On the Standard Error of the Difference between Two Independent Regression Coefficients in Moderation Analysis: A Commentary on Robinson, Tomek, and Schumacker (2013)

Andrew F. Hayes **Robert A. Agler**
The Ohio State University

In this journal, Robinson, Tomek, and Schumacker (2013) argued that researchers interested in whether the regression coefficient for X in a model estimating Y from X differs between two groups should conduct two separate regressions rather than rely on moderated multiple regression (MMR). They advocate a standard error of the difference between coefficients they claim results in a more powerful test of moderation than MMR without affecting Type I error rate. We show analytically and demonstrate through simulation that, consistent with prior research, their standard error estimator results in substantial underestimation of the standard error and should not be used.

Establishing the boundary conditions of independent variable *X*'s effect on dependent variable *Y*—the factor or factors *Z* that influence or predict the size of *X*'s effect on *Y*—advances our understanding of a given phenomenon more so than does establishing merely that *X* is related to *Y* on average. Are grades in high school (*X*) more predictive of later superior performance in college (*Y*) among first-generation college students than among children of parents who attended college themselves (*Z*)? Does the relationship between frequency of television viewing at home (*X*) and attention problems in school (*Y*) depend on whether or not a child has a television in the same room where he or she studies (*Z*)? If so, we say that *Z moderates* the effect of *X* on *Y*, or that *X* and *Z interact* in predicting *Y*. In other words, the effect of *X* on *Y* depends on *Z*.

Education researchers get exposure to methods for testing such questions about moderation or interaction—the contingencies of an effect—early in their training in the form of factorial analysis of variance in which all variables but the outcome are categorical. Alternatively, this concept is introduced in a regression analysis class by showing how a continuous or dichotomous *X*'s regression weight in a model of a continuous *Y* can be estimated as a linear function of a dichotomous or continuous proposed *moderator variable Z*.

The linear moderation model, sometimes called a *moderated multiple regression model*, is frequently used in education research (see e.g., Aspelmeier, Love, McGill, Elliot, & Pierce, 2012; Farris, Lefever, Borkowski, & Whitman, 2013). It takes the form

$$Y = i_1 + b_1X + b_2Z + b_3XZ + e \,, \qquad (1)$$

where *X* and *Z* are an independent variable and a moderator variable, respectively, that are either dichotomous or continuous, *XZ* is their product, *Y* is a continuous dependent variable, and *e* is an error in estimation. The inclusion of *XZ* as a predictor in Equation 1 along with *X* and *Z* allows *X*'s effect on *Y* to be a linear function of *Z*, meaning *Z* serves as a moderator of the effect of *X* on *Y*. If *X* and *Z* are both dichotomous, ordinary least squares (OLS) estimation of the regression coefficients in Equation 1 is equivalent to a 2 × 2 factorial analysis of variance when *X* and *Z* are effect coded (e.g., 1/-1 or -0.5/0.5; see Hayes, 2013).

This regression-based approach to moderation analysis has received much attention in the methodology literature. This is no doubt in part because moderation is such an important concept in most any substantive area, but this attention also reflects various confusions and controversies about how the regression coefficients in such a model are interpreted, as well as whether and how the regression coefficient for *XZ* quantifies the relationship between the size of *X*'s effect on *Y* and moderator variable *Z* and the power of tests of moderation using this approach (see e.g., Aguinis & Stone-Romero, 1997; Cohen, 1978; Friedrich, 1982; Hayes, Glynn, & Huge, 2012; Kromrey & Foster-Johnson, 1998). This commentary adds to that literature and focuses on a special form of this model in which *Z*, the proposed moderator, is a dichotomous variable. It is written in response to an article published in this journal by Robinson, Tomek, and Schumacker (2013) that we believe offers advice that, if followed, will result in investigators conducting an inaccurate test of whether *X*'s effect on *Y* differs between the two groups coded by *Z*. We offer analytical and simulation based evidence to support the position we take.

Separate Regressions Versus Moderated Multiple Regression

If Z is coded 0 and 1, b_1 in Equation 1 estimates the linear relationship between X and Y in the group coded $Z = 0$, and $b_1 + b_3$ estimates the linear relationship between X and Y in the group coded $Z = 1$. Therefore, the difference between the two groups in the relationship between X and Y, at least as quantified with a linear regression weight, is equal to b_3. An inference about b_3 is an inference as to whether Z moderates the effect of X on Y. That is, a claim that the regression coefficient for XZ in Equation 1 is different from zero is a claim that the linear relationship between X and Y differs between the two groups coded with Z. We will refer to this analytical strategy throughout as the *moderated multiple regression* (MMR) approach to moderation analysis.

An alternative approach has been used in some published studies in the education field (e.g., Loes, Salisbury, & Pascarella, 2013; Roksa & Potter, 2011), which we will refer to as the *separate regressions* approach. This method involves conducting two independent regression analyses, one for the group coded $Z = 0$, and the one for the group coded $Z = 1$, with the goal of seeing if the relationship between X and Y differs in the two regressions. If one were to discard all the cases from the data coded $Z = 1$ and then regressed Y on X, as in

$$Y = i_1 + b_1X + e \tag{2}$$

it can be shown that b_1 in Equation 2 is equivalent to b_1 in Equation 1. That is, in (2), b_1 quantifies the relationship between X and Y in the group coded $Z = 0$, just as does b_1 in Equation 1. Similarly, if one were to exclude cases with $Z = 0$ and estimate

$$Y = i_2 + b_2X + e \tag{3}$$

on the remaining cases, b_2 in Equation 3 estimates the effect of X on Y in the group coded $Z = 1$. It also turns out that b_2 in Equation 3 is equivalent to $b_1 + b_3$ in Equation 1. Of interest when using the separate regression approach to moderation analysis is whether b_1 in Equation 2 differs from b_2 in Equation 3 according to a formal statistical test.

In the MMR strategy, b_3 is equivalent to the difference between b_2 and b_1 from Equations 2 and 3 and directly quantifies the difference in these two regression weights. Conveniently, such a regression analysis also yields a standard error for b_3 that can be used for statistical inference, and any regression program will generate it along with a t and p-value whether the analyst wants it or not. The standard error estimator built into all commonly-used regression routines such as in SAS, SPSS, and other packages is

$$se_{b_3} = \sqrt{\frac{MS_{residual}}{n\left(1 - R^2_{XZ}\right)\hat{V}(XZ)}} \ , \tag{4}$$

where $MS_{residual}$ is the mean squared residual from Equation 1, n is the sample size, R^2_{XZ} is the squared multiple correlation from a linear regression analysis estimating XZ from X and Z, and $\hat{V}(XZ)$ is the estimated variance of XZ (see e.g., Darlington, 1990). Under the standard assumptions of regression (normally, independently, and identically distributed errors in estimation) b_3 / se_{b_3} is distributed as $t(n - 4)$ where n is the sample size. A p-value for the test of the null hypothesis that β_3, the population counterpart of b_3, is equal to zero can be derived from the t distribution, or a c% confidence interval can be constructed as $b_3 \pm t_{crit} se_{b_3}$, where t_{crit} is the value of t that cuts off the upper $(100 - c) / 2$% of the t distribution from the rest.[1]

Robinson et al. (2013) recommend avoiding the MMR approach when interest is on testing whether the regression slope estimating Y from X is different in two groups. Rather, they advocate the separate regressions approach, which unlike the MMR approach does not automatically generate a standard error for $b_2 - b_1$. Instead, b_1 and b_2 are estimated in separate regressions, each of which yields a standard error for its respective regression slope, and the standard error for the difference between b_2 and b_1 must be calculated by hand or some other way. They advocate

$$se_{b_2 - b_1} = \sqrt{\frac{n_1\hat{V}(b_1) + n_2\hat{V}(b_2)}{n_1 + n_2 - 2}} \ , \tag{5}$$

as the estimator of the standard error of the difference[2], where n_1 and n_2 are the sizes of the two groups and $\hat{V}(b_1)$ and $\hat{V}(b_2)$ are estimated sampling variances—the squared standard errors—of b_1 and b_2 from the separate regressions of Y on X. With this standard error estimated, a t test of the difference between

regression coefficients can be conducted by deriving the *p*-value attached to $(b_2 - b_1)/se_{b_2-b_1}$ using the *t* distribution with $n_1 + n_2 - 2$ degrees of freedom. Alternatively, a *c*% confidence interval can be constructed as $(b_2 - b_1) \pm t_{crit} se_{b_2-b_1}$, where t_{crit} is the value of *t* that cuts off the upper (100 – *c*) / 2% of the *t* distribution from the rest..

Robinson et al. (2013) argue that the separate regressions approach to estimating the standard error of $b_2 - b_1$ is better than the MMR approach using Equation 4 for the standard error of b_3. They show by way of two examples that although $b_2 - b_1$ from Equations 2 and 3 correspond to b_3 in Equation 1, the standard error in Equation 5 is smaller than the standard error that Equation 4 yields. Thus, they reason, the use of the standard error estimator in Equation 5 produces a test with higher power while still keeping the Type I error in line with the nominal significance level.

On the surface their argument seems compelling. Indeed, Equation 5 will produce a smaller standard error than Equation 4, so the *p*-value using Equation 5 is smaller than when using Equation 4, and the confidence interval for the difference in regression weights is correspondingly narrower. This is apparent following the two examples Robinson et al. (2013) report. We concede that this is generally true, and the reader can verify this using any data that they have available to them that this is not specific to the two examples Robinson et al. provide. Furthermore, we agree that the approach Robinson et al. advocate is a more powerful way of testing the difference between two independent regression slopes and therefore will reduce the risk of a Type II error.

The difficulty we have with their recommendation is that it yields a more powerful test in a way that most researchers would consider unacceptable. The problem with this estimator is that it is downwardly biased, and it yields an estimate of the standard error that is systematically inaccurate and inappropriately small. Using a standard error that is too small is certainly one approach to increasing power, but use of the standard error they describe does not control Type I error rate as Robinson et al. (2013) claim. The cost of the power that the use of Equation 5 buys is *elevated* Type I error and confidence intervals with coverage that do not match the confidence level desired, contrary to Robinson et al.'s claim. Although investigators can differ in how much they weigh the cost of Type I over Type II errors, we believe that most would find that the elevated Type I error rate that the use of Equation 5 yields is too high a price to pay for the power it buys.

In this remainder of this paper, we illustrate the downward bias analytically and through simulation, and show that the result is inflated Type I error rates and confidence intervals with coverage that does not coincide with the confidence level. We also show that the standard error of b_3 from the MMR approach is generally a better estimator of the standard error of the difference between the two within-group regression slopes, so long as the assumption of homoscedasticity is met. When it is violated, the problem produced by heteroscedasticity in the MMR approach is easily remedied by using a standard error estimator in moderated multiple regression that does not assume homoscedasticity.

We acknowledge that we are not the first to make much of this argument we advance in the pages that follow. Indeed, the potential controversy this commentary may prompt has largely already been settled elsewhere (see Brame, Paternoster, Mazerolle, & Piquero, 1998; Cohen, 1983; Paternoster, Brame, Mazerolle, & Piquero, 1998). We will review some of the evidence here. Understandable though it is, Robinson et al.'s (2013) recommendation suggests that the resolution of this debate has not yet disseminated widely, at least not through the education discipline.

In the discussion below, we make frequent reference to two examples that Robinson et al. (2013) rely on in their paper. The first example is based on a study examining the relationship between academic self-efficacy (*X*) and academic achievement (*Y*) as a function of ethnicity (*Z*). The second is based on data available on the web looking at the relationship between a continuous measure of cancer risk (*X*) and intentions to get screened for cancer (*Y*) as a function of a dichotomous measure of risk (*Z*).[3] Robinson et al. (2013) report analyses based on the MMR and the separate regression approaches in their Tables 2 and 3. For the reader's convenience, we include excerpts from those two tables in our Table 1, along with additional information relevant to the argument and analyses we report below.

Analytical Derivation of the Standard Error of $b_2 - b_1$ in Separate Regressions

The regression slopes b_1 and b_2 in Equations 2 and Equations 3 are *random variables*. They can each be thought of as values from random draws from their respective sampling distributions and will deviate from their "true" or "population" values in part by the luck of the draw, i.e. the same factors that influence sampling error more generally. Statistical theory informs us a lot about what the distributions of random variables such as regression coefficients look like over repeated sampling from a population. Under the standard assumptions of regression, the sampling distributions of b_1 and b_2 are roughly normal and are centered at β_1 and β_2—their corresponding population values. That is, assuming unbiased sampling, $E(b_1) = \beta_1$ and $E(b_2) = \beta_2$, and so b_1 and b_2 are typically used as point estimators of β_1 and β_2. The variances of these sampling distributions, which we denote $V(b_1)$ and $V(b_2)$, quantify how much on average b_1 and b_2 tend to differ from β_1 and β_2 in a sample of a given size. The square of the standard errors of b_1 and b_2 produced by OLS regression, $\hat{V}(b_1)$ and $\hat{V}(b_2)$, are typically used as point estimators of sampling variances $V(b_1)$ and $V(b_2)$.

As b_1 and b_2 are random variables, so too is their difference a random variable. In any study, $b_2 - b_1$ can be thought of as a random draw from the sampling distribution of $b_2 - b_1$, which is centered at $\beta_2 - \beta_1$ assuming unbiased sampling. To make an inference about the difference between β_1 and β_2, we need a point estimate, $b_2 - b_1$, and an estimate of the standard error—the square root of the sampling variance of $b_2 - b_1$. From this we can construct a confidence interval for $\beta_2 - \beta_1$ or conduct a hypothesis test. Robinson et al. recommend Equation 5 as the standard error estimator, whereas the MMR strategy uses Equation 4.

Unfortunately, the estimator of the standard error of $b_2 - b_1$ that Robinson et al. (2013) advocate does not follow analytically from rules of covariance algebra. Put bluntly, it is incorrect. The variance of the difference between two random variables is the sum of their variances minus twice their covariance (Lindgren, 1968, p. 126). In this case, the two random variables are b_1 and b_2, so

$$V(b_2 - b_1) = V(b_1) + V(b_2) - 2\mathrm{COV}(b_1, b_2)$$

But if b_1 and b_2 are estimates derived from two independent samples, then the covariance between b_1 and b_2 is zero, yielding

$$V(b_2 - b_1) = V(b_1) + V(b_2)$$

Replacing the unknown variances with point estimators of those variances yields

$$\hat{V}(b_2 - b_1) = \hat{V}(b_1) + \hat{V}(b_2) \tag{6}$$

as a point estimator of the variance of $b_2 - b_1$. The square root of Equation 6 serves as an estimator of the standard error of $b_2 - b_1$ (see Brame et al., 1998; Clogg, Petkova, & Haritou, 1995).

This analytical derivation shows that the proper standard error for the sampling distribution of $b_2 - b_1$ is the square root of the sum of the squared standard errors of b_1 and b_2:

$$se_{b_2 - b_1} = \sqrt{\hat{V}(b_1) + \hat{V}(b_2)} \tag{7}$$

There is no need to weight each of the squared standard errors by the relative sizes of the two groups as Robinson et al. (2013) recommend. Doing so produces an estimate of the standard error that is different from the one yielded by Equation 7, and is generally smaller. Holding $\hat{V}(b_1)$ and $\hat{V}(b_2)$ constant, the difference between Equation 5 and Equation 7 is a function of the relative sizes of n_1 and n_2. In the special case where the group sample sizes are equal (i.e., $n_1 = n_2$), the right-hand side of Equation 5 can be expressed as

$$\sqrt{\frac{n_1}{2n_1 - 2}}\sqrt{\hat{V}(b_1) + \hat{V}(b_2)}$$

As n_1 (and therefore n_2) grows, $\sqrt{n_1/(2n_1 - 2)}$ rapidly approaches but is never smaller than 0.707. So when sample sizes are equal, Robinson et al.'s standard error estimator (Equation 5) is about seven tenths of the size of the analytically-derived standard error estimator (Equation 7). That is, it is about 30% smaller than the standard error derived from covariance algebra.

It is noteworthy that in both examples that Robinson et al. (2013) report, the sample sizes of the two groups are about equal, and the standard error they report based on Equation 5 is indeed just about seven tenths of the standard error generated by Equation 7, as expected from the derivation above. In their first

Table 1. *Model Coefficients, Standard errors, and the Standard Error of the Slope Difference in the Two Examples Presented in Robinson et al. (2013)*

Self-Efficacy and Academic Achievement Example	b_1 (*SE*)	b_2 (*SE*)	b_3 (*SE*)
Moderated Multiple Regression (Eq. 1)	0.031 (0.006)	-0.098 (0.098)	-0.013 (0.0093)
Separate Regressions (Eqs. 2 and 3)	0.031 (0.0065) $n_1 = 104$	0.018 (0.0067) $n_1 = 105$	
			SE (slope difference)
MMR (eq. 4)			0.0093
Separate Regressions (Eq. 5)			0.0066
Separate Regressions (Eq. 7)			0.0093
Risk and Cancer Screening Example	b_1 (*SE*)	b_2 (*SE*)	b_3 (*SE*)
Moderated Multiple Regression (Eq. 1)	0.908 (0.149)	-2.317 (0.329)	-1.860 (0.217)
Separate Regressions (Eqs. 2 and 3)	0.908 (.141) $n_1 = 96^{\dagger}$	-0.951 (0.167) $n_2 = 91$	
			SE (slope difference)
Moderated Multiple Regression (Eq. 4)			0.217
Separate Regressions (Eq. 5)			0.155
Separate Regressions (Eq. 7)			0.219
Moderated Multiple Regression (HC3)			0.191

Note. Robinson et al. (2013) misreport this sample size as 95 in their Table 3.
SE = standard error. The standard error for the slope differences is the standard error of $b_2 - b_1$ for the separate regressions approach and the standard error of b_3 for the MMR approach. Though not reported in the paper, Robinson et al. (2013) mean centered risk (X) in the cancer screening example, so we have done so here as well as in the analysis generating the output in the Appendix.

example, Equation 5 yields 0.0066 compared to 0.0093 by Equation 7, a ratio of 0.701; in the second example Equation 5 yields 0.155 and Equation 7 results in 0.219, a ratio of 0.708 (see Table 1).

But even more telling with respect to the wisdom of Robinson et al.'s (2013) recommendation that the MMR approach not be used, the analytically-derived standard error expressed by Equation 7 is very similar to the standard error of b_3 using from the MMR approach, as can be seen in Table 1. That is, the standard error of b_3 in the first example using Equation 4 is 0.0093 and 0.217 in the second example, compared to the analytically derived standard errors of 0.0093 and 0.219, respectively. In other words, the MMR approach and the analytically derived standard error estimator largely agree each other using the results from the two examples Robinson et al. (2013) use, and they both differ from the standard error that Robinson et al. advocate by a predictable amount. This is not to say that the MMR is necessarily a good standard error estimator in all circumstances, as will be discussed later. But it is generally better than the estimator Robinson et al. (2013) advocate.

Simulation Evidence of the Bias in Equation 5

The analytical derivation above suggests that the standard error estimator Robinson et al. (2013) recommend in Equation 5 is negatively biased. Furthermore, for the two examples Robinson et al. report, the MMR estimator of the standard error of b_3 in Equation 4 provides a closer approximation of the analytical derivation of the standard error of the difference between b_2 and b_1. A small set of Monte Carlo simulations we report here confirms the negative bias in the estimator Robinson et al. (2013) advocate, and an additional published simulation we discuss later also confirms the bias.

One way of checking on a theoretically-derived sampling variance of an estimator is to simulate random draws from the sampling distribution of an estimator and compute the variance of the estimates

over repeated draws. By controlling the parameters of the population from which samples are drawn, it is possible to compare what is observed empirically to what an analytical derivation predicts.

To do so, we used the sample results from the Aiken (cancer risk and cancer screening) example Robinson et al. (2013) report in their Table 3 (see the second set of results in our Table 1) as population values and simulated the sampling of two regression slopes from populations defined by their results. In this example, the separate regressions approach yields b_1 = 0.908, $\sqrt{\hat{V}(b_1)}$ = 0.141, i_1 = 7.127, n_1 = 96, and b_2 = -0.951, $\sqrt{\hat{V}(b_2)}$ = 0.167, i_2 = 4.810, n_2 = 91. Thus, we define the parameters of the simulation such that β_1 = 0.908, β_2 = -0.951, $\sqrt{V(b_1)}$ = 0.141, and $\sqrt{V(b_2)}$ = 0.167. All simulations were conducted using the GAUSS statistical system, version 12 (Aptech Systems, 2011).

For a simulated low risk group, we constructed n_1 = 91 values of Y using the function

$$Y = 7.127 + \beta_1 X + e_1$$

where X was a random standard normal variable and values of e_1 were drawn from a population of random normal deviates centered as zero with standard deviation = 1.37. This use of 1.37 ensures that the standard deviation of the sampling distribution of b_1 is about 0.141. An identical approach was used to construct n_2 = 96 values of X and Y for a simulated high risk group. The function to construct Y from X was

$$Y = 4.810 + \beta_2 X + e_2$$

where X was a random standard normal variable and e_2 is drawn from a population with random normal errors centered at zero with standard deviation 1.57, yielding a standard deviation of the sampling distribution of b_2 of about 0.167.

With values of X and Y generated in each of the two groups, we regressed Y on X using OLS regression in each of the two groups separately to generate estimates of b_1 and b_2, and then calculated their difference $b_2 - b_1$. This was repeated a total of 100,000 times. The result is a sample of 100,000 estimated differences between the two slopes when sampling from two populations with values of β_1, $V(\beta_2)$, β_2, and $V(\beta_2)$ corresponding to the estimates from the risk and cancer screening example Robinson et al. (2013) use.

In each of these 100,000 samples we also estimated the standard error of the difference using Equations 5 (Robinson et al.'s approach) and 7 (the analytically derived standard error). We also implemented the MMR strategy, which estimates the slope difference as b_3 in Equation 1, with standard error estimated using Equation 4. In addition, we implemented a *t*-test of the null hypothesis that the slopes are equal, as well as constructed a 95% confidence interval for the difference, recording whether or not the confidence interval contained the population slope difference of -1.86.

The results of the simulation comparing these methods, along with another we discuss later, can be found in the first row of Table 2. Naturally, the mean of the 100,000 estimated slope differences was -1.86—the population difference. More important is the standard deviation of these 100,000 estimates of the slope difference—listed as the *Empirical SE* in Table 2—which was 0.219. It is this standard deviation that the various standard error estimators described above attempt to estimate. The empirical standard error is the same to the third decimal place as the estimated standard error of 0.219 generated when Equation 7—the analytically derived standard error—is applied to the sample results reported by Robinson et al. and in our Table 1. It also is very close to the standard error for b_3 using the MMR strategy, but it is quite different from and considerably larger than the standard error generated by Equation 5, which is the approach Robinson et al. recommend.

Table 2 also provides the mean estimated standard error of the slope difference when these three different estimators are applied to each of the 100,000 simulated data sets. As can be seen, on average, the standard error constructed using Equations 4 (MMR) and 7 (analytical slope difference) is very close to the standard deviation of the 100,000 slope differences. But the average standard error constructed using Equation 5 (Robinson et al.'s favored method) is substantially smaller, at 0.155.

These results illustrate the bias in the standard error introduced by weighting the standard errors of the within-group slopes by the group samples sizes, as Equation 5 does. On the surface, this underestimation of the standard error would seem to have little consequence for testing the null hypothesis of equality of the slopes in this example, as in every case the use of Equation 5 correctly rejects the null. However, power is so high in this example that all methods correctly do so; the rejection rate is 100% for every method. More telling is the coverage of the confidence intervals constructed for the

slope difference using different standard error estimators. As can be seen using the MMR standard error estimator (Equation 4) or the analytically derived standard error of the separate regressions slope difference (Equation 7), about 95% of the 95% confidence intervals cover the true slope difference of -1.86. That is, only about 5% of the confidence intervals fail to include the true difference, just as a properly constructed 95% interval estimate should. But only 84% of the confidence intervals constructed using Robinson et al.'s standard error in equation 5 contain the true slope difference. This is as would be expected for a standard error estimator that is negatively biased. It produces confidence intervals that are too narrow and, as a result, miss the population slope difference more than 5% of the time.

To get at Type I error, we did another simulation identical to the one just described but setting $\beta_1 = \beta_2$. In this simulation, we split the difference between β_1 and β_2 in the prior simulation and used the group sample size-weighted average of 0.003 for each. The results of this run can be found in the second row of Table 2. As can be seen, again, the average of 100,000 standard errors estimated using Equations 4 and 7 were very close to the standard deviation of the 100,000 estimated slope differences (which is still 0.220 because the standard errors are determined by sampling variance and not by the population slopes). But the average standard error using Equation 5 was too small. The result is Type I error inflation, as can be seen in Table 2. Using Equation 5 resulted in rejection of the true null hypothesis 16.4% of the time at the $\alpha = 0.05$ level of significance, as expected when using a standard error that is negatively biased. But the MMR standard error and the analytically-derived standard error rejected the true null hypothesis about 5% of the time, as they should. Confidence interval coverage reflects this, with proper coverage using Equations 4 and 7, but coverage that is well below 95% when using Equation 5. These results stand in direct conflict with Robinson et al.'s (2013) claim in various places (pp. 16, 17, 23, 24) that the use of Equation 5 does not adversely affect the Type I error rate relative to Equation 4.

Table 2. *Monte Carlo Simulation Results Examining the Performance of Four Estimators of the Difference Between Independent Regression Coefficients*

	Parameters		Standard Error Estimator				Empirical
			Eq. 5	Eq. 7	MMR Eq. 4	MMR HC3	*SE*
1	$\beta_1 = 0.908$, $\beta_2 = -0.951$	Mean *SE*	0.155	0.219	0.218	0.223	0.219
	$n_1 = 96$, $n_2 = 91$	Rej.%	100.0	100.0	100.0	100.0	
	$SDe_1 = 1.37$, $SDe_2 = 1.57$	Cov.%	83.7	95.1	95.0	95.3	
2	$\beta_1 = 0.003$, $\beta_2 = 0.003$	Mean *SE*	0.155	0.219	0.218	0.223	0.220
	$n_1 = 96$, $n_2 = 91$	Rej.%	16.4	5.1	5.2	4.9	
	$SDe_1 = 1.37$, $SDe_2 = 1.57$	Cov.%	83.6	94.9	94.8	95.1	
3	$\beta_1 = -0.200$, $\beta_2 = 0.200$	Mean *SE*	0.102	0.167	0.167	0.171	0.168
	$n_1 = 150$, $n_2 = 50$	Rej.%	88.2	66.7	66.8	64.6	
	$SDe_1 = 1.00$, $SDe_2 = 1.00$	Cov.%	77.2	94.9	95.0	95.0	
4	$\beta_1 = 0.400$, $\beta_2 = 0.400$	Mean *SE*	0.296	0.481	0.482	0.511	0.488
	$n_1 = 25$, $n_2 = 75$	Rej.%	22.7	5.4	5.0	5.1	
	$SDe_1 = 2.00$, $SDe_2 = 2.00$	Cov.%	77.3	94.6	95.0	94.9	
5	$\beta_1 = 0.200$, $\beta_2 = 0.200$	Mean *SE*	0.162	0.300	0.220	0.310	0.304
	$n_1 = 50$, $n_2 = 150$	Rej.%	29.2	5.4	15.2	5.4	
	$SDe_1 = 2.00$, $SDe_2 = 1.00$	Cov %	70.8	94.6	84.8	94.6	
6	$\beta_1 = -0.300$, $\beta_2 = -0.300$	Mean *SE*	0.161	0.219	0.301	0.225	0.220
	$n_1 = 50$, $n_2 = 150$	Rej.%	14.9	5.0	0.8	4.8	
	$SDe_1 = 1.00$, $SDe_2 = 2.00$	Cov.%	85.1	95.0	99.2	95.2	

Note. 100,000 replications; Mean *SE* = Mean standard error; Rej.% = Percentage of rejections of test of the null hypothesis that $\beta_1 = \beta_2$ using *t* distribution; *SDe* = standard deviation of the errors in estimation of *Y* from *X*; Cov.% = 95% confidence interval coverage; Empirical *SE* is the standard deviation of the 100,000 estimated slope differences.

More Simulation Evidence

In the event the reader questions whether the results of the simulations presented above are specific to the parameters we used, we provide the results from a few other simulations in Table 2 in which we varied the sample sizes of the two groups, error variances, and β_1 and β_2. The results were the same, with Equation 5 resulting in underestimation of the standard error of the slope difference, under-coverage of 95% confidence intervals, and elevated Type I error when the population slopes were equal. These problems were not in evidence using Equation 7, or when using Equation 4, with the exception of the last two sets of conditions (rows 5 and 6 of Table 2) which we discuss later.

Furthermore, we are not the only ones to have studied the relative performance of these approaches to comparing independent regression weights. Brame et al. (1998) did an extensive set of Monte Carlo simulations comparing minor variants of these approaches and their results are comparable to those we report here, with Type I error rates three to seven times the nominal significance level.[4] They conclude, in reference to their version of Equation 5, that "We found that the bias produced by this estimator is, in general, nontrivial. Consequently, we believe that researchers should abandon any use of [equation 5]" (p. 258). We concur, as presumably does Cohen (1983), who shows that Equation 5 is not the proper estimator of the standard error of $b_2 - b_1$.

Where Does Equation 5 Come From?

Robinson et al. (2013) do not provide any source justifying Equation 5 as their preferred estimator of the standard error of the difference between slopes. However, toward the end the paper they state "Our equations were verified using Kleinbaum and Kupper's 1978 textbook" (p. 23). We examined this book and found two versions of the standard error estimator. On page 101, section 8.3.2, they offer a "large sample Z test for parallelism" that is equivalent to Equation 7—the standard error estimator we derived analytically. In the prior section (8.3.1) they offer a "small-sample *t* test for parallelism" with a standard error estimator that is not equivalent to Equation 5 and is more similar to Equation 4. It yields 0.212 when applied to the cancer risk example data, which is trivially different than what Equations 4 and 7 yield in this example and very different from the 0.155 that Robinson et al. report in their Table 3 using Equation 5. This leads us to wonder whether Kleinbaum and Kupper (1978) is the source of Robinson et al.'s preferred standard error estimator. Regardless, later, Kleinbaum and Kupper (p. 192) point out when describing the MMR strategy that it yields a test of equality of regression slopes exactly equivalent to the small sample *t* test they describe in section 8.3.1.

Yet Robinson et al. (2013) are not the only ones to have advanced Equation 5 as the standard error of the difference between slopes. Others who have compared the relative performance of Equations 5 and 7 have traced Equation 5 back to an article by Wright in a 1978 volume of the *American Journal of Sociology* (Brame et al., 1998). We have not had any luck finding any source earlier than this, nor have we found any modern references advocating its use in any specific field. But it has been used in other disciplines (see, for example, the list provided by Brame et al.), so we can assume it has been used in education research and other fields as well.

Regardless of its ultimate origin, there is a certain allure to Equation 5 that might lead people to unquestionably accept it as a legitimate standard error estimator for the difference between slopes. The separate regressions approach to comparing regression slopes requires a single standard error of the difference between slopes, but two standard errors are available for use, one for b_1 and one for b_2. Clearly, some kind of "pooling" of the information from the two regressions is needed. This is similar to the problem faced by researchers needing to compare the means of two independent groups. Most every introductory statistics book (e.g., Agresti & Findlay, 2009) offers

$$se_{\bar{Y}_2-\bar{Y}_1} = \sqrt{\frac{\hat{V}(Y_1)}{n_1} + \frac{\hat{V}(Y_2)}{n_2}} \tag{8}$$

as the estimated standard error of $\bar{Y}_2 - \bar{Y}_1$, where $\hat{V}(Y_1)$ is the estimated variance of the measurements of *Y* in group 1 and $\hat{V}(Y_2)$ is the estimated variance of the measurements of *Y* in group 2. Under the assumption of between-group equality of variances in *Y*, a "pooled" standard deviation is calculated which is a weighted average of variances of *Y* in the two groups

$$s_p = \sqrt{\frac{(n_1 - 1)\hat{V}(Y_1) + (n_2 - 1)\hat{V}(Y_2)}{n_1 + n_2 - 2}} \quad (9)$$

the square of which is then substituted into Equation 8 for both $\hat{V}(Y_1)$ and $\hat{V}(Y_2)$.

Notice that Equation 9 is quite similar to the right hand side of Equation 5, with two critical differences other than the subtraction of 1 from each of the sample sizes in the numerator. First, $\hat{V}(Y_1)$ and $\hat{V}(Y_2)$ are estimates of the variances of *measurements*, not estimates of the *sampling variances of an estimator*. Second, Equation 9 is only an intermediate computation to the generation of the standard error of the mean difference. The result of Equation 9 must be substituted into Equation 8 for $\hat{V}(Y_1)$ and $\hat{V}(Y_2)$ which then yields a standard error for $\bar{Y}_2 - \bar{Y}_1$. Indeed, $\hat{V}(Y_1)/n_1$ and $\hat{V}(Y_2)/n_2$ in Equation 8 are the sampling variances of estimators (of $\bar{Y}_1$ and $\bar{Y}_2$), much like $\hat{V}(b_1)$ and $\hat{V}(b_2)$ are sampling variances of estimators in Equation 7. There is no weighting of these sampling variances in Equation 8 by sample size as in Equation 5. Thus, the analytically derived standard error for $b_2 - b_1$ in Equation 7, shown to be the more accurate estimator, overlaps theoretically and computationally more with Equation 8 than with Equation 9.

Heteroscedasticity and the Standard Error of the Slope Difference

The separate regressions approach and the MMR approach to comparing independent regression slopes differ in one important way. The two separate regressions estimators of the standard error of the difference between slopes (Equations 5 and 7) do not assume between-group equality of the variance of the errors in estimation of *Y* from *X*, unlike the MMR estimator in Equation 4. For those who prefer to avoid unnecessary assumptions, Equation 7 seems like a better choice than Equation 4, even though Equation 7 requires some extra hand computation. Yet in our limited simulation, the two methods performed almost identically.

But the results we have described thus far belie a more complex reality, as there are boundary conditions to the similarity between the standard errors generated by Equations 4 and 7. Research has shown that when the variances in the errors of estimation differ between groups and the sample sizes of the groups differ, Equation 4 yields a standard error for b_3 that is either too small or too large (Aguinis & Pierce, 1998; Dretzke, Levin, and Serlin, 1982). The result is a test of equality of the regression slopes that is either too liberal (inflated Type I error rate) or too conservative (deflated Type I error rate and reduced power). So the convenience of having a standard error readily available in regression output when using the MMR approach is tempered in part by the conditions that must be placed on its use. If no alternatives were available, one recommendation would be to use the MMR strategy when the assumption of between group equality of variance of errors is met, otherwise use Equation 7, which must be calculated by hand.

But there are alternatives available, so a compromise is possible. There is a family of *heteroscedasticity-consistent* standard error estimators that have been widely studied and perform well in regression analysis in the presence of heteroscedasticity, and regardless of the form of that heteroscedasticity. Moreover, these estimators work quite well even when the homoscedasticity assumption is met. Heteroscedasticity-consistent estimators are somewhat complex and cannot be represented without the use of matrix notation; we refer the reader to Long and Erwin (2000) or Hayes and Cai (2007) for the details. Conveniently, several of these estimators have been implemented in STATA as well as regression analysis macros for SPSS and SAS (Hayes & Cai). One of these that seems to perform particularly well—the HC3 estimator attributed to MacKinnon and White (1985)—is implemented in a freely-available and easy-to-use SPSS and SAS macro for moderation and mediation analysis (PROCESS) described and documented in Hayes (2013). When the HC3 estimator was used instead of Equation 4 in the cancer risk analysis, the resulting standard error for b_3 was 0.191 (see Table 1), which is slightly smaller than yielded by Equations 4 and 7 but larger than Equation 5. An example output from the PROCESS macro for SPSS and SAS used to conduct this analysis can be found in the Appendix.

In the simulations described earlier, we also implemented hypothesis tests and confidence intervals for each set of conditions using the HC3 estimator of the standard error of b_3 in the moderated multiple regression approach. In the first two sets of conditions (rows 1 and 2 of Table 2), the equality of variance assumption is violated (as the population error variances were different between the two groups by design) but only slightly so, whereas in the second two sets (rows 3 and 4 of Table 2) the assumption is met. In all four of these cases, the HC3 estimator using the MMR approach yielded an accurate standard error, Type I error rates were about 5%, and coverage for 95% confidence intervals was about 95%, as they should be.

The last two rows of Table 2 are most important, however, as they represent more substantial differences in estimation error variance than in the conditions delineated in the first two rows. In the conditions described in row 5, $\beta_1 = \beta_2$ but the variance of the errors in estimation is 4 times larger in the smaller group. Observe that the MMR standard error estimator for b_3 defined in Equation 4 is now quite liberal, though not as liberal as Robinson et al.'s preferred estimator, with a standard error that is on average too small, a Type I error rate is larger than 5%, and confidence interval coverage below 95%. But the analytically-derived standard error estimator in Equation 7 for the separate regressions approach did quite well, as did the MMR approach using the HC3 standard error estimator. Both of these resulted in average standard errors roughly equal to the empirical standard error, and Type I error rate and confidence interval coverage were right on at 5% and 95%, respectively.

The last row in Table 2 describes the results in a set of conditions in which $\beta_1 = \beta_2$ but the variance of errors in estimation of Y from X is now four times larger in the *larger* group. Now Equation 4 overestimates the standard error of b_3, yielding a highly conservative test with a Type I error rate well below 5% and correspondingly high confidence interval coverage. The analytically-derived standard error estimator in Equation 7 gets the standard error of the slope difference correct, with appropriate Type I error rates and coverage, as does the HC3 estimator of the standard error of b_3 using the MMR approach. The standard error estimator Robinson et al. (2013) advocate still produces a standard error that is too small relative to the empirical standard error, with corresponding negative effects on Type I error rate and confidence interval coverage.

Summary

In this paper we offer a rebuttal to Robinson et al.'s (2013) claim that separate regressions combined with the standard error estimator they recommend for comparing two independent regression coefficients should be routinely used instead of moderated multiple regression. We have shown that the standard error estimator they advocate does not follow from covariance algebra, it is downwardly biased, and results in an inflated Type I error rate and confidence intervals that do not provide adequate coverage of the population difference. Although the separate regressions approach allows one to relax the assumption of between-group equality of errors in estimation inherent in the MMR approach, Equation 7 is the proper estimator of the standard error of the difference between slopes, not Equation 5. When this assumption is met or nearly so, the moderated multiple regression strategy performs as well as Equation 7 when doing separate regressions. The disadvantage of Equation 7 is that it is not implemented in existing software and thus must be computed by hand. A compromise is the use of the HC3 estimator of the standard error of b_3 in moderated multiple regression. It does not require the assumption of between-group equality of errors in estimation, it works as well as Equation 7, and it is available in existing and widely-used software either with (in SPSS) or without (in STATA) a special macro.

It follows from the analytical derivation we outline, our simulation results, and simulations already published that Robinson et al.'s explanations for the superior performance of Equation 5 relative to Equation 4 in the latter half of their paper are neither accurate nor germane. The smaller standard errors they observe in their two examples after applying Equation 5 rather than Equation 4 is the result of using a negatively biased standard error estimator, and nothing more. Furthermore, there is no need to review published or unpublished studies, as they suggest may be necessary, in the hopes of finding all the false negatives that have resulted from the use of the moderated multiple regression strategy—effects Robinson et al. (2013) worry may have been detected had an researchers instead used the separate regressions approach and standard error they advocate.

References

Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences* (4th ed.). Upper Saddle River, NJ: Prentice-Hall.

Aguinis, H., & Pierce, C. A. (1998). Heterogeneity of error variance and the assessment of moderating effects of categorical variables: A conceptual review. *Organizational Research Methods, 1*, 296-314.

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82*, 192-206.

Aptech Systems (2011). GAUSS version 12 [Computer software]. Black Diamond: WA: Aptech Systems.

Aspelmeier, J. E., Love, M. M., McGill, L. A., Elliot, A. N., & Pierce, T. W. (2012). Self-esteem, locus of control, college adjustment, and GPA among first- and continuing-generation students: A moderator model of generational status. *Research in Higher Education, 53*, 755-781.

Brame, R., Paternoster, R., Mazerolle, P., & Piquero, A. (1998). Testing the equality of maximum-likelihood regression coefficients between two independent equations. *Journal of Quantitative Criminology, 14*, 245-261.

Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology, 100*, 1261-1293.

Cohen, A. (1983). Comparing regression coefficients across subsamples: A study of the statistical test. *Sociological Methods & Research, 12*, 77-94.

Cohen, J. (1978). Partialed products *are* interactions; Partialed powers *are* curve components. *Psychological Bulletin, 85*, 858-866.

Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.

Dretzke, B. J., Levin, J. R., & Serlin, R. C. (1982). Testing for regression homogeneity under variance heterogeneity. *Psychological Bulletin, 91*, 376-383.

Farris, J., Lefever, J. E. B., Borkowski, J. G., &Whitman, T. L. (2013). Two are better than one: The joint influence of maternal preparedness for parenting and children's self-esteem on academic achievement and adjustment. *Early Education and Development, 24*, 346-365.

Friedrich, R. J. (1982). In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science, 26*, 797-833.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* New York: The Guilford Press.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods, 39*, 709-722.

Hayes, A.F., Glynn, C.J., & Huge, M.E. (2012). Cautions regarding the interpretation of regression coefficients and hypothesis tests in models with interactions. *Communication Methods & Measures, 6*, 1-11.

Kleinbaum, D. G., & Kupper, L. L. (1978). *Applied regression analysis and other multivariate methods*. North Scituate, MA: Duxbury Press.

Kromrey, J. D., & Foster-Johnson, L. (1998). Mean centering in moderated multiple regression: Much ado about nothing. *Educational and Psychological Measurement, 58*, 42-67.

Lindgren, B. W. (1968). *Statistical theory* (2nd ed.). New York: The MacMillan Company.

Loes, C. N., Salisbury, M.H., & Pascarella, E. T. (2013). Diversity experiences and attitudes toward literacy: Is there a link? *The Journal of Higher Education, 84*, 834-865.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity-consistent standard errors in the linear regression model. *American Statistician, 54*, 217-224.

MacKinnon, J. G., & White, H. (1985). Some heteroscedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics, 29*, 305-325.

Paternoster, R., Brame, R., Mazerolle, P., Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology, 36*, 859-866.

Robinson, C. D., Tomek, S., & Schumacker, R. E. (2013). Tests of moderation effects: Difference in simple slopes versus the interaction term. *Multiple Linear Regression Viewpoints, 39*, 16-24.

Roksa, J., & Potter, D. (2011). Parenting and academic achievement: Intergenerational transmission of educational advantage. *Sociology of Education, 84*, 299-321.

Send correspondence to: Andrew F. Hayes
The Ohio State University
Email: hayes.338@osu.edu

APPENDIX

Example Output from the PROCESS Macro for the Cancer Risk Analysis using the HC3 Estimator

SPSS: PROCESS vars=dummy risk int2gr/y=int2gr/x=risk/m=dummy/model=1/hc3=1.

SAS: %PROCESS (data=aiken,vars=dummy risk int2gr,y=int2gr,x=risk,m=dummy,model=1,hc3=1);

```
***************** PROCESS Procedure for SPSS Release 2.12 ***************
          Written by Andrew F. Hayes, Ph.D.       www.afhayes.com
    Documentation available in Hayes (2013). www.guilford.com/p/hayes3
**************************************************************************
Model = 1
    Y = int2gr
    X = risk
    M = dummy
Sample size
       187
**************************************************************************
Outcome: int2gr

Model Summary
          R       R-sq          F        df1        df2          p
      .7027      .4937    81.0503     3.0000   183.0000      .0000
Model
              coeff         se          t          p       LLCI       ULCI
constant     7.1270      .1841    38.7129      .0000     6.7638     7.4902
dummy       -2.3168      .3174    -7.3000      .0000    -2.9429    -1.6906
risk          .9081      .1247     7.2818      .0000     .6620     1.1541
int_1       -1.8595      .1909    -9.7396      .0000    -2.2362    -1.4828

Interactions:
 int_1    risk        X     dummy
**************************************************************************
Conditional effect of X on Y at values of the moderator(s):
      dummy     Effect         se          t          p       LLCI       ULCI
      .0000      .9081      .1247     7.2818      .0000      .6620     1.1541
     1.0000     -.9514      .1446    -6.5812      .0000    -1.2367     -.6662

Values for quantitative moderators are the mean and plus/minus one SD from mean.
Values for dichotomous moderators are the two values of the moderator.

******************** ANALYSIS NOTES AND WARNINGS ************************

Level of confidence for all confidence intervals in output:
    95.00
NOTE: All standard errors for continuous outcome models are based on the
      HC3 estimator
```

Endnotes

1. This test is mathematically identical to the test as to whether XZ explains any variation in Y independent of X and Z. This is frequently tested by examining the change in R2 that results when XZ is added to a model containing X and Z. Under the null hypothesis that the two regression coefficients are equal, the increase in R2 due to XZ can be converted to a statistic distributed as F(1,n – 4). The F(1,n – 4) distribution is the square of the t(n – 4) distribution. The p-values for these tests will be the same.

2. Equation 5 here is equation 8 in Robinson et al. (2013).

3. It is not apparent from the data for this example, available at http://www.public.asu.edu/~atlsa/PSY531/ whether the dichotomous measure of risk is derived from a measure of risk different than the continuous measure. We assume it is, but whether it is or not does not affect any of our computations or the argument itself.

4. The version of equation 5 that Brame et al. (1998) studied weights the squared standard errors of the separately-estimated regression coefficients by their respective degrees of freedom rather than by the sample size. Furthermore, they used maximum likelihood estimation of the parameter estimates and standard errors rather than OLS. But this difference is trivial and would not have any noteworthy effects on the results of the simulation.