
Explaining and Controlling Regression to the Mean in Longitudinal Research Designs

TUTORIAL

Xuyang Zhang
J. Bruce Tomblin
University of Iowa,
Iowa City

This tutorial is concerned with examining how regression to the mean influences research findings in longitudinal studies of clinical populations. In such studies participants are often obtained because of performance that deviates systematically from the population mean and are then subsequently studied with respect to change in the trait used for this selection. It is shown that in such research there is a potential for the estimates of change to be erroneous due to the effect of regression to the mean. The source of the regression effect is shown to arise from measurement error and a sampling bias of this measurement error in the process of selecting on extreme scores. It is also shown that regression effects are greater with measures that are less reliable and with samples that are selected with more extreme scores. Furthermore, it is shown that regression effects are particularly prominent when measures of change are based on changes in dichotomous states formed from quantitative, normally distributed traits. In addition to a formal analysis of the regression to the mean, the features of regression to the mean are demonstrated via a simulation.

KEY WORDS: regression to the mean, longitudinal design, language diagnosis, recovery, improvement of performance

Research in the field of communication sciences and disorders often entails the use of repeated measures performed on a sample of individuals. In some cases these individuals are participants in an intervention, and thus the measures are aimed at documenting response to treatment. In other cases, the individuals are members of a cohort being followed to document features of natural history or outcomes. In each of these cases, the focus of the research is usually one of change over time as reflected by measures performed at particular time intervals. Research designs that use multiple measures to document change are vulnerable to an important threat to their validity. This threat is called regression to the mean. Although this term is well known, its properties, the conditions that cause it, and its impact on data have not been well documented. This tutorial is intended to provide an explicit formalization of what regression to the mean is, why and when it will occur, its effects on research outcomes, and, finally, ways it can be managed in research.

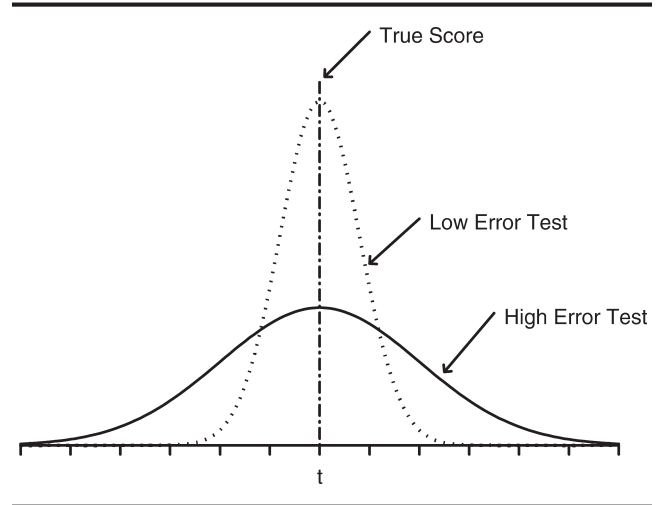
Cook and Campbell (1979), in their book on quasi-experimental design, noted that regression to the mean depends on both the reliability of a measure and on the degree to which a selected subgroup differs from the population mean. Thus, to understand regression to the mean

one must first understand the basis of test reliability or its complement, measurement error. Additionally, one must be aware that sampling a group of individuals based on values of a measure that are systematically above or below the mean creates the conditions for regression towards the mean to occur. In this tutorial, the formal mathematical mechanism for regression to the mean is provided under the assumption of normal distribution. This mathematical treatment is provided to promote understanding of the phenomenon and is not intended to imply that regression toward the mean will only occur when normality of distribution holds. Regression to the mean may occur for test scores for which distribution is not normal, but the theoretical magnitude of its effect is more difficult to calculate than under the assumption of normality. Secondly, we illustrate by simulation how regression toward the mean can result in erroneous conclusions about change in performance, that is, interpreting the change as true gains in performance. Further, we show that regression to the mean can generate the illusion of differential effects in clinical subgroups. Finally, we show how regression effects can be controlled.

Measurement Error

As we noted, measurement error provides an important condition for the occurrence of regression to the mean. Therefore, we need to explore the concept and mathematics of measurement error. In behavioral sciences, researchers and clinicians often need to measure a characteristic of clients or research participants. In such cases we assume that, at least theoretically, this characteristic has a true value. For instance, there exists at the moment of measurement a true threshold for hearing sensitivity, or a true vocabulary level. Accurately measuring these characteristics is always desirable, but all test scores contain measurement error, that is, the obtained test score is not likely to be equal to the person's true score. A person's true score at the testing time can have only one value and therefore is fixed. However, the person's obtained test score is a sample from a distribution of a random variable. This distribution specifies possible values for the person's obtained test score and a probability associated with each possible value. Thus, even if we knew a person's true score, which in reality we can't, we still wouldn't be able to specify exactly what score the person would get on a given observation. By definition, a specific realization of a random variable will vary, but the probability of getting a specific value can be specified when the distribution of the variable is known. The true score and measurement error variance are the two parameters for the distribution of obtained scores assuming normality. Figure 1 presents obtained score distributions for an individual with true score t on

Figure 1. Distribution of obtained scores given a true score. Shown are two probability density functions of obtained scores for a true score with a specific value t (the dashed vertical line). The dotted curve represents a measure of t with a smaller magnitude of measurement error—a more reliable test—than the solid curve.



two tests. These two tests have different amounts of measurement error variance. The narrow distribution corresponds to a low-error test, and the wide one corresponds to a high-error test.

In the classical measurement theory, the distribution of the test scores for a given person is assumed to be normal with a mean equal to the person's true score and variance equal to the measurement error variance, as shown in Figure 1. We acknowledge that the score distribution may not be normal for those extremely low or extremely high true scores, commonly called floor or ceiling effect, but the skewness caused by this exception should not be prominent if the test is well developed and used age-appropriately. A specific test score for a person is a random sample from his or her theoretical obtained test score distribution. This theoretical distribution may be thought of as the set of scores obtained from some large sample of measures in a circumstance where the person's true score remains constant. Thus, it can be seen that for an individual with a particular true score (T), the obtained score X is solely determined by the particular value of the measurement error on the occasion of testing. This can be stated formally as

$$X | (T = t) = t + e.$$

That is, the obtained score (X) equals the true score (t) plus error (e).

When the measurement error variance of the test is large, the distribution will be spread out, and the single observation will have more chance to be far away from the distribution mean, that is, the person's true score. On the other hand, when the measurement error variance is small, the distribution of a person's obtained test

score will be narrow, and the person's obtained score will be more likely to be close to the distribution mean (i.e., the person's true score). In the extreme case in which the measurement error variance is 0, the distribution of a person's obtained test score converges to a single value, which is the person's true score. Sampling from this distribution, the probability of getting the true score is unity and getting any other score is null. In the extreme case in which the measurement error variance is infinite, the distribution of a person's obtained test score is flat, and a sample from this distribution will have nothing to do with the person's true score.

Test Reliability

A good test should have a small measurement error variance relative to the true score variance. In fact, the ratio of true score variance to the obtained score variance has been defined as test reliability. Reliability has typically been obtained by computing the correlation coefficient between repeated measures of the same or parallel tests, referred to here as X and Y . The correlation between X and Y is mathematically equivalent to the proportion of true score variance in the obtained score variance where $X = T + e_X$ and $Y = T + e_Y$. The common assumptions for the variables in these two equations are that e_X and e_Y are not correlated to T , nor are they correlated with each other. The variance of X is assumed to be equal to the variance of Y because they are equivalent forms of the same test. Then, we have

$$\begin{aligned} \rho_{xy} &= \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \frac{\text{Cov}(T + e_X, T + e_Y)}{\text{Var}(X)} \\ &= \frac{\text{Cov}(T, T) + \text{Cov}(T, e_Y) + \text{Cov}(e_X, T) + \text{Cov}(e_X, e_Y)}{\text{Var}(X)} \\ &= \frac{\text{Var}(T) + 0 + 0 + 0}{\text{Var}(X)} \\ &= \frac{\text{Var}(T)}{\text{Var}(X)} = \text{reliability}, \end{aligned}$$

where $\text{Cov}(X, Y)$ is the covariance between X and Y , $\text{Var}(T)$ is the true score variance, $\text{Var}(X)$ is the obtained score variance, and ρ_{XY} is the correlation between X and Y . Thus, although we cannot estimate the variance of T directly, the correlation between X and Y (ρ_{XY}) is an estimate of the ratio of true score variance to the obtained score variance (reliability).

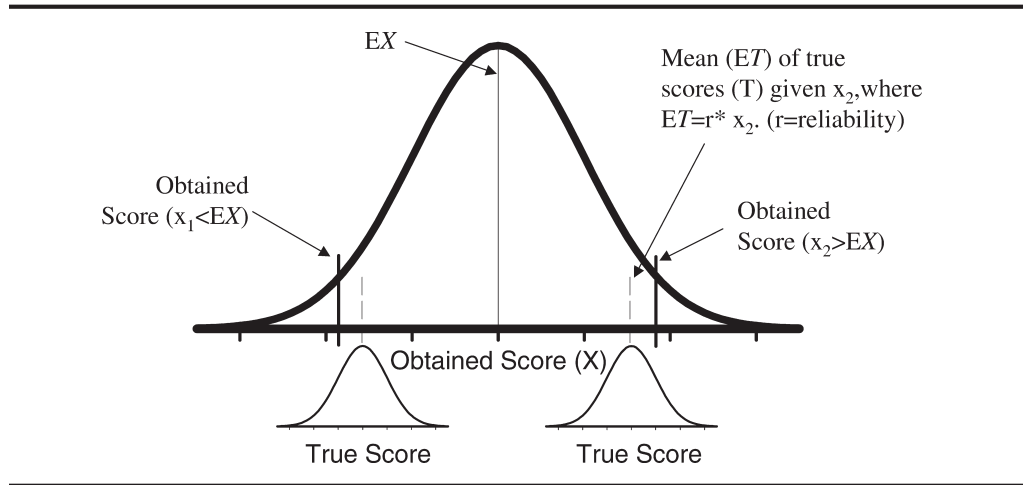
Regression Toward the Mean

As we noted earlier, measurement error is a key factor in understanding regression to the mean. Consider the situation where we have a random sample of

examinees, all of whom have the same true score depicted in Figure 1. Although we see that the obtained scores are normally distributed around this true score, the average measurement error in the sample will approach 0 when the sample size is large because overestimation and underestimation will be cancelled out in the sum of different observations.

The distribution of X given $T = t$ is a normal distribution with the mean equal to t and variance equal to the measurement error variance assuming that the test does not have a serious floor or ceiling effect. However, the reverse is not true, and it is this situation that leads to regression to the mean. For a sample of individuals with a given obtained score that is either greater or lesser than the population mean, the mean of the true scores is not equal to the obtained score. In fact, as shown in Appendix A, the conditional distribution of T given $X = x$ is a normal distribution with mean (i.e., expectation of T [ET]) equal to the product of the reliability times x . Thus, the true score mean for those individuals with the same obtained score x is closer to the population mean than x . Appendix A provides the mathematical proof for this. Specifically, the expectation of T given $X = x$ is the product of x times the test reliability. In Figure 2, we display distributions of true scores given that the obtained score is equal to x_1 and x_2 . If two groups of individuals were selected to have particular values (for one group $X = x_1$ and for the other $X = x_2$), and then the true score for each individual in these two groups was revealed, the true score mean would be shifted relative to the values of X . In each case the shift of true scores would be toward the mean of the distribution of X . As noted above, this shift is a function of the reliability of the variable X and the particular value of x_1 and x_2 , demonstrating that selecting obtained scores that deviate from the mean comprises a biased sample of measurement error such that the underestimation and overestimation of the true score cannot cancel each other. This bias is systematic in that more individuals' obtained scores have error away from the population mean than toward the population mean. Recall that the true score for any individual represents the hypothetical mean of that person's universe of obtained scores (i.e., the mean of all possible obtained scores). Because the obtained score distribution is normal with a mean equal to the person's true score, sampling from this distribution is more likely to yield a score close to the mean (i.e., the true score) than a score away from the mean. Thus, if this person is re-assessed (i.e., sampled from the distribution again) our best guess would be a score (y) equal to the person's true score, which is closer to the population mean than x . Thus, the person's next obtained score would more likely shift toward the population mean than away from the population mean. This is in fact the phenomenon of regression towards the mean (Blommers & Forsyth, 1977, pp. 492–496).

Figure 2. Shift of true score means relative to given values of obtained scores. Also depicted are the distribution of obtained score X and two distributions of true scores, with means denoted by dashed vertical lines, for individuals sampled because of a specific obtained score x_1 and x_2 (bold vertical lines).



We have just shown that an individual's obtained score is likely to regress toward the mean on a second testing. It should not be surprising then to find that regression toward the mean also shows up when a range, rather than a specific value, of obtained scores is concerned (i.e., classifying individuals into subgroups according to their obtained scores). Assume that different subgroups were defined on X (e.g., $X < a$ or $X \geq a$) and only the subgroup with $X < a$ was subsequently given another alternate form of the test (call the score from this form Y). The question is, then, what should we expect on Y for this subgroup of individuals, assuming that X and Y are from two equivalent test forms and no true changes in the measured characteristic have ever occurred between the two test times? One may be tempted to say that this subgroup will get the same mean score of Y as the mean score of X , that is, $E(Y|X < a) = E(X|X < a)$. That is, the expectation of Y given $X < a$ is equal to the expectation of X given $X < a$ because the two tests are equivalent and X and Y are from the same individuals. However, this is incorrect for two joint reasons: the test is not perfectly reliable ($0 < \rho_{xy} < 1$) and this subgroup had been selected to be low on X . Whenever these two conditions are met, $E(Y|X < a) > E(X|X < a)$ rather than $E(Y|X < a) = E(X|X < a)$, meaning that the mean of Y theoretically will be greater than, not equal to, the mean of X .

The effect of regression toward the mean for a selected range of X is mathematically inherent, which is shown in Appendix B. The amount of regression toward the mean is $-(1 - \rho_{xy})E(X|X < a)$. If ρ_{xy} is very close to 1 (i.e., the test is almost perfectly reliable), $(1 - \rho_{xy})$ is close to 0 and the amount of regression, $-(1 - \rho_{xy})E(X|X < a)$, will be close to zero so that $E(Y|X < a)$ will be close to the mean of X given $X < a$, $E(X|X < a)$. If, on the other

hand, ρ_{xy} is very close to 0 (i.e., the test is unreliable), there will be a great amount of regression, $-E(X|X < a)$, so that $E(Y|X < a)$ will be close to the population mean, 0. Thus, individuals sampled because they obtained scores in a certain region above or below the mean are likely to shift toward the mean just as was shown for individuals sampled at one particular point on the distribution.

Effects of Regression on Research Results

Failure to appreciate the influence of regression to the mean can result in researchers misinterpreting their findings. In such studies, the research interest is typically concerned with the extent to which a group who was selected at Time 1 (T1) because of low or high scores on some measure (X) of a trait we refer to as M have changed with regard to the trait M at Time 2 (T2). The outcomes in these studies may be examined in two ways. One way is to compute the mean score of the group on a measure of trait M using a second measure Y at T2 and contrast it with the earlier mean score on X . Thus, change is represented as the difference between Y and X . Alternatively, the outcome at T2 may be the clinical status of each member of the group (affected, unaffected) with respect to the clinical trait M due to their having a score on Y that is above or below some threshold value (a). In this latter case the interest is whether a group selected because of scores exceeding the threshold for M contains some proportion at T2 who are no longer affected, that is, determining the rate of individuals crossing the threshold between T1 and T2. In such research, an intervention may or may not be provided during the interval between T1

and T2. If no intervention is provided, the change may be viewed as growth or recovery. If intervention is provided, the change may be viewed as response to treatment. In all cases, the measure of change is viewed as evidence for a true change in the trait being studied.

It should be clear from the description of regression to the mean in the prior section that all these research designs are vulnerable to an erroneous estimate of the magnitude of true change. In addition, however, the regression effect will produce artifacts that suggest differential amounts of change in subgroups that may also be erroneous. First, let us consider the case in which change is measured as a change in mean scores. Using the formula derived in Appendix B,

$$E(Y|X < a) = \rho_{xy} E(X|X < a) > E(X|X < a),$$

one can compute the expected mean score on the outcome measure Y , $E(Y|X < a)$, if the reliability (ρ_{xy}) of the test and the initial mean score of the group, $E(X|X < a)$, are known. Let us assume a very reliable measure, such as one with a reliability of .90, and a mean of X for the selected group at -2 , then the mean of Y , the follow-up measure for this group, will be -1.8 [the regression amount is $-(1 - .90) \times (-2) = 0.20$]. The 0.20 change from X to Y is due entirely to the regression effect and yet it would be tempting to conclude that this represents a true change in our trait M . This amount may not seem particularly large, but few of our measures are as reliable as this one, and with a reliability of .85, the change would be 0.30. Because these values are in z -score units, this 0.30 represents nearly a third of a standard deviation and, therefore, constitutes an effect size that would often be found to be statistically significant. If we look at the formula used to compute the magnitude of the regression, we also see that one of the terms refers to the average level of the group on our trait M in z -score units. In the example above, a value of -2 was selected; thus, the group was an average of 2 SD s below the mean. If the group was selected such that they averaged 3 SD s below the mean and the reliability of the measure was .85, the magnitude of the change due to regression to the mean would be 0.45, or nearly a half a standard deviation in apparent change when no change in the true trait actually occurred. This example also shows that if we were interested in comparing two clinical groups, one selected with a lenient diagnostic standard and one with a more stringent diagnostic standard, we would find that the magnitude of apparent change for the more severely affected group would be greater than for the less severely affected group.

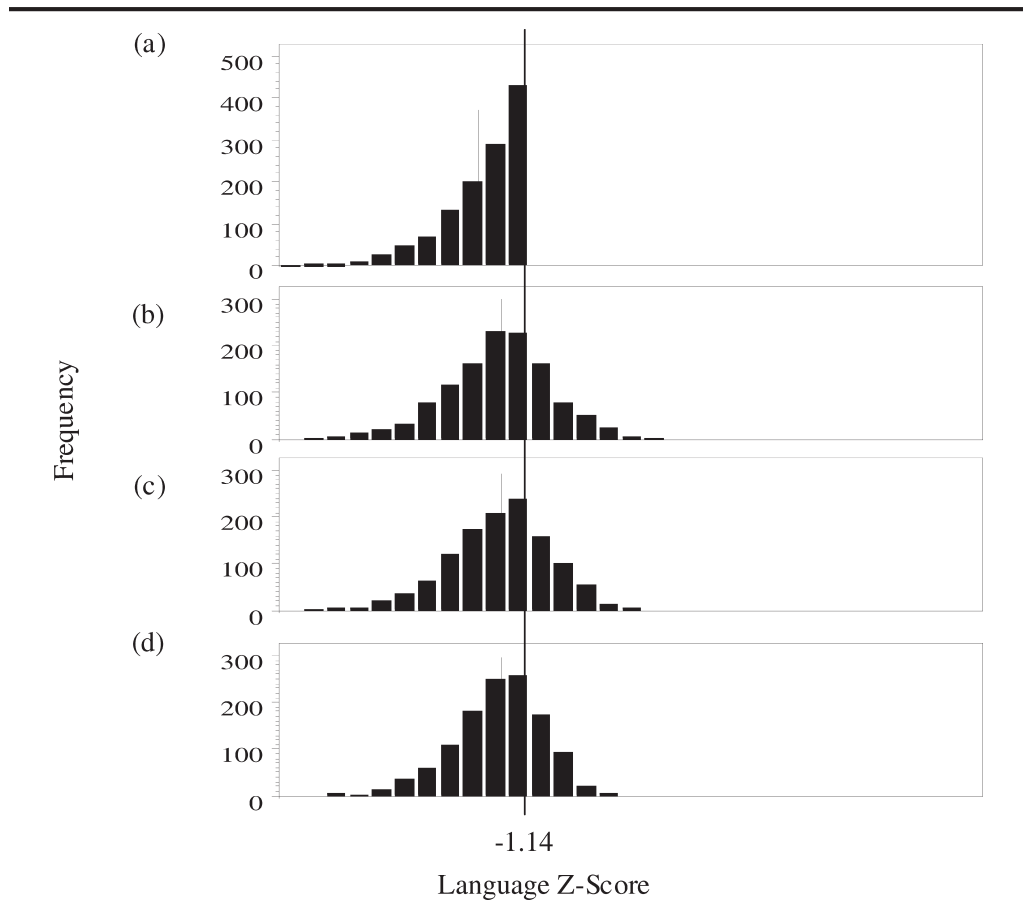
We have shown that where research is concerned with changes in mean scores, the regression effect can be modest. In research designs concerning changes in discrete clinical states, the illusory recovery rate due to regression effects can be dramatic. We can show this by

computing the illusory recovery rate, using the approach described in Appendix C. In our earlier example of a change of mean z score of 0.20, where the correlation between X and Y is .90, it can be predicted that the false recovery rate would be 35.2%. Thus, nearly a third of the original cases who were thought to be affected on the X measure were determined to be unaffected on the Y measure. This change in status, however, was not due to any changes in true scores on the trait underlying X and Y . Thus, even with highly reliable measures, regression to the mean can produce rather substantial erroneous estimates of recovery. Why is this so?

As shown in Figure 3a, this is because a group sampled because they have scores on X that are below a cut-off such as -1.14 usually does not form a regular distribution, such as a normal distribution. Instead, the density right below -1.14 is much higher than further away from this cut-off point. Among those individuals who had a score right below the cut-off, many may have a true score above the cut-off. The small difference between their obtained scores and their true scores is enough to make them false-positive cases. On the other hand, not all those affected on X but unaffected on Y were true false positives because Y is not the "gold standard." Some of them may be true positives to begin with, but became false negatives on Y . No matter which is the case, the key point here is that X and Y are measuring the same true score and the true score has never changed from X to Y . All affection status changes were due to measurement error; those individuals with scores right below the cut-off—milder cases—were more prone to the status change.

It may appear that we have reached contradictory conclusions when talking about the interaction of severity and the effects of regression to the mean. We just stated that among a group followed for recovery rate, false recovery would be concentrated in the milder cases. Earlier, we noted that a group of more severe cases is more likely to show greater rates of false change in mean scores than a group selected because they are generally milder cases. These are not contradictory statements. In the first case, we considered sampling two groups based on different values of a and showed that the mean score of the more severe group would have greater regression than the mean score of the less severe group. In the second case, we considered the situation where a single group was selected due to scores below some cut-off (a) and later evaluated to determine how many in the group had scores $Y > a$. Here the measure of change was the proportion of individuals who in the first case had scores $< a$ and moved to the other side of the cut-off on Y . In this situation, the milder cases were more likely to contribute to this change. The effect of regression to the mean can be complicated; only by thoroughly understanding the basis of this effect can one identify when and how it may affect a particular research design.

Figure 3. The empirical distributions of the simulated data for (a) diagnostic measure X , (b) baseline Z , (c) follow-up Y , and (d) the true score T for the group with $X < -1.14$ SDs ($N = 1,215$). The mean for the distribution is denoted by a vertical line within the distribution. The long vertical line across the four distributions corresponds to -1.14 .



Knowing that the results are prone to erroneous evidence of change is a first step toward managing this effect in research.

Approaches to Controlling Regression to the Mean

We have shown above that if a group is sampled because members of the group possess scores on X such that $X < a$, their subsequent mean score on Y will be higher than the mean X score. Then how can one distinguish between regression toward the mean and true improvement? As we noted in the introduction, there are two classes of longitudinal research where regression to the mean can affect results. In each of these cases, the interest is in measuring true change in the participants. One type of longitudinal design asks whether an intervention is associated with change. In such a study, regression effects can be controlled by randomly assigning participants to a treatment or control arm of the study. If the measure used to select the

participants is used as the pretreatment baseline against which posttreatment effects are compared, both groups should regress to the mean. The treatment effect will be seen as an interaction between pre–post measures and group, or as simply the group difference in the posttreatment measure if the randomization had achieved balance in the pretreatment measure. That is, the treatment effect will show up as an additional amount of the pre–post contrast unique to the treatment group.

Some longitudinal designs do not involve an intervention. In these cases it is not possible to use a control group to manage the regression effect; such studies are often concerned with patterns of growth or recovery in untreated clinical populations. As there is no logical control group available in this design, we need an alternative. A solution in such studies is to keep the measures used to select participants separate from the measure(s) used as baseline measure(s). Remember that the regression effect occurs as a result of sampling. If you sample on one measure and track change using a separate baseline measure, the bias from the sampling will not

be present on the baseline measure. Below, we show that in such a case the baseline measure will provide an unbiased estimate of initial levels of performance, even in a group that is low or high on the trait measured by the baseline measure.

To begin, let us designate three measures. These measures will be the diagnostic measure (X), the baseline measure (Z), and the follow-up measure (Y). These measures could consist of the same measurement instrument or method, or of parallel forms, but ideally they are all highly correlated and measure the trait equally well. We use X to select our participants. Therefore, as we have shown, the scores on X will be biased estimates of the true scores of this sample and thus will be prone to regression effects. Z (the baseline) should be administered at the time when the participants begin the longitudinal study. Finally, at T2, the follow-up measure Y is obtained. Although X and Z both contain measurement error, only X scores contain a biased sample of measurement error for this sample of participants. The distribution of Z scores is unbiased because no sampling operation was performed on these scores. Note that this design will only work so long as no selection of participants was done based on the baseline (Z) score. It is very tempting for a researcher to “clean up” a sample of individuals at entry into the longitudinal study using the baseline data. By doing this, the researcher will reintroduce regression effects. With such a design, one can measure the rate of change by computing the difference between Y and Z . This difference would represent unbiased estimates of true change over this time. This design also would permit a measure of recovery rate. To do this, the rate of cases based on the baseline Z measure would be compared with the rate of cases based on the outcome Y measure. Only if these rates are significantly different would we consider the existence of a significant true recovery and that this difference represents an unbiased estimate of recovery rate.

One way we can demonstrate that such a design can control for regression to mean effects is by simulation. The advantage of simulation is that we know what the true scores are for a group of hypothetical participants. Within this simulation we attempt to parallel research problems we have encountered in our research on specific language impairment. Thus, we are simulating a longitudinal follow-up of a hypothetical group of children with language impairment.

Method

Simulated Samples

In this simulation, we assumed that the true scores (T) of the trait measured by a test with a reliability of .90 followed a normal distribution. Three equivalent forms of the test were given. The target intercorrelation

between the three test scores was the test reliability (i.e., .90). The simulation sample size was 10,000, and each of the individuals was assigned a true score and three obtained scores.

The three test forms were given on three occasions (one at a time), representing a diagnostic observation (X), a baseline observation (Z), and a follow-up observation (Y). In this simulation, the true score for an individual remained the same across the three measures, but measurement error for three measures was independent. The test reliability (.90) determines the intercorrelation of X , Y , and Z . The only source of variation across these three measures, within an individual, was random measurement error.

The sample was generated using SAS/IML. The simulation process was as follows:

1. A true score was generated for each of the 10,000 individuals by sampling from a normal distribution with mean of zero and variance equal to the test reliability value (.90). In real life, each individual has an inherent true characteristic value, though it is unobservable.
2. An obtained score, X , was generated for each of the 10,000 individuals by sampling from a normal distribution with a mean equal to the individual's true score generated in Step 1 and variance equal to 0.10 (the error variance = 1 minus the test reliability, .90). Thus, $X \sim n(0, 1)$, which means X follows a standard normal distribution. Because this score was used for diagnosis, we designated it as a diagnostic measure.
3. Each of the 10,000 individuals was diagnosed based on the obtained score, X . Individuals with an obtained score (X) below -1.14 were diagnosed as having language impairment, simulating a criterion we used in diagnosis of language impairment (Tomblin, Records, & Zhang, 1996).
4. For those individuals with language impairment, two additional obtained language scores, Z and Y , were generated following the method described in Step 2. Variable Z was designated as the baseline, and Y was the follow-up. The population correlation between X , Y , and Z was set to be equal to the test reliability. To check whether the correlation was desirable, these scores were generated for those with normally developed language in the same manner as for individuals with language impairment.

Statistical Analysis

Descriptive Statistics

The means and standard deviations for the three obtained language scores (X , Z , and Y) and the correlation among these variables were calculated. These

statistics provided a check to show that the data generation routine had worked as desired.

Regression Effect on the Group Mean and Illusory Recovery

The mean of X for the group with language impairment was compared to the mean of Z and Y . The rate of change from language impairment on X to normally developed language on Z and on Y was also obtained.

Results

Descriptive Statistics

X , Y , and Z scores in this simulation were generated as samples of a standard normal distribution, $n(0, 1)$, meaning that the sample means should all be close to 0 and the standard deviations should all be close to 1. Data in Table 1 show that the sample means and standard deviations were very close to these values. Thus, the simulation routine did generate data sets that conformed to our expectation. The sample correlation of X with Y and Z was .896 and .897, respectively, which was very close to the target reliability of the test. Thus, the covariance among the measures was also as expected.

Change of Mean From Diagnostic to Baseline and Follow-Up Tests

The distributions of obtained score X , Y , and Z and the true score T for the group of 1,215 individuals with $X < -1.14$ are shown in Figure 3. Figure 3a is the distribution of X for the group selected on X (i.e., $X < -1.14$); that is why there is no density above the long vertical line corresponding to -1.14 . The mean of X for this group of individuals, designated by the short vertical line in Figure 3a, was lower than the mean of Z (baseline measure), Y (follow-up measure), and T (true score), designated by the short vertical lines in Figures 3b, 3c, and 3d, respectively. Thus, the diagnostic test is a biased estimate of the true scores for this group and the baseline and the follow-up are unbiased. The data in Figure 3 show that from the diagnostic test (X) to the baseline

Table 1. Means and standard deviations of scores generated for the measure of a hypothetical trait T for a sample of 10,000 individuals across three observations.

Variable	M	SD
X	-0.00376	0.983
Y	-0.00254	0.987
Z	-0.00434	0.986

Note. Test reliability is .90

test (Z) or the follow-up (Y), there was a noticeable amount of change in the direction of the population mean of zero, but there was little change from Z to Y . This says that with a constant true score (T) there won't be significant change from the baseline measure to the follow-up measure, but there will be a significant change from the diagnostic to the follow-up without any true score changes. Thus, the change from the diagnostic to the follow-up is a false indication of improvement.

Illusory Recovery From Diagnostic to Follow-Up

Because of regression toward the mean, shown in Figure 3, we can predict some apparent recovery from initial diagnosis to the follow-up diagnosis. That is, there will be some individuals diagnosed as having language impairment on X (diagnosis; $X < -1.14$) but normal on Y (follow-up; $Y > -1.14$) when we use the same cut-off value for diagnosis at both time points. This prediction was confirmed; results are shown in Table 2 and Figure 3. The simulation showed that 348 (28.64%) of the 1,215 initially diagnosed as having language impairment were classified as having normally developed language at baseline, and a similar number were diagnosed as normally developing at follow-up. Illusory recovery is shown as the part of the distribution right of the long vertical line corresponding to -1.14 in Figures 3b and 3c. Note that unlike the continuous variable where the follow-up test mean was very close to the true score mean, the proportion of normally developing individuals at the follow-up was much higher than the true proportion. This difference is explained below.

Discussion

The purpose of this simulation was to provide an idealized example of how regression to the mean generates patterns in data that may be erroneously interpreted as evidence of improvement. The simulation provided a concrete example of some of the points made earlier.

Table 2. Number and percentage of individuals initially diagnosed as having language impairment who regressed to language normal status at baseline and at follow-up, and the number and percentage of true language normal status among this group.

N	Normal at baseline		Normal at follow-up		True status is normal	
	n	%	n	%	n	%
1,215	341	28.07	348	28.64	262	21.56

Note. N = number impaired at diagnosis; numbers normal at baseline, at follow-up, and in true status are out of N .

First, it showed that the mean score of X was a biased estimator of the true score of the sample of children who had X scores below -1.14 . In the real world one can never know what the true scores should be, but in a simulation we have the luxury of knowing this. Thus, we can state that the basis for this bias is that the process of sampling children with low scores oversamples instances in which the obtained score on X is below the true score T . The magnitude of this bias in the estimate of the mean true score would increase as test reliability decreases (see Appendix B).

Another feature of the simulation was shown in the analysis of recovery. We noted that even when the regression effects appear to be small with regard to estimates of mean scores, change in a dichotomous outcome could appear large. This pattern is well exemplified in the case of the simulated test, which had a reliability of .90—a level of reliability that is viewed as very desirable for measurement purposes. Even with such a measure, more than a quarter of the initial cases turned out to be re-diagnosed as within normal range on Y . Thus, on a second examination, these cases appeared to have improved, when in fact they were either incorrectly diagnosed to begin with or were incorrectly classified on Y . It is difficult to achieve high levels of diagnostic accuracy when the latent trait is actually a quantitative trait. Dichotomizing a continuous trait based on scores that deviate from the population mean may result in relatively high levels of classification error.

The simulation also showed that by including a separate baseline in a longitudinal design it is possible to control for the regression effects. If the baseline measure is used as the measure of initial status and the follow-up measure is contrasted with this baseline measure, it is possible to obtain an unbiased estimate of the extent to which there was a change in the true scores of the participants. It is important to emphasize that the baseline measure is not a better measure of the latent trait M than is the diagnostic measure X . In this case they were simulated to have the same quality. The reason that the Z measure was unbiased and the X measure was biased was because the selection operation was performed using X to form the sample of individuals with language impairment. This shows that once individuals have been placed into the cohort to be followed, they cannot be excluded due to their performance on the baseline or any subsequent measure. Sampling on measures that are also used to monitor change is the key source of the regression effect and thus must be avoided.

As we noted earlier, in this simulation we knew what the true score was for each of the simulated individuals. In a real world research situation, we often want to be able to do the same thing. That is, we often want to determine which members of a sample are true cases

or which individuals have truly changed due to treatment or recovery. It would be tempting to view the design just presented as offering a solution to this. That is, because the baseline provides an unbiased estimate of the true score, can we then look at those individuals who had scores below -1.14 and say that they are true cases? In fact, we can't. Recall that Z was no better a measure of M than was X . Although the mean of Z is an unbiased estimator of the mean of the true score T , any given observation z selected because it falls below some cut-off is likely to be a biased estimate of t . Thus, we can never know the true status of an individual.

A Dilemma in Categorical Diagnosis

The simulation and mathematical proofs have shown that the mean test score on a measure such as Z is an unbiased estimator of the true score for the group with $X < -1.14$. That is, the mean of this score will be equal to the true score mean for this group. However, the simulation also revealed a paradox. If a categorical diagnosis is made on this new measure, the rate of noncases, or individuals considered normally developing, will not be equal to the rate of noncases in the true status (as shown in Table 2). The percentage of individuals diagnosed as normally developing on follow-up is much higher than it should be (348 vs. 262) when the true scores were used for diagnosis. Why is the mean score an unbiased estimator of true score mean, but the rate of normal status is not? The expectation of the obtained score is the person's true score if he or she was not selected for having a low or high obtained score. Overestimation and underestimation of a given true score will be cancelled out when the mean score is calculated. However, for the categorical (binary) diagnosis, the only thing that matters is whether the value is below or above the cut-off point. For the rate of normally developing individuals on Z to be the same as on T (true score), there must be the same number of individuals with two types of misdiagnosis: false positive and false negative. Note that individuals who have a true status of "normal" are candidates for false positive, whereas those with a true status of "impaired" are candidates for false negative. As shown in Table 2, among the 1,215 diagnosed as impaired on X , 262 had a true status of normal and 999 had a true status of impaired. Thus, the diagnosis on Z or any other measures, when applied to this sample, will generate many more false negatives than false positives and hence overestimate the rate of normal, or recovery rate. Nevertheless, if rate change across time, such as from Z to Y or one follow-up to another, is of concern, then the overestimation of recovery in the two measures should be canceled out in the difference between the two measures. In this sense,

the baseline measure is still useful for tracking status changes, though the rate at either point is not an unbiased estimator of the absolute true rate of recovery.

Conclusions

This tutorial was intended to fully explore the nature of regression to the mean and ways in which it can be managed in research designs that involve repeated measures of participants who have been sampled because they deviate from the mean on some trait. We have shown that the basis for regression to the mean is quite straightforward, but the ramifications of this effect can be considerable and sometimes subtle. An understanding of the nature of regression to the mean will allow researchers to avoid its effects on their results.

Acknowledgment

This work was supported by a clinical research center grant P0-DC-02748 from the National Institute on Deafness and Other Communication Disorders.

References

- Blommers, P. J., & Forsyth, R. A.** (1977). *Elementary statistical methods in psychology and education* (2nd ed.). Lanham, MD: University Press of America
- Casella, G., & Berger, R. L.** (1990). *Statistical inference*. Belmont, CA: Duxbury.
- Cook, T. D., & Campbell, D. T.** (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.
- Tomblin, J. B., Records, N. L., & Zhang, X.** (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research, 39*, 1284–1294.

Received March 14, 2003

Accepted July 15, 2003

DOI: 10.1044/1092-4388(2003/104)

Contact author: Xuyang Zhang, PhD, Department of Speech Pathology and Audiology, The University of Iowa, Iowa City, IA 52242. E-mail: xuyang-zhang@uiowa.edu

Appendix A (p. 1 of 2). Bivariate normality of joint distribution of X and T or X and Y and conditional distributions.

Let the true score be denoted as T and the scores from two alternate forms denoted as X and Y . Assume that the two obtained scores have been standardized into scores with mean of 0 and standard deviation of 1. Assume further that the reliability of the test is ρ_{XY} , the correlation between X and Y .

The conditional distribution of X or Y given $T = t$ is normal with mean of t and variance equal to the measurement error variance, $(1 - \rho_{XY})$, that is, $X|T=t$ or $Y|T=t \sim n(t, (1 - \rho_{XY}))$. Because the variance of X or Y is 1 and the reliability is the proportion of true score variance in the obtained score variance, the variance of T is ρ_{XY} . Thus, we have

$$f(x|T=t) = \frac{1}{\sqrt{2\pi}\sqrt{1-\rho_{XY}}} \exp\left\{-\frac{(x-t)^2}{2(1-\rho_{XY})}\right\}$$

$$f(y|T=t) = \frac{1}{\sqrt{2\pi}\sqrt{1-\rho_{XY}}} \exp\left\{-\frac{(y-t)^2}{2(1-\rho_{XY})}\right\}$$

$$f(t) = \frac{1}{\sqrt{2\pi}\sqrt{\rho_{XY}}} \exp\left\{-\frac{t^2}{2\rho_{XY}}\right\}.$$

The probability density of the X and T joint distribution is

$$f(x,t) = f(x|t)f(t)$$

$$= \frac{1}{2\pi\sqrt{\rho_{XY}(1-\rho_{XY})}} \exp\left\{-\frac{\rho_{XY}(x^2 - 2xt + t^2) + (1-\rho_{XY})t^2}{2\rho_{XY}(1-\rho_{XY})}\right\}$$

$$= \frac{1}{2\pi\sqrt{\rho_{XY}(1-\rho_{XT}^2)}} \exp\left\{-\frac{1}{2(1-\rho_{XY})}\left(x^2 + \frac{t^2}{\rho_{XY}} - 2xt\right)\right\}.$$

The correlation between X and T squared, ρ_{XT}^2 , is the variance of T or test reliability, ρ_{XY} . Hence,

$$f(x,t) = \frac{1}{2\pi\sigma_X\sigma_T\sqrt{(1-\rho_{XT}^2)}} \cdot \exp\left\{-\frac{1}{2(1-\rho_{XT}^2)}\left[\left(\frac{x}{\sigma_X}\right)^2 + \left(\frac{t}{\sigma_T}\right)^2 - 2\rho_{XT}\frac{x}{\sigma_X}\frac{t}{\sigma_T}\right]\right\}$$

(because $\sigma_X^2 = 1$, $\sigma_T^2 = \rho_{XY}$, and $\rho_{XT}^2 = \rho_{XY}$).

This is the density function for a bivariate normal distribution of X and T (Casella & Berger, 1990, p. 167). Y is equivalent to X , so the joint distribution of Y and T is bivariate normal too.

It can be shown that X and Y jointly follow a bivariate normal distribution with density function

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho_{XY}^2}} \exp\left\{-\frac{1}{2(1-\rho_{XY}^2)}(x^2 + y^2 - 2\rho_{XY}xy)\right\}.$$

Proof:

On the basis of independence between X and Y given $T = t$,

$$f(x,y|T=t) = f(x|T=t)f(y|T=t)$$

$$= \frac{1}{2\pi(1-\rho_{XY})} \exp\left\{-\frac{(x-t)^2 + (y-t)^2}{2(1-\rho_{XY})}\right\}$$

Appendix A (p. 2 of 2). Bivariate normality of joint distribution of X and T or X and Y and conditional distributions.

$$\begin{aligned}
 f(x, y, t) &= f(x, y | T = t) f(t) \\
 &= \frac{1}{2\pi(1 - \rho_{XY})} \exp \left\{ -\frac{(x-t)^2 + (y-t)^2}{2(1 - \rho_{XY})} \right\} \\
 &\quad \cdot \frac{1}{\sqrt{2\pi}\sqrt{\rho_{XY}}} \exp \left\{ -\frac{t^2}{2\rho_{XY}} \right\} \\
 &= \frac{1}{2\pi(1 - \rho_{XY})\sqrt{2\pi}\sqrt{\rho_{XY}}} \exp \left\{ -\frac{x^2 + y^2}{2(1 - \rho_{XY})} \right\} \\
 &\quad \cdot \exp \left\{ -\frac{t^2 - 2\frac{\rho_{XY}}{1 + \rho_{XY}}(x + y)t}{\frac{2\rho_{XY}(1 - \rho_{XY})}{(1 + \rho_{XY})}} \right\}
 \end{aligned}$$

$$\begin{aligned}
 f(x, y) &= \int_{-\infty}^{\infty} f(x, y, t) dt \\
 &= \frac{1}{2\pi\sqrt{1 - \rho_{XY}^2}} \exp \left\{ -\frac{1}{2(1 - \rho_{XY}^2)}(x^2 + y^2 - 2\rho_{XY}xy) \right\},
 \end{aligned}$$

which is the density function for a standard bivariate normal distribution.

Given that X and T jointly follow a bivariate normal distribution, the conditional distribution of T given X = x is

$n(\rho_{XY}x, \rho_{XY}(1 - \rho_{XY}))$, a normal distribution with mean of the test reliability times x and variance is $\rho_{XY}(1 - \rho_{XY})$. Because $0 < \rho_{XY} < 1$, the mean of the distribution, $\rho_{XY}x$, is closer to the population mean, 0, than is x.

As shown above, X and Y also jointly follow a bivariate normal distribution. The conditional distribution of Y given X = x is

$$n(\rho_{XY}x, (1 - \rho_{XY}^2)).$$

This is a normal distribution with mean of $\rho_{XY}x$ and variance of $(1 - \rho_{XY}^2)$.

Because $0 < \rho_{XY} < 1$, the mean of the distribution, $\rho_{XY}x$, is closer to the population mean, 0, than is x.

Note that the mean of the conditional distribution of T given X = x is equal to that of Y. This is why a baseline measure, an alternate form test or retest score, can be used as an unbiased estimate of the true scores for the group selected on X (diagnostic test scores). This mathematical formalism is based on the assumption of univariate normality of the distribution of X, Z, and Y given T, which also is normally distributed. This assumption makes the formulation easier, but it is not a necessary condition for regression toward the mean to occur. This statement also applies to Appendixes B and C that follow.

Appendix B. Regression toward the mean causes illusory performance improvement.

It has been shown in Appendix A that X and Y follow a bivariate normal distribution. Both X and Y have been transformed into a standard normal scale (mean of 0 and standard deviation of 1). Let the probability density of X, Y, and (X, Y) jointly be denoted as $f_X(x)$, $f_Y(y)$, and $f(x, y)$ respectively. To show that regression toward the mean caused illusory performance improvement is to show that $E(Y | X < a) > E(X | X < a)$, which reads that the expectation (i.e., mean) of Y given X < a is greater than the expectation of X given X < a.

$$\begin{aligned}
 E(Y | X < a) &= \frac{\int_{-\infty}^a \int_{-\infty}^{\infty} yf(x, y) dy dx}{\int_{-\infty}^a \int_{-\infty}^{\infty} f(x, y) dy dx} \\
 &= \frac{\int_{-\infty}^a \int_{-\infty}^{\infty} (\rho_{XY}x + e)f(x, y) dy dx}{\int_{-\infty}^a f_X(x) dx}
 \end{aligned}$$

($y = \rho_{XY}x + e$ where e is the regression residual)

$$\begin{aligned}
 &= \frac{\int_{-\infty}^a (\rho_{XY}x + e) \int_{-\infty}^{\infty} f(x, y) dy dx}{\int_{-\infty}^a f_X(x) dx} \\
 &= \frac{\int_{-\infty}^a \rho_{XY}xf_X(x) dx + \int_{-\infty}^a ef_X(x) dx}{\int_{-\infty}^a f_X(x) dx} = \rho_{XY}E(X | X < a)
 \end{aligned}$$

$> E(X | X < a)$ if $0 < \rho_{XY} < 1$ and $E(X | X < a) < 0$.

Because the population mean is 0, the mean of X with those Xs $\geq a$ excluded must be less than zero. Therefore, $0 < \rho_{XY} < 1$ and $E(X | X < a) < 0$ entails $E(Y | X < a) = \rho_{XY}E(X | X < a) > E(X | X < a)$, that is, regression toward the mean occurs for the group selected based on $X < a$.

Appendix C. Calculation of illusory recovery rate.

The density function of Y given $X < a$ is

$$\begin{aligned} f(y | x < a) &= \frac{f(y \cap x < a)}{F_x(a)} \\ &= \frac{\Pr(x < a | y) f_Y(y)}{F_x(a)} \\ &= \frac{F_{X|Y=y}(a) f_Y(y)}{F_x(a)}, \end{aligned}$$

where $F_{X|Y=y}(a)$ is the cumulative distribution function for $X | (Y = y) \sim n(\rho_{XY}y, (1 - \rho_{XY}^2))$, $f_Y(y)$ is the probability density function

for $Y \sim n(0, 1)$, and $F_x(a)$ is the cumulative distribution function for $X \sim n(0, 1)$.

The illusory recovery rate is

$$\int_a^\infty f(y | x < a) dy.$$

Note that, as long as $\rho_{XY} < 1$, the probability of Y being greater than a given $X < a$ will be greater than 0. Thus, illusory recovery will be most likely to occur when the test is not perfectly reliable.

Explaining and Controlling Regression to the Mean in Longitudinal Research Designs

Xuyang Zhang, and J. Bruce Tomblin
J Speech Lang Hear Res 2003;46:1340-1351
DOI: 10.1044/1092-4388(2003/104)

This article has been cited by 1 article(s) which you can access for free at:
<http://jslhr.asha.org/cgi/content/abstract/46/6/1340#otherarticles>

This information is current as of May 10, 2010

This article, along with updated information and services, is
located on the World Wide Web at:
<http://jslhr.asha.org/cgi/content/abstract/46/6/1340>

