Dodd, M., West, C., Paul, S. M., Schumacher, K., Tripathy, D., & Koo, P. (2007). The use of a responder analysis to identify differences in patient outcomes following a self-care intervention to improve cancer pain management. *Pain, 129,* 55–63. doi:10.1016/j.pain.2006.09.031

Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review, 18,* 46–60. doi:10.3758/s13423-010-0034-0

Novick, J. M., Hussey, E., Teubner-Rhodes, S., Harbison, J. I., & Bunting, M. F. (2013). Clearing the garden-path: Improving sentence processing through cognitive control training. *Language and Cognitive Processes*, 1–44. doi:10.1080/01690965.2012.758297

Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., Howard, R.J., & Ballard, C.G. (2010). Putting brain training to the test. *Nature, 465,* 775–8. doi:10.1038/nature09042

Protzko, J., Aronson, J., & Blair, C. (2013). How to make a young child smarter: Evidence from the database of raising intelligence. *Perspectives on Psychological Science, 8,* 25–40. doi:10.1177/1745691612462585

R Development Core Team. (2013). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41,* 1–48. doi:10.1006/cogp.1999.0735

Reddy, S. (2013, May 14). When computer games may keep the brain nimble. *Wall Street Journal*, p. D1.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., Kane, M. J., & Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology. General, 142,* 359–79. doi:10.1037/a0029082

Revolution Analytics. (2013). doMC: Foreach parallel adaptor for the multicore package (R package version 1.3.0) [Computer Software]. Palo Alto, CA. Retrieved fromhttp://cran.r-project.org/web/packages/doMC/index.html

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. doi:10.3758/PBR.16.2.225

Rudebeck, S. R., Bor, D., Ormond, A., O'Reilly, J. X., & Lee, A. C. H. (2012). A potential spatial working memory training task to improve both episodic memory and fluid intelligence. *PLoS ONE, 7,* e50431. doi:10.1371/journal.pone.0050431

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: Findings from the COGITO study. *Frontiers in Aging Neuroscience, 2*(27), 1–10. doi:10.3389/fnagi.2010.00027

Schweizer, S., Grahn, J., Hampshire, A., Mobbs, D., & Dalgleish, T. (2013). Training the emotional brain: Improving affective control through emotional working memory training. *Journal of Neuroscience, 33,* 5301–5311. doi:10.1523/JNEUROSCI.2593-12.2013

Senn, S., & Julious, S. (2009). Measurement in clinical trials: A neglected issue for statisticians? *Statistics in Medicine, 28,* 3189–3209. doi:10.1002/sim.3603

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin, 138,* 628–654. doi:10.1037/a0027473

Sprenger, A. M., Atkins, S. M., Bolger, D. J., Harbison, J. I., Novick, J. M., Weems, S. A., Chrabaszcz, J. S., Smith, V., Bobb, S., Bunting, M. F., & Dougherty, M. R. (2013). Training working memory: Limits of transfer. *Intelligence, 41,* 638–663. doi:10.1016/j.intell.2013.07.013

Sukel, K. (2013, Dec 9). *Digital "brain health" market predicted to boom*. Retrieved from http://bigthink.com/world-in-mind/digital-brain-health-market-predicted-to-boom

Thompson, T. W., Waskom, M. L., Garel, K.L. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., Chang, P., Pollard, K., Lala, N., Alvarez, G.A., & Gabrieli, J. (2013). Failure of working memory training to enhance cognition or intelligence. *PLoS ONE, 8,* e63614. doi:10.1371/journal.pone.0063614

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6,* 291–298. doi:10.1177/1745691611406923

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*(12), 1–20.

Zinke, K., Zeintl, M., Rose, N. S., Putzmann, J., Pydde, A., & Kliegel, M. (2013). Working memory training and transfer in older adults: Effects of age, baseline performance, and training gains. *Developmental Psychology*. doi:10.1037/a0032982