Published in final edited form as:

Soc Sci Res. 2015 March ; 50: 277–291. doi:10.1016/j.ssresearch.2014.12.004.

The power of a paired t -test with a covariate

E. C. Hedberg¹ and Stephanie Ayers²

¹Arizona State University Sanford School of Social and Family Dynamics and NORC at the University of Chicago

²Arizona State University Southwest Interdisciplinary Research Center

Abstract

Many researchers employ the paired t -test to evaluate the mean difference between matched data points. Unfortunately, in many cases this test is inefficient. This paper reviews how to increase the precision of this test through using the mean centered independent variable x , which is familiar to researchers that use analysis of covariance (ANCOVA). We add to the literature by demonstrating how to employ these gains in efficiency as a factor for use in finding the statistical power of the test. The key parameters for this factor are the correlation between the two measures and the variance ratio of the dependent measure on the predictor. The paper then demonstrates how to compute the gains in efficiency a priori to amend the power computations for the traditional paired t -test. We include an example analysis from a recent intervention, *Families Preparing the New Generation (Familias Preparando la Nueva Generación)*. Finally, we conclude with an analysis of extant data to derive reasonable parameter values.

Keywords

Statistical Power; paired t -tests; regression

1. Introduction

It is still common practice in social science to collect paired observations such as pretest and posttests and perform a statistical test on the average difference to determine an average change in score. This paired t -test is a one sample (population) t -test on the mean difference the stems from Gossett's work on small sample tests of means (Student, 1908) and is a classic method to test gains from pretests to posttests (Lord, 1956; McNemar, 1958). The analysis of covariance (ANCOVA) literature also has long recognized the gains in statistical power to measure differences when using covariates (Cochran, 1957; Kisbu-Sakarya, MacKinnon, & Aiken, 2013; Oakes & Feldman, 2001; Porter & Raudenbush, 1987). Previous work has examined the use of covariates to estimate differential gains based on the

© 2014 Elsevier Inc. All rights reserved.

Corresponding Author Information. E. C. Hedberg, 951 S Cady Mall, Tempe, AZ 85287, ehedberg@asu.edu, 773 909 6801.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

initial value (Garside, 1956), but to our knowledge a factor for predicting the gains in precision for the mean difference has yet to be developed.

While such pretest-posttest designs are undesirable for causal interpretations due to uncontrolled sources of gains (Shadish, Cook, & Campbell, 2002), researchers conducting observational studies may still want to plan for the ability to detect changes in a cohort of subjects. For example, surveys of older adults may wish to find evidence of increasing or decreasing depression symptoms (see, e.g., O'Muircheartaigh, English, Pedlow, & Kwok, 2014; Payne, Hedberg, Kozloski, Dale, & McClintock, 2014). Given limited resources for any study, whether observational or experimental, it is important to maximize power for detecting such changes. The factor developed in this paper is directly relevant to planning such studies. We show how studies employing a paired *t*-test will sometimes have less power, and thus require a larger sample, than an analysis that employs regression with a covariate.

The purpose of this paper is to outline how including the pretest (centered on the pretest mean) in a prediction model of gain scores produces the same mean difference with less sampling variance, thus increasing statistical power (the chance of detecting the mean difference). We then present a factor that predicts the increase in precision based on the correlation and relative variance of the posttest and pretest variables. In the spirit of Guenther (1981), we then present power analysis techniques that employ this factor. This paper recognizes that issues of measurement error are very important with these designs (Althausen & Rubin, 1971; Lord, 1956; Overall & Woodward, 1975). We explore the implications of measurement error for the methods presented here. Moreover, our discussion section incorporates how measurement error may influence which test to use. Future work will incorporate how to include measurement error in the calculations of precision gains.

The regression-based test that we explore in this paper can be achieved using any conventional regression package. The procedure is simply to calculate two new variables. The first variable is the difference between the posttest and pretest (post minus pre). The second variable is the pretest minus the pretest mean, or the mean-centered pretest. The procedure for the analysis is simply to use the first variable, the difference between the posttest and the pretest, as the dependent variable regressed onto the mean-centered pretest. The intercept of that regression model, and its standard error, provide the regression-based test. For example, if we import our hypothetical data (see Table 1) into R and ask for a paired *t*-test (Team, 2012)

```
> y <- c(56, 69, 75, 23, 45, 70, 36, 60, 58, 77)
> x <- c(48, 56, 63, 28, 44, 52, 46, 45, 57, 65)
> t.test(y, x, paired = TRUE)
Paired t-test
data: y and x
t = 2.2158, df = 9, p-value = 0.05394
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

```
-0.1360894 13.1360894
sample estimates:
mean of the differences
6.5
```

we see that the mean difference (6.5) is not statistically significant (i.e., p -value = 0.05394). We can use the same data, but instead fit a linear model (lm) of the difference between y and x ($y-x$) regressed on the mean centered x ($I(x - \text{mean}(x))$):

```
> summary(lm(y-x ~ 1 + I(x - mean(x))))
Call:
lm(formula = y - x ~ 1 + I(x - mean(x)))
Residuals:
Min 1Q Median 3Q Max
-14.4648 -2.2181 -0.7336 3.5849 10.9977
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.5000 2.6235 2.478 0.0383 *
I(x - mean(x)) 0.4625 0.2565 1.803 0.1090
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 8.296 on 8 degrees of freedom
Multiple R-squared: 0.289, Adjusted R-squared: 0.2002
F-statistic: 3.253 on 1 and 8 DF, p-value: 0.109
```

The intercept (which is the mean difference, 6.5000) is statistically significant (0.0383) using this method. In this paper we investigate why this is the case and how to estimate the power for a study that would use this analysis method.

2. Theoretical model

In this section we outline the explicit data-generating process for which a paired t -test analysis seeks to uncover. We then explore the theoretical reason why a regression approach provides a more powerful test. We assume that data are generated from a population where each i th population member increases (or decreases) their score on a given measure by a quantity denoted as δ . We call this quantity a gain, but the following applies even if the change is, on average, negative. Algebraically, this is written as

$$y_i = x_i + \delta_i, \quad (1)$$

where y is the post-test and x is the pretest. We assume that x and δ are normally distributed

$$x \sim N(\bar{x}, \sigma_x), \quad (2)$$

$$\delta \sim N(\bar{\delta}, \sigma_{\delta}), \quad (3)$$

and as a consequence, y is also normally distributed

$$y \sim N(\bar{y}, \sigma_y), \quad (4)$$

where

$$\sigma_y = \sqrt{\sigma_x^2 + \sigma_{\delta}^2 + 2\text{cov}(x, \delta)}. \quad (5)$$

The purpose of an analysis of a sample from this population is to estimate the average of δ , $\bar{\delta}$, and its sampling variability, in order to perform a test of the hypothesis that $\bar{\delta} = 0$. The average of $y - x = \delta$ is an unbiased estimate of $\bar{\delta}$ even if x and δ are correlated. It is also true that the average of δ is obtained by finding the least squares estimate of the predicted difference for the average value of x . Finally, even in the case of measurement error in y and x , assuming it has a mean of 0 and is normally distributed, the estimate of $\bar{\delta}$ is unbiased. Thus, there are many equivalent methods for finding an estimate of the mean gain. The rest of this section explores the sampling variability of the estimate of $\bar{\delta}$ as it relates to error in the measurement.

We first turn to the true sampling variability if we knew the true gains for all population members. The sampling variability of $\bar{\delta}$ is like any other variance of a mean estimate, which is the variation of the individual gains divided by the size of the population:

$$\text{var}(\bar{\delta}) = \frac{\sigma_{\delta}^2}{N}. \quad (6)$$

The sampling variance of a paired t -test is based on the variance of the difference between y and x . Thus, substituting (5) for the variance of y ,

$$\text{var}(\bar{\delta}) = \frac{2\sigma_x^2 + \sigma_{\delta}^2 + 2\text{cov}(x, \delta) - 2\text{cov}(y, x)}{N}. \quad (7)$$

It is worth noting that this variance includes twice the variance of the pretest, the variance of the true gains and the covariance of x and δ . Positive covariance between x and δ (i.e., higher pretests lead to higher gains) increases the uncertainty of the paired test, whereas negative covariance (i.e., higher pretests lead to lower gains; the more plausible pattern) increases precision. Values relating to the realized posttest, y , only enter as part of the covariance between y and x . When there is no measurement error in either y or x , (6) and (7) produce the same quantity.

In most cases, unfortunately, variables are measured with a certain amount of error. For example, instead of observing the true value x , we observe

$$x^* = x + u \quad (8)$$

and instead of observing the true value of y , we observe

$$y^* = y + w \quad (9)$$

where u and w are the measurement errors that have a mean of 0 and are normally distributed. When there is measurement error, the variance of the paired test is inflated. This is because the variance of a variable with measurement error is the sum of the variance of the true measure and the variance of its error, thus,

$$\sigma_{x^*}^2 = \sigma_x^2 + \sigma_u^2 \quad (10)$$

and

$$\sigma_{y^*}^2 = \sigma_y^2 + \sigma_w^2. \quad (11)$$

As a consequence, the true variance from the paired test, vis-à-vis the true sampling variance of, $\bar{\delta}$ is

$$\text{var}(\overline{y^* - x^*}) = \frac{2(\sigma_x^2 + \sigma_u^2) + \sigma_\delta^2 + 2\text{cov}(x, \delta) + \sigma_w^2 + 2\text{cov}(y, x)}{N} \quad (12)$$

Which adds $2\sigma_u^2 + \sigma_w^2$ (twice the variance of the measurement error in x and the variance of the measurement error in y) to the numerator. With this added variance, the true variance of the gains (6) is smaller than the estimated variance of the gains from a paired t-test (12). Thus, researchers should seek an approach that can remove at least some of this added error variance.

While it is difficult to remove measurement errors in x without prior knowledge or an instrument variable, a regression approach can remove error in measurement from y through the residual term. As a result, the variance of the test statistic using the regression approach detailed below will provide a smaller variance that is closer to the true gain variance than a paired approach. In cases where there is no measurement error in x , our approach will estimate the true sampling variance. In the case of measurement error on both y and x , the suggested procedure will at least provide a more powerful test. In the discussion, we will showcase the consequences of this method when applied to perfectly measured y and x .

3. Review of test variances

In this section we move from the theoretical statistical model to the formulation of a practical factor that predicts gains in precision through the use of regression. This requires an examination of the paired- and regression-based approaches to estimating the variance of the gains. Below we present the variances for each test: the variance of the paired t -test and the variance of the intercept when the difference is regressed on the mean centered covariate. Details about these variances are presented in the Appendix. The factor that is ultimately detailed is the ratio of these variances.

Paired t -test variance

Suppose we have a set of observations on two variables, y and x , and we calculate a difference $d_i = y_i - x_i$. Hypothetical example data are presented in Table 1. Further suppose that we wish to test whether the mean of the difference was statistically different than 0. This test is typically accomplished with the following so-called paired t -test familiar to most introductory statistics students³

$$t = \frac{\bar{d}}{\sqrt{N^{-1} s_d^2}}, \quad (13)$$

which in the case of our hypothetical data is

$$\frac{6.500}{\sqrt{8.6056}} = \frac{6.500}{2.934} = 2.216.$$

This test has a two-tailed p -value (with 9 degrees of freedom) equal to 0.0539; not statistically significant by convention. To examine this test further, note that the denominator in the above test is the square root of the variance of the sample mean difference divided by N , $N^{-1} s_d^2$, which we will call var_d^{paired} :

$$var_d^{paired} = N^{-1} (\sigma_y^2 + \sigma_x^2 - 2\text{cov}(y, x)), \quad (14)$$

which is the familiar variance of y plus the variance of x minus twice the covariance (Thorndike, 1942). Note that expression (14) differs from expression (7) in that we substitute σ_y^2 for $\sigma_x^2 + \sigma_\delta^2 + 2\text{cov}(x, \delta)$ because we do not know from the data the true variance of the gains and how it relates to the covariate in the presence of measurement error. Next, we consider an alternative method.

The difference regressed on a mean-centered covariate

The same estimate of $\bar{d}$ can be achieved with the following bivariate regression

$$y_i - x_i = \bar{d} + \beta (x_i - \bar{x}) + \varepsilon_i, \quad (15)$$

and the test for whether $\bar{d} = 0$ is a test of the intercept,⁴

$$t = \frac{\bar{d}}{\sqrt{N^{-1} (1 - r_{yx}^2) s_y^2}}, \quad (16)$$

where r_{yx} is the correlation between y and x . With the example data, the value of $\bar{d}$ is the same, except this test produces a smaller standard error and thus a statistically significant t -test with 8 degrees of freedom (p -value is 0.0383):

³ s is the estimate of the standard deviation σ .

⁴Here, r is an estimate of the correlation ρ .

$$\frac{6.5}{\sqrt{6.8829}} = 2.4776.$$

This is because the variance of the intercept, when x is centered on its mean, is

$$var_d^{regression} = N^{-1} \left((1 - \rho_{yx}^2) \sigma_y^2 \right), \quad (17)$$

which is the variance of the *residuals* from the regression of y on x divided by N . In other words, by introducing a covariate, we are reducing the variation in the difference scores by a factor of $1 - \rho_{yx}^2$. This reduction in variance produces a smaller standard error because we are using the information in x to model the assumed correlation between x and δ .

Thus, instead of regressing the difference on the covariate, another way to estimate the mean difference is to center both the pre and post data on the pretest mean. This moves the mean of x to 0 and the mean of y to d . Since regression lines intersect the means of y and x , the intercept of the model is the mean of y or d . Substantively, the interpretation is straightforward: the intercept is the average deviation from the mean of x for an average value of x . In pretest-posttest parlance, it is the average change for the average pretest. This is presented visually in Figure 1.

4. Formulating the paired test inflation factor (*PTIF*) for power analyses

Most power analysis programs' regression routines focus on the ability to detect a slope or correlation (see, e.g., Faul, Erdfelder, Lang, & Buchner, 2007). Since the mean difference in this case is estimated by the intercept, most programs are not equipped to give a power estimate for this design. However, it is possible with modern software such as R, Stata, SAS, or SPSS, to estimate power with a noncentrality parameter. This section provides guidance on estimating this noncentrality parameter and how to use software to estimate the power for a given design.

Power for a given test is the chance of detecting an expected effect given a sample size, acceptable level of uncertainty (e.g., $\alpha = 0.05$), and variance of the effect (Aberson, 2011; Cohen, 1988). Power relates to a comparison of Type I error (likelihood of falsely rejecting the null hypothesis) and Type II error (likelihood of falsely accepting the null hypothesis). Power is 1 minus the Type II error. Figure 2 presents two distributions. The dashed line is the familiar sampling distribution under the null hypothesis. This is the central distribution. We typically break our acceptable level of Type I error (α) into each tail of the distribution.

In Figure 2 we have shaded 2.5% of the right tail to represent an alpha level of 0.05. With 10 degrees of freedom, this puts the critical value at 2.26. Alternatively, the solid line represents the sampling distribution assuming that there is an effect and that it produces a t -test of 2. This is the noncentral distribution and its mean is noted as λ , which represents the expected test statistic (which in Figure 2 is 2). Type II error is the shaded portion to the left of the critical value, labeled β , and power is the complement area of the noncentral distribution. The stronger the test, the more the noncentral distribution moves away from the

critical value increasing power. We can quantify the power for a t -test by evaluating the non-central t -distribution using the expected t -test result as the non-centrality parameter, λ , with the following formula:

$$Power = 1 - H \left[t_{\frac{\alpha}{2}, df}^{critical}, df, \lambda \right] + H \left[-t_{\frac{\alpha}{2}, df}^{critical}, df, \lambda \right], \quad (18)$$

where H is the reverse cumulative non-central distribution function evaluating the critical value of t with uncertainty level α , df degrees of freedom, and noncentrality parameter λ .

In the case of the regression approach outlined in this paper, the noncentrality parameter can be directly calculated in non-standard units

$$\lambda = \frac{\bar{d}}{\sqrt{N-1 (1 - \rho^2) \sigma_y^2}}, \quad (19)$$

with $N - 2$ degrees of freedom. Unfortunately, the variance of y is typically unknown beforehand, making a priori power computations difficult. However, researchers can often expect a certain effect size, a relative variance of y to that of x (see the discussion), and correlation between y and x .

In the case of a simple paired t -test, the noncentrality parameter is simply the effect size times the square root of the sample size, $\lambda = \Delta \sqrt{N}$. A common measure of the standardized effect size for a paired t -test is simply the expected t -test itself divided by the square root of the sample size. Which, lacking substantive information, follows the “shirt sizes” of 0.2 for small, 0.5 for medium, and 0.8 for large (Cohen, 1992). It must be stressed that such “shirt sizes” are of no practical use and must only be considered complete guesses. In the discussion we present methods to estimate these important parameters from extant data. Returning to our example data in Table 1, the paired t -test produced a result of 2.2158. This is an effect size of

$$\Delta = \frac{t}{\sqrt{N}} = \frac{2.2158}{\sqrt{10}} = 0.7007,$$

with a noncentrality parameter equal to the t -test (2.2158) and a critical value (with 9 degrees of freedom and $\alpha = 0.05$) of 2.2621. Thus, the power calculation is

$$Power = 1 - H[-2.2621, 9, 2.2158] + H[2.2621, 9, 2.2158],$$

which results in 0.51.⁵ Since the test is below the critical value, the result is not statistically significant.

Power for the regression approach can be achieved with a factor that can be applied to an effect size for use in power analysis. Because $var_d^{regression}$ will be smaller than var_d^{paired} , we can conceptualize this as a paired test inflation factor (*PTIF*)

$$PTIF = \frac{N^{-1} (\sigma_x^2 + \sigma_y^2 - 2\rho_{yx}\sigma_x\sigma_y)}{N^{-1} ((1 - \rho_{yx}^2) \sigma_y^2)} \times \frac{N - 2}{N - 1}, \quad (20)$$

which is the ratio of the variances times the ratio of the degrees of freedom. If y and x are both standard normal (the variances are equal), a reasonable approximation is

$$PTIF = \frac{2 - 2\rho_{yx}}{(1 - \rho_{yx}^2)} \times \frac{N - 2}{N - 1}. \quad (21)$$

If there is an expectation about the ratio of the variance of y to the variance of x , which we will call v , the following formula is simpler to use

$$PTIF = \frac{1 + v - 2\rho_{yx} \sqrt{v}}{(1 - \rho_{yx}^2) v} \times \frac{N - 2}{N - 1}. \quad (22)$$

Thus, the three pieces of information needed for a priori expectations of the benefits of using this method are: the correlation between y and x (ρ_{yx}), the ratio of the variance of y to x , v , and the sample size, N .

With a reasonable guess as to the ratio of the variances, v , it is possible to make an a priori power computation using the design effect and effect size. The procedure is to take the expected effect size of the paired test, Δ , multiply by the square root of the sample size, and then multiply by the square root of the *PTIF*, which requires the correlation and ratio of the variances. Thus, the noncentrality parameter for this test is

$$\lambda = \Delta \sqrt{N} \sqrt{\frac{1 + v - 2\rho_{yx} \sqrt{v}}{(1 - \rho_{yx}^2) v} \cdot \frac{N - 2}{N - 1}}, \quad (23)$$

which has $N - 2$ degrees of freedom.

⁵Calculated in R using the following syntax:

```
df <- 9 #enter the degrees of freedom
alpha <- 0.05 #set alpha
ncp <- 2.2158 #enter the noncentrality parameter
1-pt(qt(alpha/2,df, lower.tail = FALSE),df,ncp) + pt(qt(alpha/2,df),df,ncp)
```

5.1 Example *PTIF* and power calculation

Turing back to the example data in Table 1, the variance of the traditional paired difference is $2.9335^2 = 8.6054$, and the variance of the difference from the regression based approach is $2.6235^2 = 6.8828$, which leads to a ratio of about 1.25. This means that the variance from the paired test is 25% larger than the variance for the regression based approach. If we knew the variances of y and x along with the bivariate correlation, ρ_{yx} , we could predict this result with the formula presented above (the correlation of the example data is 0.8959)

$$PTIF = \frac{116.2667 + 309.8778 - 2 \times 0.8959 \times 10.7827 \times 17.6033}{(1 - 0.8959^2) 309.8778} \times \frac{10 - 2}{10 - 1} = 1.25.$$

If we recognize that the ratio of the variances of y and x is $\frac{309.8778}{116.2667} = 2.6652$, we can also calculate the *PTIF* with this ratio and replace the variance of x with 1

$$PTIF = \frac{1 + 2.6652 - 2 \times 0.8959 \times \sqrt{2.6652}}{(1 - 0.8959^2) 2.6652} \times \frac{10 - 2}{10 - 1} = 1.25.$$

Thus, if we add the square root of the *PTIF* to the formula the t -test becomes

$$\lambda = 0.7007 \sqrt{10} \sqrt{1.25} \approx 2.4775.$$

which is the t -statistic from the regression model (without rounding, the result is exact), now with $N - 2 = 8$ degrees of freedom. The power of this test is about 0.59.⁶ That is, in the example data, power increased about 14 percent.

5.2 Intuition about the *PTIF*

Figure 3 plots the *PTIF* as a function of the correlation and variance ratio. As shown, the *PTIF* decreases with the correlation but can also increase as the ratio of the variance between y and x changes. This effect quickly becomes large if the variance of y is smaller than the variance of x . Knowing the variance ratio is key for this a priori expectation, and we provide methods to make an educated guess in the discussion. Finally, the normal approximation can be quite misleading if the variances of x and y are not the same.

In most cases the regression approach will offer a more powerful design for a given sample size. These are cases in which the correlation between y and x is weak and the variances of y and x are similar. We present power curves in Figure 4 comparing the ability of each procedure to detect an effect size of 0.20 for difference values of ρ_{yx} (r) and v . The

⁶Again, in R (note the difference degrees of freedom, from 9 to 8):

```
df <- 8
alpha <- 0.05
ncp <- 2.4775
1-pt(qt(alpha/2, df, lower.tail = FALSE), df, ncp) + pt(qt(alpha/2, df), df, ncp)
```

horizontal line represents 0.80 power. The paired test requires 156 cases to achieve 0.80 power. In comparison, the regression approach often achieves power of 0.80 with fewer cases but is most beneficial in the first subplot with the lower correlation of 0.3 and slightly lower variance of y compared to x . When the number of observations is small, however, the larger critical value associated with $N - 2$ degrees of freedom may actually decrease power.

6. Example of increased power from an intervention

We now turn to an example with actual data from an intervention, *Families Preparing the New Generation (Familias Preparando la Nueva Generación)(FPNG)*. *FPNG* is an efficacy trial examining a culturally specific parenting intervention, *FPNG*, designed to increase or boost the effects of *keepin'it REAL*, an efficacious classroom-based drug abuse prevention intervention targeting middle school students (Marsiglia, Williams, Ayers, & Booth, 2013) through culturally specific materials (Williams, Ayers, Garvey, Marsiglia, & Castro, 2012). Overall, *FPNG* was designed to: (1) empower parents in aiding their youth to resist drugs and alcohol; (b) bolster family functioning and parenting skills to increase pro-social youth behavior; and (c) enhance communication and problem-solving skills between parents and youth. Parents of middle school adolescents attended 8 weekly sessions, receiving the manualized curriculum and participating in group activities.

One aspect of parenting expected to change through participation in *FPNG* was parental self-agency (Dumka et al., 1996). This measure is based on 5 select items from Dumka, Stoerzinger, et al. (1996): I feel sure of myself as a mother/father; I know I am doing a good job as a mother/father; I know things about being a mother/father that would be helpful to other parents; I can solve most problems between my child and me; and When things are going badly between by child and me, I keep trying until things begin to change. Each item had 5 responses: (1) Almost never or never, (2) Once in a while, (3) Sometimes, (4) A lot of the time (frequently), and (5) Almost always or always. These items were collected in three waves. In the fall (September – November) of the school year, a pre-intervention questionnaire (Wave 1) was administered. After completion of the intervention, at week 8, parents completed a short-term questionnaire (Wave 2). A third round of questionnaires (Wave 3) was completed in the spring (March–May) of the following school year, approximately 15 months post-intervention completion.

The example analysis presented in this paper involved the 29 members of the second cohort who answered each of the 15 items (5 scale items in each wave). We performed a principal component factor analysis on the five items from the first wave. The results are presented in Table 2, which shows each item loaded well onto the factor with an eigenvalue of 2.72. We then used the scoring coefficients (also presented in Table 2) to score each wave. Thus, the factor scores for waves 1, 2, and 3 are comparable and the means can be compared. Table 3 presents the means, standard deviations, and variances for each wave, along with the differences between waves 2 and 1, 3 and 1, and 3 and 2. The first wave, as expected, has a mean of 0 and variance of 1. The mean for Wave 2 is lower than that for Wave 1 and has a higher variance (1.11 compared to 1). Wave 3, in contrast, has a higher mean and lower variance. Finally, Table 4 presents the correlations across each wave.

With the variances and correlations presented in Tables 3 and 4, respectively, *PTIF* statistics can be calculated for each wave comparison. These *PTIF* ratios are presented in Table 5. As expected, the *PTIF* is smallest for the Wave 2 to Wave 1 comparison, since it has the largest correlation and largest variance ratio. The largest *PTIF* is for the Wave 3 to Wave 2 comparison, which is also expected since it has the lowest correlation and smallest variance ratio.

The benefits of this method are also evident in Table 5. None of the paired test results are significant, but the regression measures each have a larger absolute value for the *t*-test. This regression based *t*-test is also equal to the paired test times the squareroot of the *PTIF*. In the case of the Wave 2 to Wave 3 gains, the regression result produced a significant result whereas the result from the paired test was not significant. Practically, the implication is that without using the regression method, a false conclusion that this program had no effect would be reached.

7. Discussion

In this part of the paper we discuss several implications of this analysis strategy. First, we discuss under which conditions this approach is useful. Next, we offer guidance on the key design parameters.

7.1. Appropriate Analysis Conditions

An important question to ask, given the results in this paper, is when to use this approach. Will it always be beneficial; will it sometimes underestimate the true sampling variance? In order to explore these issues, we performed a Monte Carlo simulation to estimate variances under six conditions:

1. δ correlated with x , but with no measurement error
2. δ correlated with x , with measurement error in y but not x
3. δ correlated with x , with measurement error in both y and x
4. δ uncorrelated with x , but with no measurement error
5. δ uncorrelated with x , with measurement error in y but not x
6. δ uncorrelated with x , with measurement error in both y and x

In each simulation, we make draws from the normal distribution to produce x (with a mean of 0 and standard deviation of 1) for 30 observations. We then compute a value for δ by making a draw from the normal distribution (with a mean of 1 and standard deviation of 0) and adding to that a specified correlation multiplied by x . In the case of correlations between x and δ , we set the correlation to -0.9 ;⁷ in the case of no correlation, we set the correlation to 0. Thus, δ is created with

$$\delta_i = z_{i,1} + \rho \times x_i.$$

⁷Note that the realized correlation was often much smaller due to random variation.

We then calculate y following (1), which is x plus the individual gain. In cases where we introduce measurement error into y or x , we simply add another draw from the normal distribution to y or x , after the true y , x , and δ are set.

Once the set of 30 cases is created, we estimate the sampling variance for the paired test and the regression approach. In addition, we estimated the average of the true δ for each simulation. For each of the 6 scenarios, we performed this simulation 500 times. Using the results of the 500 simulations, we calculated the standard deviation of the true δ and the mean standard errors from the paired approach and the regression approach. Thus, we are comparing three quantities: the true sampling distribution of δ , the average estimated sampling distribution from the paired approach, and the average sampling distribution from the regression approach.

The results from the exercise are presented in Figure 5. In scenario 1, with a correlated δ and without measurement error, we see that the regression approach *underestimates* the true sampling variability of δ . This is also true, but to a lesser extent, for scenario 4 with an uncorrelated δ . In other cases in which δ is correlated with x but measurement error is introduced (scenarios 2 and 3), we see that the regression approach does a good job replicating the true sampling variability of δ whereas the paired approach overestimates the sampling variability. However, in cases where δ is not correlated with x , we see that the regression approach does not show an improvement over the paired test in reducing the estimated sampling distribution.

The results of these simulations have several implications for research. The paired approach is only appropriate when the researchers are certain that the variables are measured without error. In most cases, this heroic assumption is never met. This leads to the conclusion that the regression approach will either give the correct sampling distribution of the gain or at least overestimate it.

7.2 Reasonable expectations

Power analyses are essentially arguments. Any reasonable argument that is believed must be based on reasonable expectations. Thus, choices of the parameters employed in a power exploration for this type of analysis must be reasonable. First, the correlation between y and x , ρ_{xy} , can be approximated from test-retest reliability of the outcomes employed. If the literature lacks this information, it can be approximated with extant data (see below). More importantly, the techniques employed in this paper involve the new parameter v , the ratio of the variances of y and x , and thus some discussion of how to estimate this parameter is in order. We offer both theoretical and empirical guidance for this parameter.

7.2.1 Theoretical guidance on v —The first method to estimate the v parameter (the ratio of the variances of y and x) is theoretical. For this, we consider what the ratio of variances is in the population. Using (5) we can express v as

$$v = 1 + \frac{\sigma_{\delta}^2}{\sigma_x^2} + \frac{2\rho_{x\delta}\sigma_{\delta}}{\sigma_x}. \quad (24)$$

Table 6 provides values of v for different levels of variation in x and δ , and correlations between x and δ . The average value of v is about 3.67 and it ranges from 0.2 to 24.2. Of course, very small standard deviations of x and/or δ can produce very large results. Care must be taken when considering σ_δ , which is the variation in individual gains. We will typically expect gains to vary less than pretests, so the first two columns in table 6 are the likely values. Also difficult to estimate (guess) is the correlation of gains to pretest, $\rho_{x\delta}$. One possibility is to employ the variances and reliabilities of the observed y and x (see expression 11 in Thomson, 1924). The standard recommendation is to derive these values from a pilot sample or extant data as we do in the next section.

7.2.2 Empirical Guidance on v and ρ_{yx} —The best method to find estimates of v and ρ_{yx} is to employ previously collected data about the outcome measures measured at two points in time. This can either be from a previous project, pilot data, public data sources. Whatever the source, the procedure is simple: organize the data so that each unit has a single row and two columns: one for the posttest and one for pretest. Next, using a standard statistical package or a standard spreadsheet program, calculate the correlation between posttest and pretest as an estimate of ρ_{yx} . Then, calculate the variance of the posttest and pretest as estimates of σ_y^2 and σ_x^2 , respectively. Finally, the estimate of v is the ratio $\hat{\sigma}_y^2/\hat{\sigma}_x^2$. Below we present a small number of examples from these popular data sources. As with any power analyses, we encourage analysis of prior data sources to gain information about the expected parameters.

Table 7 presents examples of these calculations from three data sources: the 1998 cohort of the Early Childhood Longitudinal Study (ECLS, United States Department of Education. Institute of Education Sciences. National Center for Education, 2014), The National Longitudinal Study of Adolescent to Adult Health (a.k.a. Add Health, Harris & Udry, 2014), and The Evaluation of the Gang Resistance Education and Training Program in the United States (GREAT, F.-A. Esbensen, 2006). The items presented are not the exhaustive list of items or questions that are longitudinal in these data sources, but offer instead a small list of possible outcomes of interest for illustration.

Data from ECLS tracks a kindergarten cohort from 1998 on various academic and health related outcomes. Table 7 presents parameters from reading and math achievement tests as well as body mass index (BMI). Each row in the table represents a different period of growth, sometimes from fall to spring of school year and later on several years between grades. Examining the reading outcome, we see that the correlations between the posttest and pretest are high (about 0.80), and that the effect sizes () are also large. The last parameter in the table, however, v , is the column of interest.

We see that in kindergarten, the ratio of the variances is near 2, but as the cohort ages this parameter gets smaller. From 3rd grade to 5th grade, it falls below 1. We see a similar pattern with the math scores, with large ratios in the early years and smaller ratios in the later years. This is in contrast with the BMI measures. While we still find large correlations between posttest and pretest, the effect sizes do not follow a linear pattern. Moreover, whereas we found large ratios in academic achievement in the earlier grades, measure of BMI showcase smaller ratios in the earlier years.

The next panel in Table 7 presents results from Add Health. The Add Health study tracks high school students and their health and social outcomes. The outcomes presented here include incidents of smoking (tobacco and then marijuana) in the past 30 days from the survey date. The first follow up survey was a year after the initial interview. For these outcomes, the posttest to pretest correlations are small and do not seem to follow a pattern. The effect sizes generally indicate that smoking tobacco increases during high school then starts to decrease afterwards. Similar to academic achievement in ECLS, the ratios of the variances are larger in the earlier grades.

Finally, Table 7 presents results from the GREAT intervention. These data are different from the survey data sources in that they contain both a control and treatment group. While all effects represent changes from 7th to 8th grade, each row in this section represents either the treatment or control group. The items from these data are attitudinal scales (1= strongly disagree to 5 = strongly agree). As expected, the correlations between posttests and pretests with such scales are lower than with the refined standardized scales from ECLS (about 0.35). With the first two attitude items presented (“It’s exciting to get into trouble” and “It’s OK to steal”) the results are very similar in terms of the correlations. However, the ratio for both items is larger for the treatment group. This is in contrast with the gang related item, in which case the intervention group had a lower ratio. It is worth noting that the effect size for the gang item is also lower for the intervention group compared to the treatment group. This is consistent with the results of this study, which indicated that the intervention had little immediate effects (F. A. Esbensen, Osgood, Taylor, Peterson, & Freng, 2001).

8. Conclusion

Power analysis is an important aspect of any study design. In this paper we have presented a method for testing paired differences that is more powerful than the standard *t*-test. More importantly, we have developed a method to gauge the power of these designs before data collection. It is common knowledge that a more sensitive test of the difference between two variables on paired observations can be achieved using a mean-centered covariate. This paper conceptualized the gains in power as a factor, the *PTIF*, which can be used in power computations. We also explored how measurement error impacts these phenomena.

Power analyses indicate the regression approach is most beneficial with smaller correlations between the two variables and similar variances. The limitation of this method is that the power analysis requires more information, namely the correlation between *y* and *x*, and the ratio of the variances of *y* to *x*. An example from an actual study was presented using data on gains in parent self-agency due to an intervention in which the paired test approach did not yield significant results whereas the regression approach did. Finally, we also provide example analyses of extant data to gain insight into parameter values.

Acknowledgements

This research was supported by funding from the National Institutes of Health/National Institute on Minority Health and Health Disparities (NIMHD/NIH), award P20 MD002316 (F. Marsiglia, P.I.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIMHD or the NIH.

Appendix

Formulations of estimated variances

The denominator in the paired t-test test is the square root of the variance of the sample

mean difference, $\frac{1}{N}\sigma_d^2$, which we will call var_d^{paired} :

$$var_d^{paired} = \frac{1}{N} \times \frac{\sum_i ((y_i - x_i) - (\bar{y} - \bar{x}))^2}{N - 1}.$$

Pulling quantities out of the summations, we can write this variance in an extended format that will prove useful for comparison

$$var_d^{paired} = \frac{1}{N} \times \frac{\sum_i y_i^2 + \sum_i x_i^2 - 2\sum_i x_i y_i + 2\bar{x}\sum_i y_i - 2\bar{y}\sum_i x_i + N\bar{x}^2 + N\bar{y}^2 - 2N\bar{x}\bar{y}}{N - 1}.$$

The same result can be obtained using transformed y and x values. This transformation would subtract the mean of x, $\bar{x}$, from both y and x and running the same regression model

$$d_i = (y_i - \bar{x}) - (x_i - \bar{x}) = \bar{d} + \varepsilon_i$$

We can then rewrite the extended version of the variance of d for the population if we

consider that $\bar{x} = 0$, $\sum_i x_i = 0$, and $\sum_i x_i^2$ is the sum of squares for x. Given those identities we can rewrite the variance of d as,

$$var_d^{paired} = \frac{1}{N} \left(\frac{\sum_i y_i^2}{N} + \frac{\sum_i x_i^2}{N} - 2 \frac{\sum_i x_i y_i}{N} - \bar{y}^2 \right).$$

We can make this formula even simpler if we recognize that $\frac{\sum_i y_i^2}{N} - \bar{y}^2$ is the variance of y,

or σ_y^2 , and again with $\bar{x} = 0$ that $\frac{\sum_i x_i^2}{N}$ is the variance of x, or σ_x^2 , and finally that $\frac{1}{\sigma_x^2} \frac{\sum_i x_i y_i}{N}$ is

the regression slope β of x predicting y, so $\frac{\sum_i x_i y_i}{N} = \beta \sigma_x^2 = \left(\rho \frac{\sigma_y}{\sigma_x} \right) \sigma_x^2 = \rho \sigma_x \sigma_y$. The simplified formula is then

$$var_d^{paired} = N^{-1} \left(\sigma_y^2 + \sigma_x^2 - 2\rho\sigma_x\sigma_y \right),$$

which is the familiar variance of y plus variance of x minus twice the covariance as found in expression 1 in Thorndike (1942).

The variance of the intercept for the regression $d_i = d + \beta(x_i - \bar{x}) + \varepsilon_i$, which we call $var_d^{regression}$, is

$$var_d^{regression} = \frac{\sum_i \varepsilon_i^2}{N-2} \left(\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right) = \frac{\sum_i \varepsilon_i^2}{N-2} \left(\frac{1}{N} + \frac{(0-0)^2}{\sum_i (x_i - \bar{x})^2} \right) = \frac{1}{N} \times \frac{\sum_i \varepsilon_i^2}{N-2},$$

which is simply the variance of any predicted value without the term involving the deviation

from the mean of x since the mean of x is 0. However, $\sum_i \varepsilon_i^2$ in this case will be smaller than in the case of regression on the difference without a covariate, because $\varepsilon_i = (y_i - x_i) - ((\bar{y} - \bar{x}) + \beta(x_i - \bar{x}))$ instead of $(y_i - x_i) - (\bar{y} - \bar{x})$, making the variance

$$var_d^{regression} = \frac{1}{N} \times \frac{\sum_i ((y_i - \bar{y}) - ((\bar{y} - \bar{x}) + \beta(x_i - \bar{x})))^2}{N-2},$$

or in extended form

$$var_d^{regression} = \frac{1}{N} \times \frac{\sum_i y_i^2 - 2\bar{y}\sum_i y_i + 2\bar{y}\sum_i x_i - 2\beta\sum_i x_i y_i + 2\beta\bar{x}\sum_i y_i + N\bar{y}^2 - N2\bar{y}\beta\bar{x} + \beta^2\sum_i x_i^2 - 2\beta^2\bar{x}\sum_i x_i + N\beta^2\bar{x}^2}{N-2}.$$

As before, we can then rewrite the extended version of the variance of d

$$var_d^{regression} = \frac{1}{N} \left(\frac{\sum_i y_i^2}{N} - \bar{y}^2 - 2\beta \frac{\sum_i x_i y_i}{N} + \beta^2 \frac{\sum_i x_i^2}{N} \right).$$

Also, we can make this formula simpler if we recognize the same identities as before:

$$var_d^{regression} = N^{-1}(\sigma_y^2 - 2\sigma_x^2\beta^2 + \beta^2\sigma_x^2),$$

which is equivalent to

$$var_d^{regression} = N^{-1} \left(\sigma_y^2 - \left(\rho \frac{\sigma_y}{\sigma_x} \right)^2 \sigma_x^2 \right).$$

Since $\sigma_y^2 - \left(\rho \frac{\sigma_y}{\sigma_x} \right)^2 \sigma_x^2$ is the variance of the residuals, which is $(1 - \rho^2) \sigma_y^2$, the variance is as we would expect it: the variance of the residuals divided by N , or

$$var_d^{regression} = N^{-1} \left((1 - \rho^2) \sigma_y^2 \right).$$

References

- Aberson, CL. Applied power analysis for the behavioral sciences. Routledge; 2011.
- Althausen RP, Rubin D. Measurement error and regression to the mean in matched samples. *Social Forces*. 1971; 50(2):206–214.
- Cochran WG. Analysis of covariance: its nature and uses. *Biometrics*. 1957; 13(3):261–281.
- Cohen, J. Statistical power analysis for the behavioral sciences. Psychology Press; 1988.
- Cohen J. A power primer. *Psychological Bulletin*. 1992; 112(1):155. [PubMed: 19565683]
- Dumka LE, Stoerzinger HD, Jackson KM, Roosa MW. Examination of the cross-cultural and cross-language equivalence of the parenting self-agency measure. *Family Relations*. 1996:216–222.
- Esbensen F-A. Evaluation of the Gang Resistance Education and Training (GREAT) Program in the United States, 1995–1999. 2006 Retrieved from: <http://doi.org/10.3886/ICPSR03337.v2>.
- Esbensen FA, Osgood DW, Taylor TJ, Peterson D, Freng A. HOW GREAT IS GREAT? RESULTS FROM A LONGITUDINAL QUASI-EXPERIMENTAL DESIGN*. *Criminology & Public Policy*. 2001; 1(1):87–118.
- Faul F, Erdfelder E, Lang A-G, Buchner A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*. 2007; 39(2):175–191. [PubMed: 17695343]
- Garside R. The regression of gains upon initial scores. *Psychometrika*. 1956; 21(1):67–77.
- Guenther WC. Sample size formulas for normal theory t tests. *The American Statistician*. 1981; 35(4):243–244.
- Harris KM, Udry JR. National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994–2008 [Public Use]. 2014 Retrieved from: <http://doi.org/10.3886/ICPSR21600.v15>.
- Kisbu-Sakarya Y, MacKinnon DP, Aiken LS. A Monte Carlo comparison study of the power of the analysis of covariance, simple difference, and residual change scores in testing two-wave data. *Educational and Psychological Measurement*. 2013; 73(1):47–62.
- Lord FM. The measurement of growth. *Educational and Psychological Measurement*. 1956; 16(4):421–437.
- Marsiglia FF, Williams LR, Ayers SL, Booth JM. Familias: Preparando la Nueva Generación: A Randomized Control Trial Testing the Effects on Positive Parenting Practices. *Research on Social Work Practice*. 2013
- McNemar Q. On growth measurement. *Educational and Psychological Measurement*. 1958
- O’Muircheartaigh C, English N, Pedlow S, Kwok P. Sample design, sample augmentation, and estimation for Wave II of the National Social Life, Health and Aging Project (NSHAP). *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*. 2014
- Oakes JM, Feldman HA. Statistical Power for Nonequivalent Pretest-Posttest Designs The Impact of Change-Score versus ANCOVA Models. *Evaluation Review*. 2001; 25(1):3–28. [PubMed: 11205523]
- Overall JE, Woodward JA. Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin*. 1975; 82(1):85.
- Payne C, Hedberg E, Kozloski M, Dale W, McClintock MK. Using and Interpreting Mental Health Measures in the National Social Life, Health, and Aging Project. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*. 2014; 69(Suppl 2):S99–S116.
- Porter AC, Raudenbush SW. Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology*. 1987; 34(4):383.
- Shadish, WR.; Cook, TD.; Campbell, DT. Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Cengage learning; 2002.
- Student B. The probable error of a mean. *Biometrika*. 1908; 6(1):1–25.
- Team RC. R: A language and environment for statistical computing. 2012
- Thomson GH. A formula to correct for the effect of errors of measurement on the correlation of initial values with gains. *Journal of Experimental Psychology*. 1924; 7(4):321.
- Thorndike RL. Regression fallacies in the matched groups experiment. *Psychometrika*. 1942; 7(2):85–102.

- United States Department of Education. Institute of Education Sciences. National Center for Education, S. Early Childhood Longitudinal Study [United States]: Kindergarten Class of 1998–1999, Kindergarten-Eighth Grade Full Sample. 2014 Retrieved from: <http://doi.org/10.3886/ICPSR28023.v1>.
- Williams LR, Ayers SL, Garvey MM, Marsiglia FF, Castro FG. Efficacy of a culturally based parenting intervention: Strengthening open communication between Mexican-heritage parents and adolescent children. *Journal of the Society for Social Work and Research*. 2012; 3(4):296. [PubMed: 23805361]

Highlights

- The paired-t test is a basic but popular statistical test
- Use of regression to improve power in t-tests is also common
- We provide guidance on computing power for such a design

Figure 1.
Plot of example data from Table 1 with regression line (solid) and dashed reference lines

Figure 2.
Example of Type I and Type II errors as they relate to power

Figure 3.
Values of the paired test inflation factor as a function of the correlation and variance ratio

Graphs by r and v

Figure 4.

Power computations for an effect size of 0.2 using the traditional paired approach and the regression approach for difference values of the correlation and variance ratio.

Figure 5.
Results of Simulations Depicting Sampling Variances Under Different Conditions

Table 1

Hypothetical example data

	<i>y</i>	<i>x</i>	<i>d</i>
	56	48	8
	69	56	13
	75	63	12
	23	28	−5
	45	44	1
	70	52	18
	36	46	−10
	60	45	15
	58	57	1
	77	65	12
Mean	56.9	50.4	6.5
Standard deviation	17.6033	10.7827	9.2766
Variance	309.8778	116.2667	86.0556

Table 2Factor analysis on Wave 1 *FPNG* Parental Agency Scale

	Factor loadings	Scoring coefficients
I feel sure of myself as a mother/father.	0.780	0.287
I know I am doing a good job as a mother/father.	0.870	0.320
I know things about being a mother/father that would be helpful to other parents.	0.666	0.245
I can solve most problems between my child and me.	0.734	0.270
When things are going badly between my child and me, I keep trying until things begin to change.	0.609	0.224
Notes: N = 29, Eigenvalue = 2.72		

Table 3

Summary statistics on *FPNG* Parental Agency Scale for Waves 1–3

	Wave 1	Wave 2	Wave 3	d_{W2-W1}	d_{W3-W1}	d_{W3-W2}
Mean	0.0000	−0.1142	0.2088	−0.1142	0.2088	0.3230
Standard deviation	1.0000	1.0550	0.8414	1.0060	1.0820	1.1447
Variance	1.0000	1.1131	0.7079	1.0119	1.1708	1.3103

Notes: N = 29

Table 4Correlations among *FPNG* Parental Agency Scales for Waves 1–3

	Wave 1	Wave 2	Wave 3
Wave 1	1.0000		
Wave 2	0.5219	1.0000	
Wave 3	0.3192	0.2876	1.0000
Notes: N = 29			

Table 5Comparison of results from *FPNG* Parental Agency Scale Analysis

	<i>PTIF</i>	Paired <i>t</i> -test (N-1 degrees of freedom)		Regression <i>t</i> -test (N-2 degrees of freedom)	
		<i>t</i>	<i>p</i> -value	<i>t</i>	<i>p</i> -value
Wave 2 – Wave 1	1.20	-0.61	0.55	-0.67	0.51
Wave 3 – Wave 1	1.78	1.04	0.31	1.38	0.18
Wave 3 – Wave 2	1.95	1.52	0.14	2.12	0.04
Notes: N = 29					

Table 6

values of ν for differing levels of variation in gains and pretest, and correlations between gains and pretest

		σ_g					σ_g				
$\rho_{x,g}$	σ_x	0.250	0.500	0.750	1.000	$\rho_{x,g}$	σ_x	0.250	0.500	0.750	1.000
-0.900						0.250					
	0.250	0.200	1.400	4.600	9.800		0.250	2.500	6.000	11.500	19.000
	0.500	0.350	0.200	0.550	1.400		0.500	1.500	2.500	4.000	6.000
	0.750	0.511	0.244	0.200	0.378		0.750	1.278	1.778	2.500	3.444
-0.500	1.000	0.613	0.350	0.213	0.200		1.000	1.188	1.500	1.938	2.500
						0.500					
	0.250	1.000	3.000	7.000	13.000		0.250	3.000	7.000	13.000	21.000
	0.500	0.750	1.000	1.750	3.000		0.500	1.750	3.000	4.750	7.000
-0.250	0.750	0.778	0.778	1.000	1.444		0.750	1.444	2.111	3.000	4.111
	1.000	0.812	0.750	0.812	1.000		1.000	1.312	1.750	2.312	3.000
						0.900					
	0.250	1.500	4.000	8.500	15.000		0.250	3.800	8.600	15.400	24.200
0.000	0.500	1.000	1.500	2.500	4.000		0.500	2.150	3.800	5.950	8.600
	0.750	0.944	1.111	1.500	2.111		0.750	1.711	2.644	3.800	5.178
	1.000	0.938	1.000	1.188	1.500		1.000	1.513	2.150	2.912	3.800
	0.250	2.000	5.000	10.000	17.000						
	0.500	1.250	2.000	3.250	5.000						
	0.750	1.111	1.444	2.000	2.778						
	1.000	1.062	1.250	1.562	2.000						

Table 7Empirical values of ρ_{yx} , $\rho_{\delta x}$, and ν , from select items using public datasets

Source, Measure, and Time	ρ_{yx}	$\rho_{\delta x}$	ν	
Early Childhood Longitudinal Study 1998 Cohort				
Reading test standardized score				
Fall Kindergarten to Spring Kindergarten	0.83	0.19	1.42	1.91
Fall 1st to Spring 1st grade	0.83	0.17	1.80	1.84
Spring 1st to Spring 3rd grade	0.73	-0.19	2.56	1.36
Spring 3rd to Spring 5th grade	0.85	-0.38	1.56	0.88
Spring 5th to Spring 8th	0.78	-0.24	1.07	1.14
Math test standardized score				
Fall Kindergarten to Spring Kindergarten	0.83	0.14	1.53	1.76
Fall 1st to Spring 1st grade	0.82	0.07	1.74	1.64
Spring 1st to Spring 3rd grade	0.78	0.08	2.42	1.86
Spring 3rd to Spring 5th grade	0.87	-0.24	1.94	1.02
Spring 5th to Spring 8th	0.85	-0.45	1.30	0.81
Child Body Mass Index				
Fall Kindergarten to Spring Kindergarten	0.86	-0.17	0.10	1.11
Fall 1st to Spring 1st grade	0.90	0.02	0.20	1.23
Spring 1st to Spring 3rd grade	0.87	0.25	0.88	1.81
Spring 3rd to Spring 5th grade	0.92	0.23	0.97	1.48
Spring 5th to Spring 8th	0.73	-0.02	0.55	1.80
ADD Health				
Times Smoked Tobacco Last 30 days				
7th to 8th Grade	0.02	-0.56	0.54	2.01
8th to 9th grade	-0.12	-0.72	0.15	1.19
9th to 10th grade	-0.08	-0.72	0.15	1.10
10th to 11th grade	-0.11	-0.72	0.17	1.22
11th to 12th grade	0.04	-0.71	-0.05	0.93
12th grade to 1 year after	0.17	-0.68	-0.11	0.86
Times Smoked Marijuana Last 30 days				
7th to 8th Grade	-0.01	-0.54	-0.35	2.52
8th to 9th grade	-0.07	-0.71	-0.06	1.15
9th to 10th grade	0.06	-0.69	0.01	0.98
10th to 11th grade	-0.04	-0.75	0.09	0.86
11th to 12th grade	0.00	-0.75	0.15	0.80
12th grade to 1 year after	0.05	-0.75	0.24	0.72
Gang Resistance Education and Training (GREAT) program (7th grade to 8th grade)				
It's exciting to get into trouble				
Control students	0.35	-0.48	0.24	1.33
Treatment students	0.34	-0.47	0.25	1.43

Source, Measure, and Time	ρ_{yx}	$\rho_{\delta x}$	ν
It's OK to steal if it's the only way to get something			
Control students	0.36	-0.47	0.24 1.36
Treatment students	0.35	-0.46	0.27 1.43
Gangs interfere with goals			
Control students	0.36	-0.52	0.17 1.17
Treatment students	0.28	-0.58	0.10 1.09