Multiple Linear Regression Viewpoints

A Publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression: The General Linear Model

MLRV

Volume 33 • Number 1 • Fall 2007

Table of Contents

Multiple Linear Regression Viewpoints

Editorial Board

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the AERA Special Interest Group on Multiple Linear Regression: General Linear Model through the **University of Alabama at Birmingham**.

Subscription and SIG membership information can be obtained from:
Jeffrey B. Hecht, MLR:GLM/SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854.
jbhecht@niu.edu

MLRV abstracts appear in CIJE, the ERIC system, and microform copies are available from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106. *MLRV* is listed in the *EBSCO Librarians Handbook*.

The Use and Impact of Adjusted R^2 Effects in Published Regression Research

Lesley F. Leach **Robin K. Henson**
University of North Texas

This paper empirically evaluates the reporting of adjusted effect sizes (e.g., adjusted R^2, omega^2) in published multiple regression studies by (a) documenting the frequency of adjusted effect reporting and interpretation, (b) identifying the types of corrected effects reported, and (c) estimating the degree of "shrinkage" present across regression analyses by using the information found in published journal articles to calculate corrected effects based on various formulae. Adjusted effects were infrequently reported in the literature, and interpretation of adjusted effects that were reported was rare.

Researchers are becoming increasingly aware that interpretation of effect sizes is critical in evaluating empirical results (Henson & Smith, 2000; Henson, 2006; Kirk, 1996; Rosnow & Rosenthal, 1989; Thompson, 1996; Thompson & Snyder, 1997). The American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson & APA Task Force on Statistical Inference, 1999) stated:

> It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual *p*-value or, better still, a confidence interval. . . *Always* provide some effect-size estimate when reporting a *p*-value. (p. 599, italics added).

The Task Force went on to state, "Always present effect sizes for primary outcomes . . .It helps to add brief comments that place these effect sizes in a practical and theoretical context" (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599).

This directive was a substantial step beyond the fourth edition of the APA's Publication Manual, which only *recommended* reporting of effect sizes in research (APA, 1994). Several empirical studies demonstrated, however, that this recommendation had little impact on the number of effect sizes reported in articles and it affected the interpretation of effect sizes even less (cf. Henson & Smith, 2000; Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000).

The fifth edition of the APA manual (APA, 2001) incorporated the Task Force's directive, stating "For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section" (p. 25). The current APA manual also called the "failure to report effect sizes" a "defect in the design and reporting of research" (p. 5). At least 23 journals have followed suit, requiring the inclusion of effect sizes with statistical results (Onwuegbuzie, Levin, & Leech, 2003).

The use of effect sizes has been widely discussed in the literature vis-à-vis null hypothesis significance tests (NHST). A discussion of issues surrounding the use of NHSTs is beyond the scope of this paper. Harlow, Mulaik, and Steiger (1997) present a balanced discussion of the debate for interested readers. Huberty and Pike (1999) and Huberty (2002) document the historical development of statistical testing and effect sizes, respectively.

Indeed, Pedhazur and Schmelkin (1991) noted that, "Probably few methodological issues have generated as much controversy among sociobehavioral scientists as the use of [statistical significance] tests" (p. 198). Elsewhere, Pedhazur (1997) indicated that the "controversy is due, in part, to various misconceptions of the role and meaning of such [statistical significance] tests in the context of scientific inquiry" (p. 26). These "misconceptions" have been attacked for considerable time (see e.g., Berkson, 1942; Tyler, 1931), and yet they persist in modern research practice (Cohen, 1994; Finch, Cumming, & Thomason, 2001). Nevertheless, current methodological practice is increasingly emphasizing the need for effect size indices and more accurate interpretation of NHSTs (Kline, 2004).

Some researchers recommend using effect sizes and NHSTs together (Fan, 2001; Huberty, 1987). Moreover, some critics of NHSTs have argued that effect sizes should be reported whether or not the results are statistically significant (Rosnow & Rosenthal, 1989; Thompson, 1999). As Roberts and Henson (2002) stated, ". . .one remaining point of debate concerns whether effect sizes should be reported (a) for all null hypothesis tests, even non-statistically significant ones, or (b) only after a finding is first determined to be statistically significant" (pp. 242-243).

Types of Effect Sizes: Corrected and Uncorrected Indices

There are many different effect size indices from which researchers can choose, but most can be grouped into two broad categories: (a) measures of standardized differences (e.g., Cohen's *d*, Hedges' *g*) and (b) variance–accounted-for measures (e.g., R^2, η^2) (Kirk, 1996; Kline, 2004; Olejnik & Algina, 2000; Onwuegbuzie, Levin, & Leech, 2003). Reviews of various effect size indices are provided by Olejnik and Algina (2000), Snyder and Lawson (1993), and Yin and Fan (2001).

In addition, effect sizes can be further classified as "uncorrected" or "corrected" measures (Thompson, 2002). For example, R^2 is commonly used in multiple regression applications and is the most prevalent effect size index documented in the literature – most likely due to the fact that practically all statistical computer packages routinely provide R^2 as part of regression output (Kirk, 1996). Studies have shown, however, that R^2 systematically overestimates the proportion of explained variance to total variance expected in the population or future samples (Carter, 1979; Fan, 2001; Snyder & Lawson, 1993; Thompson, 1990, 1999; Yin & Fan, 2001). That is, general linear model analyses such as multiple regression commonly utilize the ordinary least squares (OLS) estimation method to obtain the greatest possible effect size. Analyses using this estimation method capitalize on *all* the variance in a sample, including the variance attributable to sampling error that is unlikely to be present in future samples or the population (Thompson & Kieffer, 2000). Because the effect size accounts for error unique to the sample data, the resulting "uncorrected" R^2 is often found to be a biased estimate of the variance explained in the population (Roberts & Henson, 2002; Yin & Fan, 2001) and future samples (Thompson, 1990).

To statistically remove the bias associated with sampling error, various adjustment formulae can be used to "shrink" the effect size by the theoretical amount of sampling error present in a given sample (Snyder & Lawson, 1993). The amount of shrinkage is determined using the factors that affect sampling error. Theoretically, sampling error increases (a) as sample size decreases, (b) as the number of variables in the model increases (and, by extension the number of predictors increase), and (c) as the population effect decreases (Thompson, 1999; Vacha-Haase & Thompson, 2004). Because adjustment formulae limit the influence of the factors that increase sampling error, these "corrected" effects provide a better estimate of the population squared multiple correlation coefficient (Carter, 1979; Larson, 1931; Pedhazur, 1997). But as this paper demonstrates, corrected effects are rarely reported, and the failure to report such corrected effects may impact result interpretation.

Purpose

Because corrected effects can be more accurate estimates of the effect in the population or future samples, the purposes of the present study were to (a) document the frequency of corrected effect reporting and interpretation, (b) identify the types of corrected effects that are reported, and (c) estimate the degree of shrinkage present when authors do not give corrected effects. Information found in the reviewed articles was used to calculate corrected effects based on various formulae (Snyder & Lawson, 1993; Yin & Fan, 2001). This analysis facilitated inspection of interpretation differences resulting from effect size adjustment and permitted empirical investigation of the amount of correction provided by the various corrected effect formulae. Because R^2 and adjusted R^2 are typically reported with regression results in statistical software packages, this paper addressed only multiple regression applications.

Adjusted R^2 Formulae

There are many formulae available for calculating corrected effect sizes. Table 1 outlines various formulae presented by Snyder and Lawson (1993) and Yin and Fan (2001), which shrink R^2 based on the number of predictors (k), sample size (n), and the obtained effect (R^2) as an initial estimate of the population effect. The adjustment formulae fall into two different categories based on their purposes: (a) population effect estimates and (b) future sample effect estimates. Population effect estimates approximate the association strength expected to be realized in the population (Yin & Fan, 2001), while those in the future sample category estimate the effect likely to be found upon replication of the study with a new sample (Snyder & Lawson, 1993). One could expect greater shrinkage to be more likely with future sample estimates because "[they] must adjust for sampling error present in both the present study and some future study" (Snyder & Lawson, 1993, p. 340). Conversely, adjusted effect estimates of the population parameter only adjust for the sampling error influencing the present study's data and, consequently, will generally be less conservative than estimates of the effect in future samples.

Table 1. Various Adjusted R^2 Formulae.

Population Effect Estimates		Future Sample Effect Estimates	
Index	Formula	Index	Formula
Smith	$1-\left(\frac{n}{n-k}\right)\left(1-R^2\right)$	Lord-1	$1-\frac{n+k+1}{n-k-1}\left(1-R^2\right)$
Ezekiel	$1-\left(\frac{n-1}{n-k-1}\right)\left(1-R^2\right)$	Lord-2	$1-\frac{(n+k+1)(n-1)}{(n-k-1)n}\left(1-R^2\right)$
Wherry-2	$1-\left(\frac{n-1}{n-k}\right)\left(1-R^2\right)$	Darlington-Stein	$1-\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)\left(1-R^2\right)$
Olkin-Pratt	$R^2-\frac{k-2}{n-k-1}\left(1-R^2\right)-\left(\frac{2(n-3)}{(n-k-1)(n-k+1)}\right)\left(1-R^2\right)^2$	Browne[a]	$\frac{(n-k-3)\rho^4+\rho^2}{(n-2k-2)\rho^2+\rho}$
Pratt	$1-\frac{(n-3)(1-R^2)}{(n-k-1)}\left[1+\frac{2(1-R^2)}{n-k-2.3}\right]$	Claudy-1[b]	$\left(2\rho-R\right)^2$
Claudy-3	$1-\frac{(n-4)(1-R^2)}{(n-k-1)}\left[1+\frac{2(1-R^2)}{n-k+1}\right]$	Claudy-2	$1-\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n-1}{n}\right)\left(1-R^2\right)$
		Rozeboom-1	$1-\left(\frac{n+k}{n-k}\right)\left(1-R^2\right)$
		Rozeboom-2[a]	$\rho^2\left[1+\left(\frac{k}{n-k-2}\right)\left(\frac{1-\rho^2}{\rho^2}\right)\right]^{-1}$

Note. *n*=sample size. *k*=number of predictor variables. Adapted from Yin & Fan (2001) and, Snyder & Lawson (1993). [a]$\rho 2$ was estimated with the Ezekiel value. [b]ρ was estimated with the square root of the Ezekiel value. Negative Ezekiel values were replaced with zeros.

Ultimately, the decision of which R^2 adjustment formula to use depends on the generalizations that the researcher wishes to make. As Snyder and Lawson (1993) observed, "Most researchers ground their work in empirical findings from previous samples and usually desire that their work generalize to future samples" (p. 341). Researchers seeking this goal would be wise to consider corrected effect size estimates for future samples. If, however, the researcher wishes to develop population expectations, a population effect estimate may be more appropriate (Snyder & Lawson, 1993; Yin & Fan, 2001). If replicability is indeed the hallmark of scientific inquiry, then the sample effect that best represents the effect expected in the population or future samples should be of primary analytic interest. Accordingly, we argue that these corrected effects should be both reported and interpreted whenever possible.

Method

We examined regression applications in four journals –*Journal of Applied Psychology* (v.86[5] – 87[4])*, Journal of Educational Psychology* (v. 93[4]-94[3])*, Journal of Experimental Education* (v. 95[1]-96[1])*, and Journal of Educational Research* (v. 70[2]-71[1])– over a one-year time span. We considered only the first three regression analyses presented in each article. Additional analyses were not considered so that articles containing an above-average number of regression analyses would not overly impact the results. The frequencies of uncorrected and corrected effects were coded as well as the interpretation of the effects. We considered an effect to be interpreted if the author included a statement explaining the effect in relation to the dependent variable. For example, Klein, Conn, and Sorra (2001) interpreted R^2 by noting, "Together management support and financial resource availability explained 19%. . .of the variance in implementation of polices and practices (β = .36, p < .05)" (p. 819). This and similar statements were coded as interpreted.

Table 2. Reporting and Interpretation Frequency of Uncorrected and Corrected Effect Sizes.

Journal	No. of Articles Using Mult. Regression	No. of Mult. Regression Analyses	No. not Reporting an Effect Size	No. Reported	No. Interpreted	No. Reported	No. Interpreted
Journal of Applied Psychology	9	22	0 (0.00)	16 (72.73)	7 (31.82)	9 (39.13)	0 (0.00)
Journal of Educational Psychology	11	28	12 (42.85)	15 (53.57)	9 (32.14)	1 (3.57)	1 (3.57)
Journal of Educational Research	4	9	3 (33.33)	4 (44.44)	3 (33.33)	3 (33.33)	2 (22.22)
Journal of Experimental Education	1	2	0 (0.00)	2 (100.00)	0 (0.00)	0 (0.00)	0 (0.00)
Total	25	61	15 (24.59)	37 (60.65)	19 (31.15)	13 (20.97)	3 (4.92)

Note. The first, second, and third multiple regression analyses were considered from each article. Percentages are presented in parentheses under selected frequencies. The number of uncorrected and corrected effects may sum to greater than the total number of analyses due to the fact that some analyses reported both types of effects.

Results

Reporting Frequency

Reporting frequencies of uncorrected and corrected effects are displayed in Table 2. Overall, 61% of the analyses reported an uncorrected effect size. Fewer interpreted the effects, however, numbering roughly 50% of the total uncorrected effects reported. These results are relatively consistent with previous studies' findings addressing uncorrected effect sizes and their interpretation (cf. Henson & Smith, 2000; Kirk, 1996; Thompson & Snyder, 1997; Vacha-Haase, Nilsson, Reetz, Lance & Thompson, 2000).

Corrected effects occurred much less often in the literature, showing up in only 21% of the reviewed articles. Interpretation of the corrected effects was even rarer at 5% of all adjusted effects reported. Adjusted R^2 was reported more frequently than other types of corrected effects to the near exclusion of other options (ω^2 was reported in one instance), but the formulae used to calculate adjusted R^2 were not reported in the literature. Nevertheless, one could reasonably surmise that most of the authors likely used the Ezekiel formula (sometimes incorrectly attributed to Wherry [Yin & Fan, 2001]) because it is the formula used by the popular SAS and SPSS statistical software packages to calculate adjusted R^2 (Kirk, 1996; Yin & Fan, 2001). Although use of the Ezekiel correction is better than no correction, the near complete dependence on it in the present review begs the issues of (a) whether authors are reporting adjusted R^2 by default because it is provided in statistical output and (b) whether authors are aware of other correction options.

Adjusted R^2 using Various Formulae

For comparative purposes, we calculated adjusted effect sizes for all analyses that included an uncorrected effect size. We used information provided by the journal authors to adjust R^2 using each of the formulae listed in Table 1. Tables 3 and 4 present the uncorrected R^2, followed by the values calculated for the population and future sample adjustment formulae, respectively. These calculations demonstrate the amount of correction for each formula in relation to the uncorrected effect size, number of predictors, and sample size. Tables 5, 6, and 7 present the degrees of shrinkage for each of the adjustment formulae categorized by sample size, uncorrected R^2, and number of predictors, respectively.

Discussion

Reporting and Interpretation Frequency of Adjusted Effects

As noted, the reporting and interpretation of adjusted effects was rare. In fact, only 3 of the 62 (4.92%) regression analyses reviewed in this study reported and interpreted an adjusted effect. Given the

Table 3. Adjusted R2 Using Population Effect Adjustment Formulae

N	*k*	Reported Adj. R^2	R^2	Smith	Ezekiel	Wherry-2	Olkin-Pratt	Pratt	Claudy-3
578	2	-	.01	.0066	.0066	.0083	.0066	.0066	.0083
1340	3	-	.02	.0178	.0178	.0185	.0178	.0178	.0186
473	1	-	.03	.0279	.0279	.0300	.0281	.0281	.0302
1261	4	-	.03	.0269	.0269	.0277	.0270	.0270	.0277
99	2	-	.05	.0304	.0302	.0402	.0316	.0309	.0417
463	8	-	.12	.1045	.1045	.1065	.1049	.1049	.1069
463	8	-	.12	.1045	.1045	.1065	.1049	.1049	.1069
465	1	-	.13	.1281	.1281	.1300	.1286	.1286	.1305
62	1	-	.14	.1259	.1257	.1400	.1309	.1296	.1456
465	1	-	.14	.1381	.1381	.1400	.1387	.1387	.1405
465	1	-	.14	.1381	.1381	.1400	.1387	.1387	.1405
1515	6	-	.14	.1366	.1366	.1372	.1367	.1367	.1373
1515	6	-	.14	.1366	.1366	.1372	.1367	.1367	.1373
1515	6	-	.14	.1366	.1366	.1372	.1367	.1367	.1373
99	2	-	.17	.1529	.1527	.1614	.1559	.1555	.1647
343	10	-	.17	.1451	.1450	.1476	.1458	.1457	.1483
664	12	-	.18	.1649	.1649	.1662	.1653	.1653	.1666
37	2	-	.19	.1437	.1424	.1669	.1536	.1499	.1784
463	8	-	.20	.1859	.1859	.1877	.1866	.1866	.1884
187	11	-	.22	.1713	.1710	.1757	.1727	.1725	.1772
62	3	-	.24	.2014	.2007	.2142	.2073	.2062	.2207
99	3	-	.26	.2369	.2366	.2446	.2408	.2404	.2487
170	8	-	.26	.2235	.2232	.2280	.2255	.2253	.2301
24	3	-	.29	.1886	.1835	.2224	.2064	.1979	.2442
36	3	-	.37	.3127	.3109	.3318	.3262	.3236	.3467
45	3	-	.42	.3786	.3776	.3924	.3898	.3885	.4044
35	4	-	.50	.4355	.4333	.4516	.4500	.4481	.4672
289	7	.54	.55	.5388	.5388	.5404	.5405	.5405	.5421
412	7	-	.59	.5829	.5829	.5839	.5841	.5841	.5851
289	7	.61	.62	.6106	.6105	.6119	.6122	.6122	.6136
25	3	-	.63	.5795	.5771	.5964	.5999	.5978	.6181
170	8	-	.64	.6222	.6221	.6244	.6249	.6249	.6272
288	8	.65	.66	.6503	.6503	.6515	.6518	.6518	.6531
25	3	-	.67	.6250	.6229	.6400	.6444	.6427	.6605
35	4	-	.74	.7065	.7053	.7148	.7182	.7176	.7270
35	3	-	.75	.7266	.7258	.7344	.7380	.7376	.7462
38	4	-	.80	.7765	.7758	.7824	.7855	.7852	.7916
	M	.54	.31	.2870	.2864	.2938	.2917	.2910	.2989
	SD	.06	.24	.2382	.2379	.2392	.2414	.2413	.2426

Note. Population adjusted effect estimates were calculated for analyses that reported uncorrected R^2 values. Analyses that did not include an uncorrected effect size were not included. *n*=sample size. *k*=number of predictor variables.

fact that adjusted effects theoretically provide the researcher with a more realistic picture of the treatment effect, this result is surprisingly low.

Comparison of Various Adjustment Formulae

For those analyses not reporting an adjusted effect, we calculated and compared adjustments using the each of the fourteen adjustment formulae. Of the adjusted R^2 formulae estimating the population effect, the Ezekiel formula provided the most conservative correction for sampling error while the

Table 4. Adjusted R2 Using Future Sample Effect Adjustment Formulae.

N	*k*	Reported Adj. R^2	R^2	Lord-1	Lord-2	Darlington -Stein	Browne	Claudy-1	Claudy-2	Rozeboom -1	Rozeboom -2
578	2	-	.01	-.0003	.0014	.0014	.0081	.0038	.0048	.0031	.0043
1340	3	-	.02	.0141	.0149	.0149	.0185	.0157	.0163	.0156	.0158
473	1	-	.03	.0218	.0238	.0238	.0297	.0260	.0279	.0259	.0260
1261	4	-	.03	.0223	.0231	.0230	.0276	.0240	.0246	.0238	.0241
99	2	-	.05	-.0094	.0008	-.0002	.0389	.0154	.0198	.0108	.0180
463	8	-	.12	.0851	.0871	.0867	.1076	.0901	.0906	.0891	.0908
463	8	-	.12	.0851	.0871	.0867	.1076	.0901	.0906	.0891	.0908
465	1	-	.13	.1225	.1244	.1244	.1295	.1263	.1281	.1263	.1263
62	1	-	.14	.0827	.0975	.0965	.1363	.1121	.1252	.1118	.1124
465	1	-	.14	.1326	.1344	.1344	.1395	.1363	.1381	.1363	.1363
465	1	-	.14	.1326	.1344	.1344	.1395	.1363	.1381	.1363	.1363
1515	6	-	.14	.1320	.1326	.1326	.1375	.1332	.1337	.1332	.1332
1515	6	-	.14	.1320	.1326	.1326	.1375	.1332	.1337	.1332	.1332
1515	6	-	.14	.1320	.1326	.1326	.1375	.1332	.1337	.1332	.1332
99	2	-	.17	.1181	.1270	.1261	.1607	.1363	.1436	.1358	.1367
343	10	-	.17	.1150	.1176	.1166	.1509	.1220	.1217	.1202	.1231
664	12	-	.18	.1473	.1485	.1482	.1686	.1504	.1507	.1498	.1508
37	2	-	.19	.0471	.0728	.0658	.1651	.1016	.1150	.0974	.1043
463	8	-	.20	.1683	.1701	.1697	.1901	.1723	.1733	.1719	.1726
187	11	-	.22	.1130	.1178	.1138	.1849	.1281	.1233	.1225	.1309
62	3	-	.24	.1352	.1491	.1451	.2176	.1649	.1722	.1627	.1659
99	3	-	.26	.1977	.2058	.2043	.2472	.2144	.2202	.2138	.2145
170	8	-	.26	.1773	.1821	.1796	.2368	.1893	.1892	.1869	.1901
24	3	-	.29	.0060	.0474	.0152	.2347	.1013	.0940	.0871	.1078
36	3	-	.37	.2125	.2344	.2233	.3467	.2570	.2652	.2555	.2560
45	3	-	.42	.3068	.3222	.3160	.4071	.3374	.3457	.3371	.3360
35	4	-	.50	.3333	.3524	.3367	.4952	.3714	.3736	.3710	.3671
289	7	.54	.55	.5244	.5260	.5256	.5515	.5277	.5289	.5277	.5275
412	7	-	.59	.5738	.5748	.5746	.5923	.5758	.5767	.5758	.5757
289	7	.61	.62	.5984	.5998	.5994	.6247	.6011	.6022	.6011	.6009
25	3	-	.63	.4890	.5095	.4943	.6533	.5266	.5332	.5291	.5200
170	8	-	.64	.5998	.6021	.6009	.6519	.6045	.6056	.6044	.6038
288	8	.65	.66	.6381	.6393	.6389	.6677	.6406	.6414	.6406	.6403
25	3	-	.67	.5443	.5625	.5489	.7026	.5774	.5836	.5800	.5710
35	4	-	.74	.6533	.6632	.6551	.7922	.6715	.6743	.6729	.6669
35	3	-	.75	.6855	.6945	.6898	.7826	.7020	.7070	.7031	.6994
38	4	-	.80	.7394	.7463	.7411	.8597	.7519	.7544	.7529	.7487
	M	.54	.31								
	SD	.06	.24	.2489	.2565	.2528	.3076	.2649	.2676	.2640	.2646
				.2341	.2340	.2331	.2581	.2332	.2335	.2343	.2316

Note. Future sample adjusted effect estimates were calculated for analyses that reported uncorrected R^2 values. *n*=sample size. *k*=no. of predictor variables.

Table 5. Degree of Shrinkage Categorized by Number of Predictors

Population Effect Estimates												
	Smith		Ezekiel		Wherry-2		Olkin-Pratt		Pratt		Claudy-3	
K	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1-2[a]	.0120	.0148	.0122	.0152	.0048	.0079	.0097	.0118	.0104	.0130	.0022	.0053
3-4[b]	.0391	.0266	.0404	.0279	.0267	.0180	.0284	.0218	.0300	.0239	.0153	.0127
5-7[c]	.0063	.0034	.0063	.0035	.0054	.0030	.0055	.0027	.0055	.0027	.0045	.0022
8-9[d]	.0182	.0094	.0183	.0095	.0159	.0082	.0169	.0091	.0169	.0091	.0146	.0079
10+[e]	.0296	.0173	.0297	.0174	.0269	.0157	.0288	.0168	.0288	.0169	.0260	.0152

Future Sample Effect Estimates																
	Lord-1		Lord-2		Darlington-Stein		Browne		Claudy-1		Claudy-2		Rozeboom-1		Rozeboom-2	
K	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1-2	.0392	.0454	.0315	.0372	.0326	.0394	.0059	.0082	.0229	.0281	.0177	.0242	.0240	.0295	.0222	.0272
3-4	.1062	.0738	.0919	.0632	.1010	.0717	-.0050	.0330	.0773	.0504	.0735	.0516	.0781	.0531	.0790	.0499
5-7	.0146	.0078	.0136	.0073	.0138	.0075	-.0001	.0031	.0126	.0069	.0119	.0066	.0126	.0069	.0127	.0070
8-9	.0411	.0213	.0387	.0200	.0396	.0209	.0064	.0134	.0355	.0180	.0349	.0184	.0363	.0188	.0353	.0177
10+	.0649	.0381	.0620	.0363	.0638	.0382	.0219	.0121	.0565	.0320	.0581	.0348	.0592	.0346	.0551	.0308

Note. N = 37 analyses. Shrinkage = uncorrected R^2 – adjusted R^2. [a]n = 9. [b]n = 13. [c]n = 6. [d]n = 6. [e]n=3.

Table 6. Degree of Shrinkage Categorized by Sample Size.

Population Effect Estimates												
Study	Smith		Ezekiel		Wherry-2		Olkin-Pratt		Pratt		Claudy-3	
N	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1-30[a]	.0656	.0311	.0688	.0327	.0438	.0207	.0464	.0323	.0505	.0361	.0224	.0203
31-50[b]	.0414	.0159	.0427	.0164	.0280	.0116	.0298	.0145	.0314	.0155	.0155	.0097
51-100[c]	.0225	.0096	.0228	.0098	.0119	.0095	.0187	.0088	.0195	.0088	.0077	.0091
100+[d]	.0115	.0121	.0115	.0122	.0097	.0113	.0071	.0068	.0107	.0118	.0089	.0109

Future Sample Effect Estimates																
Study	Lord-1		Lord-2		Darlington-Stein		Browne		Claudy-1		Claudy-2		Rozeboom-1		Rozeboom-2	
N	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1-30	.1836	.0873	.1569	.0745	.1772	.0848	-.0002	.0483	.1282	.0527	.1264	.0605	.1313	.0622	.1304	.0452
31-50	.1132	.0439	.0977	.0375	.1060	.0409	.0112	.0361	.0825	.0307	.0764	.0302	.0829	.0318	.0845	.0305
51-100	.0691	.0214	.0560	.0201	.0576	.0215	.0119	.0068	.0434	.0189	.0358	.0200	.0450	.0192	.0425	.0189
100+	.0265	.0263	.0247	.0253	.0252	.0262	.0050	.0106	.0147	.0135	.0217	.0242	.0229	.0243	.0220	.0226

Note. N = 37 analyses. Shrinkage = uncorrected R^2 – adjusted R^2. [a]n = 3. [b]n = 7. [c]n = 5. [d]n = 22.

Table 7. Degree of Shrinkage Categorized by Uncorrected R^2.

	Population Effect Estimates											
	Smith		Ezekiel		Wherry-2		Olkin-Pratt		Pratt		Claudy-3	
R^2	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
.01-.15[a]	.0065	.0065	.0066	.0065	.0036	.0049	.0059	.0060	.0060	.0061	.0029	.0053
.16-.30[b]	.0366	.0261	.0374	.0275	.0205	.0179	.0320	.0214	.0335	.0237	.0213	.0140
.31-.50[c]	.0544	.0018	.0561	.0124	.0381	.0104	.0413	.0101	.0433	.0106	.0239	.0086
.51+[d]	.0231	.0154	.0239	.0163	.0170	.0097	.0150	.0082	.0156	.0089	.0086	.0032

	Future Sample Effect Estimates															
	Lord-1		Lord-2		Darlington-Stein		Browne		Claudy-1		Claudy-2		Rozeboom-1		Rozeboom-2	
R^2	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
.01-.15	.0189	.0193	.0160	.0159	.0162	.0162	.0039	.0045	.0125	.0120	.0103	.0110	.0130	.0129	.0129	.0115
.16-.30	.0955	.0751	.0842	.0632	.0896	.0725	.0223	.0142	.0699	.0480	.0677	.0510	.0732	.0521	.0521	.0461
.31-.50	.1458	.0286	.1270	.0260	.1380	.0306	.0137	.0093	.1081	.0234	.1018	.0262	.1088	.0236	.0236	.0247
.51+	-.0604	.0447	-.0532	.0376	-.0581	.0432	-.0229	.0210	-.0047	.0318	-.0443	.0297	-.0462	.0308	-.0308	.0344

Note. N = 37 analyses. Shrinkage = uncorrected R^2 – adjusted R^2. [a]n = 14. [b]n = 10. [c]n = 3. [d]n = 10.

Table 8. *Selected Study R^2 Adjustments*

			Population Effect Estimates							
Study *n*	*k*	R^2	Smith	Ezekiel	Wherry-2	Olkin-Pratt	Pratt	Claudy-3		
38	4	.80[a]	.7765	.7758	.7824	.7855	.7852	.7916		
45	3	.42[b]	.3786	.3776	.3924	.3898	.3885	.4044		
25	3	.63[c]	.5795	.5771	.5964	.5999	.5978	.6181		
463	8	.12[d]	.1045	.1045	.1065	.1049	.1049	.1069		
24	3	.29[e]	.1886	.1835	.2224	.2064	.1979	.2442		
			Future Sample Effect Estimates							
Study *n*	*k*	R^2	Lord-1	Lord-2	Darlington-Stein	Browne	Claudy-1	Claudy-2	Rozeboom-1	Rozeboom-2
38	4	.80	.7394	.7463	.7411	.8597	.7519	.7544	.7529	.7487
45	3	.42	.3068	.3222	.3160	.4071	.3374	.3457	.3371	.3360
25	3	.63	.4890	.5095	.4943	.6533	.5266	.5332	.5291	.5200
463	8	.12	.0851	.0871	.0867	.1076	.0901	.0906	.0891	.0908
24	3	.29	.0060	.0474	.0152	.2347	.1013	.0940	.0871	.1078

Note: [a]Roth, Speece, & Cooper (2002); [b]Marks, Sabella, Burke, & Zaccaro (2002); [c]Gefland, Nishii, Holcombe, Dyer, Ohbuchi, & Fukuno (2001);, [d]Harackiewicz, Barron, Tauer, & Elliot (2002).

Claudy-3 formula offered the least conservative correction. Table 3 illustrates this trend for these adjustments. One can infer that most adjusted R^2 effects presented in the literature offer a conservative estimate since the Ezekiel correction is used in the SAS and SPSS (Kirk, 1996) software packages commonly used by researchers. The uninitiated researcher may not know, however, that these software packages use a formula that estimates only the population parameter.

As illustrated in Table 4, the majority of the future sample effect estimates provided even more conservative estimates than those predicting the population effect. Of these adjustment formulae, the Lord-1 provided the most conservative estimate of the future sample effect while the Browne formula provided the most liberal overall. It is important to note that nine of the 62 adjustments using the Browne formula actually resulted in corrected effects that were greater than the uncorrected effects – a logically impossible result. That is, a sample cannot possess less sampling error than a population that, by

definition, has no sampling error at all. This phenomenon with the Browne formula begs further investigation. It would seem prudent to use caution with the Browne formulae for correction of effects of greater magnitude.

In shrinking R^2, adjustment formulae consider the three factors that affect sampling error: (a) sample size, (b) number of variables in the model, and the (c) uncorrected effect size (as an estimate of the population effect). It naturally follows that these three factors would affect the degree of correction provided by the adjustments to R^2.

Table 5 provides evidence of the number of predictor's impact on the degree of shrinkage. Generally, as the number of predictors increased, the degree of shrinkage increased as well. Our results may be somewhat inconsistent, however, as the analyses with 5-7 predictors did not always show the upward trend as expected. This may be due to the fact that the majority of the analyses in this group had large sample sizes. In fact, no analysis in this group reported a sample size less than 289 subjects. Consequently, this group may not be representative of the adjustment trend based on the number of predictors typically found in the research literature; the large sample sizes may have lessened the degree of correction for this group.

Thompson and Melancon (1990) reported that "with a very large effect size, or a large sample size, or both, it will matter less which, if any, statistical corrections the researcher applies in estimating effect sizes" (p. 11). This proposition is supported by the data in Table 6. As sample size increased, the amount of correction lessened, although in varying amounts based on the formulae. This is logical given that as sample size increases, sampling error – the issue for which adjustments are made – decreases. More specific evidence of this fact is provided in Table 7. Given the case with a large sample size (n = 463) and a small effect size ($R^2 = 0.12$), the correction was relatively small.

Thompson and Melancon (1990) also noted the converse – that statistical corrections tend to be greater when effect sizes and sample sizes are small. This can be noted generally in Tables 6 and 7. As sample size decreased, adjustments generally increased. Again, it is interesting to note that a smaller sample size typically results in greater theoretical sampling error in a sample. Because adjustments to R^2 are determined by the degree of sampling error, it follows that one could expect a large correction given a small sample size.

The correction for one case detailed in Table 8 provides specific evidence for this proposition. With a small sample size (n=24) and a moderate effect size (R^2=0.29), the adjustment was relatively large. Moreover, our results indicate that, in this case, result interpretation may have been different if the author had calculated adjusted R^2. The uncorrected R^2 presented in the journal article indicated that the treatment explained 29% of the dependent variable variance. When corrected using the Lord-1 and Darlington-Stein formulae, however, R^2 shrunk to near zero (.0060 and .0152, respectively.) In other words, after having corrected the effect size for sampling error expected upon replication of the study with a new sample, the treatment accounted for virtually no variance in the dependent variable, a fact that was obscured by the uncorrected R^2.

As demonstrated by the previous case, it is quite possible to overestimate the importance of a result if effect sizes are not adjusted to account for the influence of sampling error. Accordingly, researchers should report and interpret corrected effect measures in their results. Not only do corrected effects provide a better estimate of the effect in the population, they can provide information concerning the replicability of the results. When a researcher uses corrected effect sizes, we recommend that he or she take into account the various formulae, their purposes, and their relative degrees of correction. Such choices have the potential to directly impact results and their interpretation.

References

*(References marked with an asterisk indicate studies included in review)

American Psychological Association. (1994). Publication Manual of the American Psychological Association (4th ed.). Washington, DC: Author.

American Psychological Association. (2001). Publication Manual of the American Psychological Association (5th ed.). Washington, DC: Author.

*Bartol, K. M., Durham, C. C., Poon, J. M. L. (2001). Influence of performance evaluation rating segmentation on motivation and fairness perceptions. Journal of Applied Psychology, 86, 1103–1119.

Berkson, J. (1942). Tests of signficance considered as evidence. Journal of American Statistical Association, 37, 325-335.

*Braaksma, M. A. H., Riglaarsdam, G., van den Bergh, H. (2002). Observational learning and the effects of model-observer similarity. Journal of Educational Psychology, 94, 405–415.

*Broekkamp, H., van Hout-Wolters, B. H. A. M., Rijlaarsdam, G., & van den Bergh, H. (2002). Importance in instructional text: Teachers' and students' perceptions of task demands. Journal of Educational Psychology, 94, 260–271.

*Cappella, E., Weinstein, R. S. (2001). Turning around reading achievement: Predictors of high school students' academic resilience. Journal of Educational Psychology, 93, 758–771.

Carter, D.S. (1979). Comparison of different shrinkage formulas in estimating the population multiple correlation coefficients. Educational and Psychological Measurement, 39, 261-266.

Cohen, J. (1994). The earth is round (p<.05). American Psychologist, 49, 997-1003.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. The Journal of Educational Research, 94, 275–282.

Finch, S., Cumming, G., Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Pscyhology: Little evidence of reform. Educational and Psychological Measurement, 61, 181-210.

*Gelfand, M. J., Nishii, L. H., Holcombe, K. M., Dyer, N., Ohbuchi, K., & Fukuno, M. (2001). Cultural influences on cognitive representations of conflict: Interpretations of conflict episodes in the United States and Japan. Journal of Applied Psychology, 86, 1059–1074.

*Gentry, M., Gable, R. K., & Rizza, M. G. (2002). Students' perceptions of classroom activities: Are there grade-level and gender differences? Journal of Educational Psychology, 94, 539–544.

*Griffiths, Y. M. & Snowling, M. J. (2002). Predictors of exception word and nonword reading in dyslexic children: The severity hypothesis. Journal of Educational Psychology, 94, 34–43.

*Harackiewica, J. M., Barron, K. E., Tauer, J. M., & Elliot, A.J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. Journal of Educational Psychology, 94, 562–575.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). What if there were no significance tests? Mahwah, NJ: Erlbaum.

*Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking Ability Tests in a Selection setting: Implications for practice effects, training performance, and turnover. Journal of Applied Psychology, 87, 243–254.

*Helwig, R., Anderson, L., & Tindal, G. (2001). Influence of elementary student gender on teachers' perceptions of mathematics achievement. The Journal of Educational Research, 95, 93–102.

Henson, R. K. (2006). Effect size measures and meta-analytic thinking in counseling
psychology research. The Counseling Psychologist, 34, 601-629.

Henson, R. K., & Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the APA Task Force Report and current trends. Journal of Research and Development in Education, 33, 285-296.

*Hill, N. E. (2001). Parenting and academic socialization as they relate to school readiness: The roles of ethnicity and family income. Journal of Educational Psychology, 93, 686–697.

*Hinds, P. J., Patterson, M., & Pfeffer, J. (2001). Bothered by abstraction: The effect of expertise on knowledge transfer and subsequent novice performance. Journal of Applied Psychology, 86, 1232–1243.

Huberty, C. J. (1987). On statistical testing. Educational Researcher, 16, 4-9.

Huberty, C. J. (2002). A history of effect size indices. Educational and Psychological Measurement, 62, 227-240.

Huberty, C. J., & Pike, C. J. (1999). On some history regarding statistical testing. In B. Thompson (Ed.), Advances in social science methodology (Vol. 5, pp. 1-22). Stamford, CT: JAI Press.

*Jockin, V., Arvey, R. D., & McGue, M. (2001). Perceived victimization moderates self-reports of workplace aggression and conflict. Journal of Applied Psychology, 86, 1262–1269.

*Johnstone, K. M., Ashbaugh, H., & Warfield, T. D. (2002). Effects of repeated practice and contextual-writing experiences on college students' writing skills. Journal of Educational Psychology, 94, 305–315.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746–759.

*Kitsantas, A. (2002). Test preparation and performance: A self-regulatory analysis. The Journal of Experimental Education, 70, 101-113.

*Klein, K. J., Conn, A. B., & Sorra, J. S. (2001). Implementing computerized technology: An organizational analysis. Journal of Applied Psychology, 86, 811–824.

Kline, R.B. (2004). Beyond significance testing: Reforming data analysis methods in behavioral research. Washington, D.C.: American Psychological Association.

Larson, S.C. (1931). The shrinkage of the coefficient of multiple correlation. Journal of Educational Psychology, 22, 45-55.

*Marks, M. A., Sabella, M. J., Burke, C. S., & Zaccaro, S. J. (2002). The impact of cross-training on team effectiveness. Journal of Applied Psychology, 87, 3–13.

*McGregor, H. A. & Elliot, A. J. (2002). Achievement goals as predictors of achievement-relevant processes prior to task engagement. Journal of Educational Psychology, 94, 381–395.

*Miller, D. C. & Byrnes, J. P. (2001). To achieve or not to achieve: A self-regulation perspective on adolescents' academic decision making. Journal of Educational Psychology, 93, 677–685.

*Okpala, C. O., Okpala, A. O., & Smith, F. E. (2001). Parental involvement, instructional expenditures, family socioeconomic attributes, and student achievement. The Journal of Educational Research, 95, 110–115.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparitive studies: Applications, interpretations, and limitations. Contemporary Educational Psychology, 25, 241-286.

Onwuegbuzie, A.J., Levin, J.R., & Leech, N.L. (2003). Do effect-size measures measure up?: A brief assessment. Learning disabilities: A contemporary journal, 1(1): 37-40.

Pedhazur, E. J. (1997). Multiple Regression in Behavioral Research (3rd ed.). Fort Worth: Harcourt Brace.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillside, NJ: Erlbaum.

*Pomerantz, E. M., Altermatt, E. R., & Saxon, J. L. (2002). Making the grade but feeling distressed: Gender differences in academic performance and internal distress. Journal of Educational Psychology, 94, 396–404.

Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. Educational and Pscyhological Measurement, 62, 241-253.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276–1284.

*Roth, F. P., Speece, D. L., & Cooper, D. H. (2002). A longitudinal analysis of the connection between oral language and early reading. The Journal of Educational Research, 95, 259–272.

*Schleicher, D. J., Day, D. V., Mayes, B. T., Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. Journal of Applied Psychology, 87, 735–746.

*Settles, I. H., Sellers, R. M., & Damas, A., Jr. (2002). One role or two? The function of psychological separation in role conflict. Journal of Applied Psychology, 87, 574–582.

Snyder, P. & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334–349.

Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. Journal of Personality, 62, 157–176.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25, 26–30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. Educational Researcher, 26, 29–32.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them [Electronic version]? Theory & Psychology, 9, 165–181.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? Journal of Counseling and Development, 80(1), 64-72.

Thompson, B., & Kieffer, K. M. (2000). Interpreting statistical significance test results: A proposed new "what if" method. Research in the Schools, 7(2), 3–10.

Thompson, B., & Melancon, J. G. (1990, November). Bootstrap versus statistical effect size corrections: A comparison with data from the finding embedded figures test. Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans.

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in the Journal of Experimental Education [Electronic version]. Journal of Experimental Education, 66, 75–83.

Tyler, R. W. (1931). What is statistical significance? Educational Research Bulletin, 10, 115-118, 142.

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. Theory & Psychology, 10, 413-425.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. Journal of Counseling Psychology, 51, 473-481.

*Wilkins, J. L. M. & Ma, X. (2002). Predicting student growth in mathematical content knowledge. The Journal of Educational Research, 95, 288-298.

Wilkinson, L., & American Psychological Association (APA) Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations [Electronic version]. American Psychologist, 54, 594-604.

Yin, P. & Fan, X. (2001). Estimating R^2 shrinkage in multiple regression: A comparison of analytical methods. The Journal of Experimental Education, 69, 203-224.

Send correspondence to: Robin K. Henson
University of North Texas
Email: rhenson@unt.edu

Performance of the Roy-Bargmann Stepdown Procedure as a Follow Up to a Significant MANOVA

W. Holmes Finch
Ball State University

The Roy-Bargmann procedure has been suggested as a post hoc procedure for a significant MANOVA result. This method, which is based on application of the univariate General Linear Model, requires the researcher to order the dependent variables a priori in terms of their contextual importance. Subsequently, if the MANOVA is found to be statistically significant for a categorical independent variable, these response variables are tested individually using univariate analyses in the sequence established a priori. Thus, the first variable is treated as the dependent variable in an ANOVA and if the groups' means are found to differ significantly on this variable, it serves as a covariate while the second variable in the sequence is the response and the means of the groups are once again compared. This testing continues for each of the response variables, with variables higher in the sequence serving as covariates for those lower in importance. The current study was designed to examine the Type I error rate and power of this post hoc approach under a variety of data conditions. Results show that factors such as distribution of the dependent variables, equality (or lack thereof) of the covariance matrices and sample size all have a significant impact on Type I error and power. Furthermore, both Type I error rates and power of variables later in the sequence are influenced by variables earlier in the sequence.

Multivariate Analysis of Variance (MANOVA) is a popular tool used by social science researchers and others, allowing for the analysis of multiple dependent variables with one or more independent factors. The null hypothesis being tested by MANOVA is $\mu_1 = \mu_2 = \mu_3$ where μ_k is the vector of means for group *k*. When this hypothesis is rejected due to a significant test statistic, researchers may be interested in to which groups or dependent variable(s) the result applies. Given that rejection of the very general null hypothesis of the MANOVA indicates that there is some difference among the *k* groups on one or more of the *p* dependent variables. In order to gain a more complete understanding of the nature of such a significant effect, a researcher may want to use a follow up analysis designed to illuminate the significant result in terms of group differences on the response variables (Tabachnick & Fidell, 2007; Stevens, 1996). A number of such approaches have been discussed in the literature, including the Simultaneous Test Procedure (STP) (Gabriel, 1968), Descriptive Discriminant Analysis (DDA) (Huberty, 1994), a Step Down procedure (SD) (Roy, 1958), two groups multivariate comparisons (Stevens, 1972) and the use of univariate Analysis of Variance (ANOVA). It should be noted that with the exception of the latter approach, all of these methods retain the general multivariate flavor of the original analysis, albeit in very different ways. Indeed, several authors (e.g. Stevens, 1996) argue that whatever follow up to MANOVA is finally used, it needs to be based upon a multivariate platform. Nonetheless, most researchers who make use of MANOVA will have specific questions regarding the nature, in terms of both the response variables and the groups, of the significant differences signaled by the multivariate analysis. This study was designed to examine the performance of one of these follow up methods that may be effective in characterizing a significant MANOVA result.

Analysis of Variance (ANOVA)

Perhaps the most straightforward approach to investigating a significant MANOVA result is through the application of individual univariate ANOVA analyses for each of the dependent variables separately. This approach is facilitated by common statistical software packages such as SAS and SPSS, which print the univariate results with the multivariate. Despite this ease of use, the use of univariate ANOVA in this way has generally been rejected as a viable alternative for following up a significant MANOVA result because, as Enders (2003) points out, the univariate ANOVA does not accurately maintain the nominal Type I error rate in most cases (generally being too conservative), even when a correction such as Bonferroni or Holm is used. Indeed, Maxwell (1992) found that using such alpha corrections with univariate ANOVA to maintain the nominal experiment-wise Type I error rate only works when either the MANOVA null hypothesis is totally false, the MANOVA null hypothesis is totally true or the MANOVA null hypothesis is false for all but one of the dependent variables. In all other cases, using ANOVA to investigate a significant MANOVA will yield an incorrect Type I error rate. Keselman, Huberty, Lix, et al

(1998) assert that if the researcher is interested in a multivariate hypothesis then the follow up to a significant MANOVA should be multivariate in nature.

Two groups multivariate comparison

Another MANOVA follow up approach that has been suggested in the literature is the two groups multivariate comparison. As outlined by Stevens (1972), the two groups method involves using the Hotelling's T^2 test statistic (Hotelling, 1931) to compare all possible pairs of groups on the entire set of dependent variables simultaneously. As an example, if the MANOVA involved one independent categorical variable with 3 groups measured on 3 dependent variables, the two groups analysis would involve the multivariate comparison of groups 1 and 2, groups 1 and 3, and groups 2 and 3 on the 3 response variables simultaneously. Stevens (1972) compared this approach with DDA, SD and multivariate contrasts and found that the results, in terms of identifying group differences, were fairly similar, though the two groups approach did not allow for direct identification of group differences on individual variables. Thus, when two groups are found to differ the difference is for the entire set of responses, so that if a researcher would like to identify clearly for which of the response variables groups differ, (s)he could not do so using this method.

Simultaneous Test Procedures

Gabriel (1968) expanded upon earlier work of Roy and Bose (1953) to develop multivariate simultaneous confidence intervals for group differences using any one of the common MANOVA test statistics, such as Roy's greatest root, Pillai's Trace, Hotelling's trace and Wilks' Lambda. Various recommendations have been made regarding which of these is the most generally appropriate for use in the STP context (Sheehan, 1995; Elliott, 1993; Mudholkar, Davidson & Subbaiah, 1974; Olson, 1974). Simulation research on the power and Type I error control of the STP based on one or more of these statistics found that violations of the assumptions of normality and homogeneity of covariance matrices had a significant impact on their performance, such that no one approach could be identified as optimal in all cases (Sheehan, 1995; Bird & Hadzi-Pavlovic, 1983; Elliott, 1993). These studies also found that the more restrictive the hypotheses being tested, the lower the power of the STP procedure, regardless of the statistic used to form the confidence intervals. Indeed, Sheehan (1995) concluded that this approach was too conservative to investigate group differences on individual variables.

Descriptive Discriminant Analysis

Several authors have recommended the use of DDA as an appropriate follow up procedure for a significant MANOVA (Enders, 2003; Huberty, 1994; Tabachnick & Fidell, 2007; Stevens, 1996). The fact that DDA is a multivariate technique very closely related to MANOVA addresses concerns voiced by Keselman, et al. (1998) and Stevens (1996) that when the research questions of interest are multivariate in nature, then the analyses used to address these questions be multivariate as well. DDA involves the identification of a linear combination of the dependent variables that maximizes the differences among the groups, as expressed in the between and within sums of squares and cross products matrices, $\mathbf{S_B}$ and $\mathbf{S_W}$, respectively. These linear combinations of the original set of dependent variables can then be used to characterize the nature of the difference(s) among groups identified by the significant MANOVA. Interpretation of these discriminant functions can be carried out in at least two ways, using either Structure Coefficients (SC) (Huberty, 1994), which are values representing the correlations between individual observed variables and the overall discriminant function or standardized discriminant function coefficients (Rencher, 1992). While there is some disagreement as to which approach is preferable, in both cases the magnitudes of either of these values reflect the relative importance of each observed variable in defining the discriminant function, making them useful for identifying potentially important contributors (from among the set of dependent variables) to the observed differences among the groups, as well as providing some insight into the conceptual nature of the discriminant functions themselves (Enders, 2003; Huberty, 1994; Stevens, 1972; Rencher, 1992).

The fact that SC's may be used in practice to ascertain which of the observed variables are related to the one or more discriminant functions (Tabachnick & Fidell, 2007) makes the issue of their interpretation very important. By examining the magnitudes of SC's, a researcher can determine which of the dependent variables are most associated with the linear combination(s) that best differentiate the groups in question. Those with the largest SC's are deemed to be most associated with the group differences, and are thus interpreted more fully in terms of how they might differ from group to group. This difference is

often characterized by examining the means of the groups on those variables with sufficiently large SC values (Tabachnick & Fidell, 2007). A difficulty with employing this method is that there is not a formal hypothesis test available for the SC's, and thus various rules of thumb have been recommended as cutoffs for identifying "sufficiently large" values. Schneider (2004) conducted a simulation study using cut points of 0.3, 0.4 and 0.5 and found that the general ability of DDA to identify statistically meaningful variables, in terms of group differentiation, was lower for higher cut offs, except where the differences in group means was characterized by a large effect size. Schneider also found that the use of these various cutoff values led to the false identification of "important" variables at rates reaching as high as 0.8 in many cases. In other words, a researcher using these rules of thumb, would be likely to incorrectly conclude that particular variables were "important" in defining the discriminant function when they were in fact not.

Stepdown Analysis

The Stepdown analysis (SD) examined here was introduced by Roy (1958) and Roy and Bargmann (1958) and extended by others (Marden & Perlman, 1990; Mudholkar & Subbaiah, 1988; Kabe, 1984) and is based on the General Linear Model (GLM) in the form of Analysis of Covariance (ANCOVA). It has been recommended as a follow up procedure for a significant MANOVA by several authors, including Tabachnick and Fidell (2007), Stevens (1996), and Mudholkar and Subbaiah (1980). In addition, the SD technique can be used to derive a test of overall significance in the multivariate context as demonstrated by Kabe (1984), among others. The SD procedure involves a multi-step application of univariate linear models involving the dependent variables, much in the way that ANOVA does. However, there are some major differences between SD and ANOVA that make the former method potentially appealing as a MANOVA follow up, from a statistical perspective. Indeed, from one perspective the omnibus MANOVA test may not be required, given that an a priori ordering of the dependent variables is made by the researcher, implying a very specific set of hypotheses to be tested. Nonetheless, given the recommendations by the authors cited above to use it as a way to elucidate a significant MANOVA result, it is in this context that the current study was conducted.

As mentioned above, the SD procedure involves several steps. In the first step the researcher places the dependent variables in descending order of theoretical importance. It is crucial to note that this ordering is based on contextual issues and not on statistics. Indeed, step one should always be done prior to any data collection so that sample estimates do not affect how variables are ordered. In step two, an ANOVA is conducted in which group means on the most important response variable are compared. In step three, this variable serves as a covariate and the means of the groups on the second most important response variable are compared using ANCOVA. This stepping procedure continues for each dependent variable in turn, with response variables at a higher level of importance serving as covariates for those of lesser importance.

It has been shown that in the 2 groups case, an overall SD test statistic can be constructed that is equivalent to Hotelling's T^2 (Mudholkar & Subbaiah, 1980) allowing for an omnibus multivariate test. By including responses as covariates in subsequent analyses, correlations among the variables are accounted for in a way that they are not in the ANOVA analyses described above (Mudholkar & Subbaiah, 1975). In order to control the Type I error rate, Bock and Haggard (1968) suggested using a variation of the Bonferroni approach by dividing the overall *a* so that more important variables are given a larger share of the rejection region than less important ones, while maintaining the desired overall level of significance. It would also be possible to simply assign each test the same portion of the rejection region using Bonferroni's correction more directly (Tabachnick & Fidell, 2007).

Mudholkar and Subbaiah (1980) cited several potential advantages to using the SD procedure as a follow up to a significant MANOVA result: 1) simplicity, 2) detailed results for specific variables and groups, 3) useful with small samples and 4) results for large samples that are equivalent to the omnibus likelihood ratio test. Another potential strength of the SD is that a researcher can assign different levels of *a* to the dependent variables, reflecting their substantive importance in the investigation of which the MANOVA is a part (Subbaiah & Mudholkar, 1978; Mudholkar & Subbaiah, 1976; Bock & Haggard, 1968). Mudholkar and Subbaiah (1988) also suggest that under the SD procedure, test statistics such as the Studentized Range can be used for paired group comparisons, providing an added degree of analytic flexibility.

While there are clear benefits to the researcher using the SD technique, it does have some weaknesses as well, perhaps foremost of which is the need to order the response variables in terms of theoretical importance. It has been clearly demonstrated that the qualitative conclusions reached by researchers can differ substantially depending upon the variables' ordering (Stevens, 1972; Koslowsky & Caspy, 1991; Mudholkar & Subbaiah, 1988). Indeed, Stevens (1996) states that the SD procedure may not be appropriate when a clear a priori ordering of the response variables is not possible. Mudholkar and Subbaiah (1980) suggest that when such an ordering is not reasonable, SD can still be used as a post hoc procedure for MANOVA if the observed variables are first subjected to analysis using a data reduction technique creating linear combinations, such as Principal Components Analysis. These linear combinations could in turn be used with SD techniques, with combinations accounting for greater variance being placed higher in the sequence. Koslowsky and Caspy (1991) observed that by applying the SD to multiple sequences of the response variables a researcher could engage in meaningful data exploration and testing of multiple hypotheses. At the same time, Stevens (1996) points out that these various orderings are not independent of one another so that the actual Type I error rate across all of them is not known.

There has been some research examining the performance of the SD procedure in various conditions. Subbaiah and Mudholkar (1978) conducted a Monte Carlo simulation study in which they assessed the performance of the omnibus multivariate test statistic for the SD procedure. They simulated data sets from the multivariate normal distribution with 2 response variables, equal covariance matrices across 2 groups, each with samples of size 20. They found that the SD procedure was able to maintain the nominal Type I error rates (0.01 or 0.05) while attaining power values above 0.8 for most studied conditions. They also found that power was affected by the relative importance of the dependent variables. If they were of unequal importance and this is reflected by unequal *a* values, the power of the SD procedure was higher than when all variables were treated as equally important in terms of a. Finally, Subbaiah and Mudholkar reported that the precision of the hypothesis tests decreased for variables later in the ordering. Mudholkar and Srivastava (2000) used a robust method based on trimmed estimates to conduct a SD analysis when the data were not normally distributed and samples were of unequal size. They found that this approach had greater power than did the standard parametric ANCOVA approach.

In addition to these simulations, other researchers have reported results obtained when using the SD with real data. Stevens (1972) used both SD and DDA to compare 4 groups of subjects on 8 response variables. The two methods yielded very similar results in terms of identifying variables on which the groups differed, while the application of univariate ANOVA's produced a markedly different outcome. Stevens also demonstrated the impact on the SD analysis of different orderings of the dependent variables. Koslowsky and Caspy (1991) reported on a refinement of the SD technique in which multiple orderings of the response variables are tested in following up a significant MANOVA. If the qualitative results of these analyses differ, the researcher could conclude that there is overlap among the dependent variables. Such a result would, according to the authors, help the user gain a greater understanding of the interrelationships among the responses while also allowing for the examination of multiple hypotheses about how the groups in question might differ on the measures of interest. These authors acknowledge that when the number of dependent variables is 3 or more, the resulting Type I error rate could be somewhat inflated. Analytic results in Mudholkar and Subbaiah (1975) match the findings of their simulation study (Mudholkar & Subbaiah, 1978), demonstrating that the power of hypothesis tests for variables lower in importance (and thus tested later in the sequence) was lower than when the variables were placed higher in the sequence.

Focus of the Current Study

This study is designed to extend the work in studies reviewed above. Specifically, the goals of this research are to ascertain the performance of the SD method for following up a significant MANOVA in terms of both power and Type I error rate and to identify data specific factors influencing the performance of this approach. Given that the SD approach has been recommended in popular multivariate texts for use in the post hoc investigation of significant MANOVA results, and has been studied relatively infrequently in this role, it is hoped that this study helps to fill a gap in the literature. As has been reported above, a number of other approaches for following up a significant MANOVA have major problems that have been identified using Monte Carlo methods. For example, the STP appears to have low statistical power to investigate group differences on individual response variables, while DDA does not allow for hypothesis testing of individual dependent variables and cutoff values used with it appear to be somewhat

problematic to interpret. The two groups multivariate approach, while appearing to be reasonably effective in differentiating multivariate group means, does not easily allow for comparisons on individual variables. There has been relatively little work conducted in investigating the performance of the SD approach in terms of power and Type I error rates in the post hoc context, with most prior research focusing on its use in constructing an omnibus test statistic. It is hoped, therefore, that this study will add greater understanding in this regard.

Methodology

A Monte Carlo simulation study was used to investigate the performance of the SD follow up for MANOVA. In this case, focus was on the SD results, rather than those of MANOVA, so that results of the omnibus MANOVA were not used in the conduct of the simulation. In other words, while in actual practice the SD procedure would not be used unless a significant MANOVA were first obtained, because the focus of the current study was on the SD procedure, an assumption was made that the omnibus MANOVA test was found to be statistically significant. A number of factors were manipulated in this study with all combinations being crossed, and 1000 replications were generated for each. All simulations were conducted using SAS IML and PROC GLM. The following factors were manipulated:

Sample size. The total sample size conditions were 30, 60, 100 and 150. These values were selected in an effort to replicate sample size conditions appearing in published research using MANOVA.

Sample size ratio. Two sample size ratio conditions were simulated, including equal and unequal group sizes. In the unequal condition, the first group (group 1) had half the number of subjects as did group 2.

Number of dependent variables. Data were simulated with either 2 or 4 dependent variables. Variables earlier in the sequence were assumed to be more important than those later in the sequence and were tested as such. The lower value (2) was simulated so as to provide information about the simplest case possible, while the latter (4) was selected because it conforms to examples used in prior research (e.g., Fan & Wang, 1999).

Correlation. The dependent variables were simulated with pooled within groups correlations of 0.3, 0.5 and 0.8. All inter-variable correlations were the same in the 4 variables condition.

Effect size. A total of 8 effect size combinations, based on Cohen's *d* (Cohen, 1988), were simulated in this study to create univariate group separation. The effect sizes used were 0 (no group separation), 0.5 and 0.8. The latter values were chosen so as to correspond with Cohen's (1988) benchmarks of medium and large effects. The combinations used appear in Table 1. It is important to note that effect size values for variables 2, 3 and 4 in the 4 variable case are identical. In the case where the group variances were equal, this value was used as the denominator in the calculation of effect size, whereas in the unequal variances case, a pooled within-cell standard deviation was used.

Distribution of response variables. The dependent variables were distributed either as standard normal or with skewness of -1.5 and kurtosis of 3, with the latter distribution selected because it conforms to that used in earlier research and which has been shown to be associated with inflated Type I error rates in standard MANOVA testing (Finch, 2005). In order to maintain the desired correlation values among these variables in the non-normal condition, a method proposed by Fleishman (1977) was used to simulate the data.

Covariance matrices. The group covariance matrices were simulated to either by equal or unequal. In the latter condition, the variances for group 2 were simulated to 5 be times larger than that for group 1, which was set at a value of 1. The latter value matches that which has been demonstrated previously to lead to inflated Type I error rates for MANOVA test statistics (e.g., Finch, 2005; Sheehan-Holt, 1998).

The outcomes of interest were the Type I error rate and power for each variable in the sequence. In order to ascertain which of the manipulated factors had a significant impact on the performance of the SD follow up, ANOVA and variance components analysis were used, treating either Type I error rate or power as the dependent variable and the manipulated factors as the independent.

Table 1: Effect size combinations for simulation

Variable 1	Variable 2 (3 and 4)
0.5	0
0.8	0
0	0.5
0	0.8
0.5	0.5
0.8	0.8
0.8	0.5
0.5	0.8

Results

Type I error rate

The results of the ANOVA/variance components analyses for the 2 and 4 variable cases yielded similar results in terms of identifying factors that influenced the Type I error rate of the SD procedure. For this reason, results of the ANOVA are only reported for the 2 variable case. Results for the second variable in the two groups case are nearly identical to those for variables 2, 3 and 4 in the 4 variable case. Specifically, for the Type I error rate of the first response variable (variable 1), the 2-way interaction of equality of the covariance matrices and the distribution of the variables was the highest order significant interaction, and accounted for 90.9% of the observed variability. The 3-way interaction of the correlation between the response variables by equality of the covariance matrices by variable distribution was the highest order significant term for the second variable (variable 2), accounting for 46% of the variation in Type I error rates. The 2-way interaction of correlation and response variable distribution was also significant for variable 2 and accounted for an additional 20% of variation in Type I error rate. Other main effects and interactions were also statistically significant for the Type I error rates of both variables, but none accounted for more than 10% of the variation, and will therefore not be considered further.

Table 2 contains the Type I error rates for the 2 and 4 variable cases by the inter-variable correlation, covariance status (equal or unequal) and distribution of the response variables. It appears that the Type I error rate for the first variable in the set is near the nominal level of 0.05, except when the assumptions of normality and equality of covariance matrices underlying the ANCOVA have both been violated. The Type I error rates for variables 2 through 4 were found to be inflated for most of the conditions displayed in Table 2. The Type I error inflation for these variables also increased somewhat with increasing correlation among the variables. Although all Type I error conditions for the later variables in the sequence were inflated, the severity of this inflation was less when the assumptions of normality and equality of covariance matrices were both met.

Power

The ANOVA and variance components analyses for the power rates indicated that for variable 1, the statistically significant interaction between covariance matrix equality/inequality and distribution of the variables accounted for 78.5% of the variability in power. No other term in the model was both statistically significant and accounted for more than 10% of the variance in power for variable 1. Three terms were found to be statistically significant and account for more than 10% each of the variation in power for variable 2 (3 and 4). The highest order interaction, accounting for 25% of power variation, was covariance matrix equality/inequality by correlation by distribution, while effect size (15.5%) and sample size (12.1%) were the two statistically significant main effects that accounted for more than 10% of variation in power. As with the Type I error rates, the same factors contributed significantly to observed power in the 2 and 4 variable cases.

Table 3 displays the power rates by covariance matrix equality/inequality, correlation and distribution for all of the variables. One outcome to note is that across these conditions power rates generally diminish through the sequence from variable 1 to 4. Power for the first variable in the set was higher in the normal distribution condition when the groups' covariance matrices were also equal than when the data were normal but the variances were unequal. However, when the covariance matrices were unequal, the opposite pattern can be seen, where variable 1 power was greater in the skewed condition. It

Table 2: Type I error rate by correlation among dependent variables, covariance equality/inequality and response distribution: 2 and 4 variables

2 response variables				
Correlation	Covariance	Distribution	V1	V2
0.3	Equal	Normal	0.051	0.141
		Skewed	0.047	0.252
	Unequal	Normal	0.036	0.063
		Skewed	0.848	0.674
0.5	Equal	Normal	0.047	0.142
		Skewed	0.044	0.259
	Unequal	Normal	0.038	0.132
		Skewed	0.852	0.784
0.8	Equal	Normal	0.053	0.234
		Skewed	0.046	0.317
	Unequal	Normal	0.034	0.488
		Skewed	0.850	0.827

4 response variables						
Correlation	Covariance	Distribution	V1	V2	V3	V4
0.3	Equal	Normal	0.050	0.143	0.138	0.107
		Skewed	0.047	0.251	0.212	0.181
	Unequal	Normal	0.050	0.144	0.144	0.140
		Skewed	0.849	0.674	0.540	0.512
0.5	Equal	Normal	0.053	0.135	0.125	0.150
		Skewed	0.044	0.266	0.267	0.249
	Unequal	Normal	0.036	0.137	0.133	0.151
		Skewed	0.852	0.690	0.652	0.640
0.8	Equal	Normal	0.051	0.236	0.228	0.208
		Skewed	0.049	0.366	0.450	0.362
	Unequal	Normal	0.034	0.490	0.324	0.287
		Skewed	0.849	0.870	0.831	0.830

is important to keep in mind that it was in this combination of conditions (unequal covariance matrices and skewed data) that the Type I error rate for variable 1 was elevated. For this reason, these high power rates in the skewed/nonnormal condition are not particularly meaningful. With respect to variables 2, 3 and 4, a pattern similar to that for variable 1 was evident, where power was higher for the normal (versus skewed) distribution when the covariances were also equal, while the opposite pattern by distribution was evident when the covariance matrices were unequal. Again, given the aforementioned Type I error inflation when neither assumption was met, it is not meaningful (nor particularly surprising) that power in this case was also higher.

Table 4 includes power by effect size combination. Please note that the category "50" in the table refers to the condition where the first variable has an effect size difference of 0.5 between the groups for the first response variable and no difference for variable 2 (as well as 3 and 4 where applicable). In like fashion, "05" refers to the condition where the first variable has no group difference but variable 2 (3 and 4) is characterized by an effect size difference of 0.5. The remaining combinations should be interpreted similarly. Results in this table for variable 1 show that a greater effect size, reflecting larger group separation, was associated with higher power. Furthermore, the power for variable 1 was unaffected by the effect size of the accompanying variable(s). Of course, this would be expected given that it is the first variable in the sequence, and thus tested independently of the others.

Results in Table 4 demonstrate that, as with variable 1, greater effect size was associated with higher power for variables 2 through 4. However, an important outcome for these latter variables was that the effect size of the first variable had an impact on their power. For example, in the two variables condition, the power for the second variable in the 05 case (where there was not a group difference for the first response and a difference of 0.5 for the second response) was 0.365, while when the first variable was characterized by a large effect size difference and the second by this same moderate effect (85), the

Table 3: Power by correlation among dependent variables, covariance equality/inequality and response distribution: 2 and 4 variables

2 response variables				
Correlation	Covariance	Distribution	V1	V2
0.3	Equal	Normal	0.721	0.513
		Skewed	0.193	0.122
	Unequal	Normal	0.359	0.243
		Skewed	0.872	0.693
0.5	Equal	Normal	0.718	0.454
		Skewed	0.194	0.176
	Unequal	Normal	0.357	0.228
		Skewed	0.873	0.136
0.8	Equal	Normal	0.717	0.456
		Skewed	0.194	0.278
	Unequal	Normal	0.355	0.297
		Skewed	0.872	0.070

4 response variables						
Correlation	Covariance	Distribution	V1	V2	V3	V4
0.3	Equal	Normal	0.687	0.360	0.400	0.301
		Skewed	0.176	0.097	0.096	0.073
	Unequal	Normal	0.690	0.361	0.397	0.296
		Skewed	0.873	0.682	0.522	0.398
0.5	Equal	Normal	0.688	0.352	0.323	0.218
		Skewed	0.175	0.173	0.076	0.057
	Unequal	Normal	0.336	0.150	0.144	0.099
		Skewed	0.869	0.593	0.544	0.531
0.8	Equal	Normal	0.691	0.464	0.315	0.207
		Skewed	0.175	0.312	0.079	0.052
	Unequal	Normal	0.334	0.254	0.157	0.103
		Skewed	0.875	0.535	0.522	0.521

Table 4: Power by effect size: 2 and 4 variables

2 response variables				
Effect size combination	V1		V2	
50 / 05*	0.424		0.365	
80 / 08*	0.610		0.582	
85	0.610		0.188	
58	0.426		0.449	
88	0.611		0.235	
55	0.424		0.158	
4 response variables				
Effect size combination	V1	V2	V3	V4
50 / 05*	0.454	0.395	0.237	0.163
80 / 08*	0.641	0.613	0.365	0.255
85	0.643	0.199	0.115	0.092
58	0.454	0.479	0.267	0.183
88	0.643	0.263	0.177	0.132
55	0.452	0.178	0.126	0.099

*The effect size combination to the left of / is for variable 1, while to the right of / is the corresponding combination for variable 2.

Table 5: Power by sample size: 2 and 4 variables

2 response variables				
Sample size	V1		V2	
30	0.296		0.158	
60	0.509		0.263	
100	0.626		0.359	
150	0.711		0.441	
4 response variables				
Sample size	V1	V2	V3	V4
30	0.307	0.140	0.102	0.080
60	0.524	0.235	0.173	0.121
100	0.640	0.330	0.255	0.178
150	0.719	0.406	0.327	0.239

power for variable 2 was 0.188. In other words, the power for detecting mean differences for variable 2 decreased between the two examples, though the degree of group separation in the two cases was identical. This result was present across all effect size combinations for all three latter variables in the sequence. As noted above, power declined from variable 2 through variable 4 in the testing sequence. This outcome was true even though the variables were characterized by comparable effect sizes in the 4 variables condition. In other words, even though in the population group separation was equivalent for variables 2, 3 and 4, power was higher for those variables tested earlier in the sequence.

Table 5 contains power rates by sample size. In interpreting all of these power results, it is important to keep in mind that the Type I error rates presented in Table 2 were elevated above the nominal error rate of 0.05 for many of the conditions studied here. Power for all variables in both the 2 and 4 variable conditions increased concomitantly with increases in sample size. As was evident from results in Table 4, power was lower for variables tested later in the sequence than those tested earlier.

Discussion

The purpose of this study was to explore the effectiveness of the Roy-Bargmann stepdown procedure as a follow up to a significant MANOVA as has been recommended in some popular multivariate statistics texts in the social sciences (Tabachnick & Fidell, 2007; Stevens, 1996). Using this method, a researcher would conclude that dependent variables for which significant results on the individual ANCOVA's were found could be identified as "important" contributors to the group separation identified by the MANOVA (Mudholkar and Subbaiah, 1980). The results of this study indicate that when both the assumptions of normality and equality of covariance matrices are not met, the Type I error rate for the first variable in the sequence was inflated, while in all other cases it was near the nominal 0.05 level. This result should serve to encourage researchers using the SD as a post hoc to a significant MANOVA to assess the viability of both the normality and equality of covariance matrices assumptions prior to making use of the technique. However, the Type I error rates of subsequent variables in the sequence were inflated in most cases studied here, including when the assumptions underlying ANCOVA are met. This widespread Type I error inflation for the latter variables in the testing sequence must call into question the findings for power, though they have been presented in order to fully inform the reader as to the results of the study. In general, power rates declined further into the testing sequence so that even with the same effect size in the population, power was lower for testing variable 2 than for variable 1 and for variable 3 than for variable 2, etc. Furthermore, the power for later variables in the sequence was influenced by the effect size of the first variable, such that larger group differences on variable 1 were associated with lower power on variables 2, 3 and 4, regardless of the effect size characterizing these latter variables.

Implications for practice

It should be remembered that this study was conducted under the presumption that a researcher would elect to use the SD approach prior to the conduct of analysis. Thus, by implication a reasonable ordering of the variables, based on contextual issues germane to the area being studied and not statistical concerns, is possible. In cases where this ordering is not reasonable, the SD procedure would not be an

appropriate approach, and some other method would clearly seem to be better as a follow-up for MANOVA.

The results of this research must, to some degree, call into question the use of the Roy-Bargmann stepdown procedure as a follow-up to a significant MANOVA, despite recommendations for its use (e.g., Tabachnick & Fidell, 2007; Stevens, 1996). Prior research (e.g., Kabe, 1984; Mudholkar & Subbaiah, 1980) has shown that this approach works well in many conditions in terms of constructing an omnibus test of significance in the multivariate case. However, the current study does not support the use of this technique in trying to identify specific variables for which the groups might differ, other than the first one. When groups didn't differ on the latter variables in the testing sequence, the Type I error rate was generally inflated above the nominal rate, while when they did differ, power tended to be lower than for variable 1. Furthermore, in cases where the first variable in the sequence was characterized by a large effect size (0.8), power for the latter variables was even more diminished.

For these reasons, it does not seem prudent to use this technique to ascertain which variables might be significantly different between groups. This does not call into question the potential utility of this procedure for an omnibus test in the MANOVA context, as mentioned previously. In that case, the hypothesis being tested is very different from those addressed by the method as studied here. In addition, given the inflated Type I error rates for the tests of all but variable 1, practitioners who do choose to use this approach in the manner outlined here should be very wary of significant results for all but the first variable in the testing sequence, particularly in (though not limited to) cases when the assumptions of normality and homogeneity of variances are not met.

Limitations and directions for future research

There are some methodological limitations to this study that should be addressed in future research efforts. Perhaps chief among these is the somewhat limited set of effect size values used. The 6 combinations reported here were selected in order to allow for an investigation of Type I error rate and power under a variety of conditions while maintaining a manageable study size. These values selected corresponded to the benchmarks of medium and large effects laid out by Cohen (1988). It is recognized, however, that more work is needed with a greater variety of patterns in order to gain a more complete understanding of the performance of the Roy-Bargmann approach to post hoc testing for a significant MANOVA. Such a future effort should include both a different set of effect size values, particularly those smaller than 0.5, as well as a more complex pattern of values in the 4 variables condition. In the current study, variables 2, 3 and 4 all had the same effect size value for all conditions, which may not be representative of the wide array of possibilities seen in real data. At the same time, given that this study is one of the first Monte Carlo efforts to examine the performance of the Roy-Bargmann method in this MANOVA post hoc context, it was felt that in order to maintain a manageable study design and ease interpretation, this somewhat limited set of combinations was advisable. Furthermore, it is not clear that major differences in results would be expected from those presented here.

A second recommendation for future research would be to include more group conditions. As stated above, it was felt that since this study was examining this approach in a fairly different way from previous research, a simpler design was preferable. However, in order to broaden its generalizability, future efforts should include more groups.

References

Bird, K.D. & Hadzi-Pavlovic, D. (1983). Simultaneous test procedures and the choice of a test statistic in MANOVA. *Psychological Bulletin, 93*, 167-178.

Bock, R.D. & Haggard, E. (1968). The use of Multivariate Analysis of Variance in research. In D.K. Whitla (Ed.), *Handbook of measurement assessment in behavioral sciences*. Reading, MA: Addison Wesley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, Hillsdale, NJ: Lawrence Erlbaum.

Elliott, R. (1993). *Contrast selection in post hoc multivariate analysis*. Unpublished Doctoral Dissertation, Ohio University-Chillicothe.

Enders, C.K. (2003). Performing multivariate group comparisons following a statistically significant MANOVA. *Measurement and Evaluation in Counseling and Development, 36*, 40-56.

Fan & Wang. (1999). Comparing linear discriminant function with logistic regression for the two-group classification problem. *The Journal of Experimental Education*, 67(3), 265-286.

Fleishman, A.I. (1977). A method for simulating non-normal distributions. *Psychometrika, 43*, 521-532.

Finch, W.H. (2005). Comparison of the performance of nonparametric and parametric MANOVA test statistics when assumptions are violated. *Methodology*, 1, 27-38.

Gabriel, K. (1968). Simultaneous test procedures in MANOVA. *Biometrika, 55*, 489-504.
Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics*, 360-378.
Huberty, C.J. (1994). *Applied Discriminant Analysis.* New York: John Wiley and Sons, Inc.
Kabe, D.G. (1984). On the maximal invariance of MANOVA stepdown procedure statistics. *Communications in Statistics: Theory and Methods, 13*, 2571-2581.
Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R.A., Donohue, B., Kowalchuk, R.K., Lowman, L.L., Petosky, M.D., Keselman, J.C., & Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research, 68*, 350-368.
Kirk, R.E. (1995). *Experimental Design*. Pacific Grove, CA: Brooks/Cole Publishing Company.
Kowlowsky, M. & Caspy, T. (1991). Stepdown analysis of variance: A refinement. *Journal of Organizational Behavior, 12*, 555-559.
Marden, J.I. & Perlman, M.D. (1990). On the inadmissibility of step-down procedures for the Hotelling T^2 problem. *The Annals of Statistics, 18*, 172-190.
Maxwell, S.E. (1992). Recent developments in MANOVA applications. In *Advances in Social Science Methodology: Volume 2* (pp.137-168). Greenwich, CT: JAI Press.
Mudholkar, G.S., Davidson, M.L. & Subbaiah, P. (1974). Extended linear hypotheses and simultaneous tests in Multivariate Analysis of Variance. *Biometrika, 61*, 467-477.
Mudholkar, G.S. & Srivastava, D.K. (2000). A class of robust stepwise alternatives to Hotelling's T^2 tests. *Journal of Applied Statistics, 27*, 5999-619.
Mudholkar, G.S. & Subbaiah, P. (1975). A note on MANOVA multiple comparisons based upon step-down procedure. *The Indian Journal of Statistics, 37*, 300-307.
Mudholkar, G.S. & Subbaiah, P. (1980). A review of step-down procedures for Multivariate Analysis of Variance. In R.P. Gupta, (Ed.), *Multivariate Statistical Analysis (pp.161-178)*. Amsterdam: North-Holland.
Mudholkar, G.S. & Subbaiah, P. (1988). On a fisherian detour of the step-down procedure for MANOVA. *Communications in Statistics: Theory and Methods, 17*, 599-611.
Olson, C.L. (1974). Comparative robustness of six tests in Multivariate Analysis of Variance. *Journal of the American Statistical Association, 69*, 894-908.
Rencher, A.C. (1992). Interpretation of canonical discriminant functions, canonical variates, and principal components. *The American Statistician, 46*, 217-225.
Roy, J. (1958). Step-down procedure in multivariate analysis. *Annals of Mathematical Statistics, 29*, 1177-87.
Roy, S.N. & Bose, R.C (1953). Simultaneous confidence interval estimation. *Annals of Mathematical Statistics, 24*, 513-536.
Roy, S.N. & Bargmann, R.E. (1958). Tests of multiple independence and the associated confidence bounds. *The Annals of Mathematical Statistics, 29*, 491-503.
Schneider, M. K. (2004). Comparison of the usefulness of within-group and total-group structure coefficients for identifying variable importance in descriptive discriminant analysis following a significant MANOVA: Examination of the two-group case. *Multiple Linear Regression Viewpoints, 30*, 8-18.
Sheehan-Holt, J.K. (1998). MANOVA simultaneous test procedures: The power and robustness of restricted multivariate contrasts. *Educational and Psychological Measurement, 58*, 861-881.
Sheehan, J.K. (1995). A comparison of the Type I error and power of selected MANOVA simultaneous test procedures. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA, April 18-22.
Stevens, J.P. (1972). Four methods of analyzing between variation for the K-group MANOVA problem. *Multivariate Behavioral Research, 7*, 499-522.
Stevens, J.P. (1996). *Applied Multivariate Statistics for the Social Sciences*. Mahwah, NJ: Lawrence Erlbaum and Associates, Publishers.
Subbaiah, P. & Mudholkar, G.S. (1978). A comparison of two tests for the significance of a mean vector. *Journal of the American Statistical Association, 73*, 414-418.
Tabachnick, B.G. & Fidell, L.S. (2007). *Using Multivariate Statistics*. Boston: Pearson.

Send correspondence to: W. Holmes Finch
Ball State University
Email: whfinch@bsu.edu

The Use of Propensity Score Analysis to Address Issues Associated with the Use of Adjusted Means Produced by Analysis of Covariance

John W. Fraas Ashland University
Isadore Newman University of Akron
Scott Pool Ashland University

It is common for researchers in the field of education to engage in research that involves two groups (e.g., control and experimental) and not have the opportunity to randomly assign the participants to the groups. The challenge facing educational researchers is how to analyze the differences between two groups when randomization is not possible and selection bias is an issue. An analytic technique commonly used by educational researchers to address this challenge is analysis of covariance. The use of this technique, however, raises two concerns: (a) The inclusion of the covariates in the analysis of the criterion variable may change the construct represented by the criterion variable, and (b) the analytic technique employed does not match the research question, which is a Type VI error. A technique referred to as propensity score analysis, which is also designed to deal with selection bias in the comparison of non-randomized group means, may address these two concerns. How this technique can be applied by educational researchers is presented.

Various analytic methods have been proposed to control nuisance variables in the analysis of two or more groups (Kirk, 1982). One approach is to randomly assign participants to groups. It is common, however, for researchers in the field of education to engage in research that involves two groups (e.g., control and experimental) and not have the opportunity to randomly assign the participants to the groups. As noted by Halperin and Jorgensen (1994), there are two broad classes of research designs where such randomization is not possible (a) observational studies as identified by Cochran (1983) and (b) quasi-experimental designs as discussed by Campbell and Stanley (1963).

McNeil, Newman, and Kelly (1996) stated that "some statisticians . . . take the position that lack of random assignment disallows a meaningful conclusion" (p.155). They further state, however, that it is their "position . . . that research and decisions must be made in the real world. Random assignment of [subjects to] groups is ideal, but insight can be gained when this is not possible" (p.155).

One analytic approach suggested by Kirk (1982) that can be used to analyze data obtained from a non-randomized research design involved the sub-classification of the participants on key covariates and the inclusion of these sub-classifications in the analysis. It should be noted, however, that the sub-classification procedure can be difficult when multiple key covariates are identified (Halperin & Jorgensen, 1994). As the number of covariates increases, the number of sub-classifications grows exponentially. It is quite possible that some of the sub-classifications will contain none of the study's participants.

A second analytic approach often used by educational researchers when random assignment of subjects is not possible is analysis of covariance (ANCOVA) (Huitema, 1980; McNeil, et al., 1996). In analysis of covariance, the researcher estimates and statistically tests the amount of unique variation in the dependent variable accounted for by the variable or variables representing group membership. One concern with the use of ANCOVA to analyze the difference between group means in non-randomized designs, which is the focus of this article, was discussed by Tracz, Nelson, Newman, and Beltran (2005). Tracz et al. stated that:

> It is important to remember that the outcome or dependent variable in ANCOVA is an adjusted score. . . . After the effects of the covariate have been statistically controlled or removed from the dependent variable . . . , the error variance is all that remains. This residualized or adjusted dependent variable is no longer the same as the original dependent variable. (*p*. 20)

Thus, when the covariates are included in the analysis of the criterion variable, the criterion variable may change as a measure of the construct.

A second concern with the use of ANCOVA deals with the lack of congruency between the research question and the analytic technique, which is referred to as a Type VI error (Newman, Deitchman, Burkholder, & Sanders, 1976; Newman, Fraas, Newman, & Brown, 2002). If the research question deals with student achievement, but the analytic technique analyzes adjusted scores due to the inclusion of covariates, the analytic technique may not produce results that directly address the research question.

One technique that may allow researchers to address these two concerns is propensity score analysis. The next section of this article presents the concept of propensity score analysis and its application to a hypothetical set of data.

Propensity Score Methodology

Rosenbaum (2002) and Rosenbaum and Rubin (1983; 1984) presented an analytic method that used propensity scores to adjust the comparison of non-randomized group means for selection bias due to systematic differences on a set of covariates. Rosenbaum and Rubin (1984) stated:

> There exists a scalar function of covariates, namely the propensity score, that summarizes the information required to balance the distribution of the covariates. Specifically, subclasses formed from the scalar propensity score will balance all . . . covariates. In fact, often five subclasses constructed from the propensity score will suffice to remove over 90% of the bias due to each of the covariates. (*p*. 516)

As noted by Yanovitzky, Zanutto, and Hornik (2005):

> The diagnostics and fitting of the propensity score model are done independent of the outcome and, thus, approximate random assignment of the subjects to treatment . . . . Propensity score methods seek to create comparison groups which are similar (or balanced) on all confounders [covariates] but different on their levels of treatment. (*pp*. 210-211)

We believe this characteristic of the propensity score technique allows researchers to address the selection bias concern with respect to the covariates while not changing the construct represented by the criterion variable. In addition, propensity score analysis may allow the researcher to better match the analytic tool and the research question. Thus, a researcher would not be as likely to commit the Type VI error that involves the analysis of adjusted means when the research question of interest deals with unadjusted means.

Steps Used to Conduct Propensity Score Analysis

Propensity score analysis can be understood by reviewing the steps used to conduct such an analysis. The means of conducting propensity score analysis presented in this article is not meant to be an exhaustive discussion of the various ways researchers can implement the technique, but rather the discussion is meant to provide insight into how this analytic technique may allow researchers to address the two concerns previously mentioned regarding the use of ANCOVA.

Yanovitzky et al. (2005) presented six steps researchers may follow to conduct a propensity score analysis.

> *Step 1--Select the covariates*. The researcher must select, a priori, a set of covariates based on theoretical grounds and previous empirical studies. These covariates are used to estimate the propensity scores used to form sub-groups of participants.
>
> *Step 2--Assess the initial imbalance in the covariates*. The researcher gauges the initial imbalance in each of the covariates with respect to the groups. For covariates with interval or ratio level of measurement, an independent-samples t test can be used, while for dichotomous covariates a z test of differences in proportions can be employed. Assessing the initial imbalance in the covariates is useful for two reasons. First, it allows the researcher to determine if the balance is adequate, that is, the degree of balance one would expect in a completely randomized experiment (see Rosenbaum & Rubin, 1984; Zanutto, Lu, & Hornik, 2005). If the balance is adequate, the researcher does not need to employ the propensity score analytic technique and the group means on the criterion variable can be directly analyzed. Second, the assessment of the initial imbalance serves as a benchmark against which the propensity score methodology has increased the balance in the covariates.
>
> *Step 3--Estimate the propensity scores*. If an imbalance exists between the groups with respect to a number of the covariates, the propensity scores are estimated for each of the participants in the study. These propensity scores can be estimated using a variety of methods. Researchers could use discriminant analysis, probit models, or logistic regression models with the dependent variable being the group variable (e.g., control and experimental) and the covariates serving as the independent variables (D'Agostino, 1998; Rosenbaum & Rubin, 1984). McCaffrey, Ridgeway, and Morral (2004) described the use of generalized boosted regression models, which is a multivariate nonparametric regression technique, to estimate the propensity

scores. Yanovitzky et al. (2005) noted that the use of logistic regression models was the most common method used to generate the propensity scores.

Step 4--Stratify the propensity scores. Once the propensity scores are estimated, they are stratified into four or five levels with equal or nearly equal numbers of subjects in the categories. As noted by Cochran (1983), stratifying into more than 4 or 5 groups usually gains very little.

Step 5--Assess the balance on the covariates across the treatment groups. Once the propensity scores are stratified, the researcher needs to verify that the propensity score groups remove any initial bias on the covariates. Yanovitzky et al. (2005) suggested that this verification procedure can be conducted through the use of a two-way analysis of variance (ANOVA), where the two factors are the treatment groups and the propensity score groups and each covariate is used as the criterion variable. Balance is assumed to be achieved when the treatment main effect and the interaction effect are not statistically significant. Yanovitzky et al. noted that:

> If these two conditions are not met, the propensity score should be re-estimated by adding interaction terms and/or non-linear functions (e.g., quadratic or cubic) of imbalanced covariates to the propensity score model. . . . Steps 3 through 5 are repeated until balance is achieved or no further improvement in balance can be made. (*p*. 214)

Step 6--Estimate and statistically test the difference between the treatment means. In this step, the differences between the treatment means on the criterion variable are calculated and statistically tested for (a) each propensity score group and (b) across all propensity score groups. The statistical tests of the difference between the means in each propensity score group can be conducted with the use of t tests..

As noted by Yanovitzky et al. (2005), the overall estimate of the treatment effect is calculated by averaging the differences between means of the treatment groups across all propensity score groups. The overall treatment effect is calculated as follows:

$$\hat{\delta} = \sum_{k=1}^{4} \frac{n_k}{N}\left(\overline{Y}_{ek} - \overline{Y}_{ck}\right) \tag{1}$$

where, $\hat{\delta}$ is the estimated treatment effect; the propensity score groups (1 through 4) are represented by *k*; *N* is the total number of participants; n_k. is the number of participants in the propensity score group *k*; and the means of the criterion variable for the experimental and control groups within a specific propensity score group are $\overline{Y}_{ek}$ and $\overline{Y}_{ck}$, respectively.

The estimated standard error for the estimated treatment effect is calculated as follows:

$$\hat{s}\left(\hat{\delta}\right) = \sqrt{\sum_{k=1}^{4} \frac{n_k^2}{N^2}\left(\frac{s_{ek}^2}{n_{ek}} + \frac{s_{ck}^2}{n_{ck}}\right)} \tag{2}$$

where, n_k is the number of participants in the *k* propensity score group; *N* is the total number of participants; the sample variances of the experimental and control groups are s_{ek}^2 and s_{ck}^2, respectively; and the number of participants in the experimental and control groups are n_{ek} and n_{ck}, respectively.

The *t* value for the estimated treatment effect is calculated by dividing the estimated treatment effect ($\hat{\delta}$) by its standard error ($\hat{S}(\hat{\delta})$).

Table 1. Descriptive Statistics for the Criterion Variable and Covariates

	Treatment Group[a]			
	Control		Experimental	
Variable	Mean	*SD*	Mean	*SD*
Posttest	40.95	7.27	44.91	6.49
Pretest	23.76	8.29	29.51	7.38
OPT	227.82	26.49	231.54	23.57
Ability	116.02	13.89	113.63	11.31

Note: [a]The sample sizes for the control and experimental groups are 123 and 129, respectively.

Table 2. Comparison of Differences between Control and Experimental Groups on Covariates Before and After Propensity Group Formation

	Pre-Propensity Group Formation	Post-Propensity group formation
Variable	*p*	*p*
Pretest	<0.01	0.41
OPT	0.24	0.66
Ability	0.13	0.90

An Illustration of Propensity Score Analysis

To illustrate the application of propensity score analysis, consider a set of hypothetical data collected from a nonrandomized design. For this example the criterion variable (posttest) is a quantitative variable consisting of scores on a math test administered to the students at the completion of the study. Once the criterion variable was identified, the propensity analysis was conducted by completing the six steps previous presented.

Step 1. The following three covariates were identified:

1. The math pretest covariate, which is labeled pretest, consisted of math scores obtained from a test administered prior to the implementation of the instructional methods.
2. The Ohio Math Proficiency Test (OPT) covariate was composed of student scores on the Ohio Math Proficiency Test administered prior to the implementation of the instructional methods.
3. The covariate labeled ability consisted of student scores on a cognitive ability test administered to the students prior to their exposure to the methods of instruction.

In addition to these three covariates, a dichotomized independent variable was formed to identify the instructional method. For this independent variable, which was labeled treatment, the values of zero (control group) and one (experimental group) were assigned indicating the instructional method to which the students were exposed. The mean and standard deviation values of the criterion variable and the three covariates for both the control and experimental groups are listed in Table 1.

Step 2. The initial imbalances between the treatments on the covariates were determined through the use of independent-samples t tests applied to the pretest, OPT, and ability means for the control and experimental groups. The probability values of these three statistical tests are listed in Table 2 under the heading *Pre-Propensity Group Formation*.

Step 3. A logistic regression model was constructed for the purpose of estimating the propensity score for each student. The treatment variable served as the criterion variable for the model and the covariates and their two-way interaction variables were considered as possible predictor variables. The first procedure used in the construction of the model consisted of entering the three covariates (i.e., pretest, OPT, and ability). The next procedure allowed the three two-way interaction variables formed from the three covariates (i.e., pre-by-OPT, pre-by-ability, and OPT-by-ability) to be entered into the logistic regression model in a step-wise fashion. The step-wise procedure used was a forward method of entry with the criterion for entry set at .05 for the probability of the Wald test value of each two-way interaction variable coefficient.

Once the step-wise procedure was completed the final procedure used to construct the logistic regression model involved in constructing the model consisted of a review of the significance levels of the three covariates (i.e., pretest, OPT, and ability). Any covariate with a non-significant coefficient was deleted unless it was used to form a two-way interaction variable that was entered into the model. The results of the analysis of the logistic regression model,

which included the predictor variables of pretest, OPT, ability, pre-by-ability, and OPT-by-ability, are listed in Table 3.

Step 4. The logistic regression model developed in Step 3 was used to estimate a probability for each of the 252 participants in the study. Each probability value represented the probability that the corresponding participant would be a member of the treatment group (i.e., the group assigned a value of one in the treatment variable) based on that participant's covariate values. These 252 probability values, which are referred to as propensity scores, were stratified into four equal groups of 63 participants as follows:

1. A participant with propensity score less than or equal to the first quartile value was placed in Propensity Group 1.
2. A participant with propensity score greater than the first quartile value but less than or equal to the second quartile value was placed in Propensity Group 2.
3. A participant with propensity score greater than the second quartile value but less than or equal to the third quartile value was placed in Propensity Group 3.
4. A participant with propensity score greater than or equal to the first quartile value was placed in Propensity Group 4.

Step 5. Three two-way ANOVA analyses were used to verify that the propensity score groups removed initial bias on the three covariates with the two main effects consisting of (a) the two treatment groups and (b) the four propensity score groups. The probability value of the F test of the treatment main effect for each of the three analyses is listed in Table 2 under the column entitled Post-Propensity Group Formation. Recall that the column in Table 2 entitled Pre-Propensity Group Formation contains the probability values of the statistical tests of the differences between the covariate means for the control and experimental groups for the three covariates. Since the post-propensity group formation probability values are substantially higher than the pre-propensity group probability values, the propensity group formation is considered to have reduced the initial bias between the treatments with respect to the covariates.

As previously stated Yanovitzky et al. (2005) suggested that balance between the treatment groups with respect to the covariates is assumed to be achieved when both the treatment main effect and the interaction effect between the treatment group and propensity group variable are not statistically significant for the analysis of each covariate. As indicated by the p values listed in the column entitled Post-Propensity Group Formation in Table 2, none of the treatment effects was statistically significant for the three covariates. In addition, the p values for the interaction effects between the treatment and propensity group variables in the three two-way ANOVA analyses of the pretest, OPT, and ability variables were .10, .45, and .19, respectively, which indicates that none of the interaction effects was statistically significant.

Step 6. Table 4 contains the posttest mean and standard deviation values for the control and experimental groups for each of the propensity score groups. The corresponding t test values for the four mean differences are also listed. None of the t values corresponding to the four differences between the posttest means of the experimental and control groups reached the critical t value of 1.67, which corresponds to a one-tailed alpha level of .05. Thus, the difference between the control and experimental groups for each of the propensity score groups was not statistically significant.

Since the results of these four t tests resulted in the same conclusion for each propensity group, that is, none of the differences between the control and experimental posttest means was statistically significant, it was appropriate to test the difference between the overall posttest means of the two groups. The overall treatment effect and its standard error values were calculated using Equations 1 and 2, respectively. The treatment effect value was 1.17, which is also equal to the difference between the overall means listed in Table 4, and the standard error value was 0.77. The *t* value for the overall treatment effect (1.52) was calculated by dividing the treatment effect (1.17) by the standard error (0.77). Since this t test value did not reach the one-tailed critical t value of 1.65, which corresponds to

Table 3. Results for the Logistic Regression Model[a]

Variable	Coefficient	Wald	*p*
Pretest	-0.417	2.21	0.14
OPT	0.255	7.72	<0.01[b]
Ability	0.306	5.81	0.02[b]
Pre x Ability	0.005	4.14	0.04[b]
OPT x Ability	-0.002	8.21	<0.01[b]
Constant	-38.375	6.77	<0.01[b]

Note: [a]Δ(-2 Log likelihood) = 68.995, χ^2 = 98.96, df = 5, p <.01, Cox Snell R^2 =.24, Nagelkerke R^2 = .32.
[b]Statistically significant at the two-tailed α = .05.

Table 5. Results of the ANCOVA Analysis of the Posttest Scores Using MLR Model 2[a]

Variable	Coefficient	*t*	*p*
Treatment[b]	1.49	2.63	<0.01[c]
Pretest	0.40	9.00	<0.01[c]
OPT	0.09	6.27	<0.01[c]
Ability	0.07	2.66	<0.01[c]
Constant	1.83	0.64	0.53

Note: [a]R^2 =.692, df_n = 4, df_d = 247, F =139.00, , p<.01, $\overline{R}^2$ = .687
[b]The proportion of unique variation in the posttest variable accounted for by the treatment variable is .009.
[c]Statistically significant at one-tailed α =. 05.

Table 4. Estimated Treatment Effects on Posttest Math Scores Using Propensity Score Groups

Propensity Score Group	Treatment	Group Size	Mean (SD)	*t*
Group 1	Control	51	38.53 (9.15)	0.63[a]
	Experimental	12	40.25 (5.08)	
Group 2	Control	36	41.17 (5.28)	0.17[a]
	Experimental	27	41.44 (7.40)	
Group 3	Control	27	43.56 (4.11)	0.30[a]
	Experimental	36	43.92 (5.10)	
Group 4	Control	9	46.00 (3.97)	1.25[a]
	Experimental	54	48.33 (5.35)	
Overall	Control	123	42.32[b]	1.52[c]
	Experimental	129	43.49[b]	

Note: [a]Not significant at the one-tailed alpha level of .05 (critical *t* value = 1.67 for comparisons within Propensity Score Groups 1 through 4).
[b]The means are the overall estimates averaged over the propensity score groups. The standard error used to calculate the *t* test value for difference between the two overall estimates is 0.77.
[c]Not significant at the one-tailed alpha level of .05 (critical *t* value = 1.65 for overall comparison).

the significance level of .05, the propensity analysis indicated that the overall treatment effect was not statistically significant.

It is important to note that if one or more of the differences between the posttest means of the control and experimental groups were statistically significant, it would not be appropriate to test the overall posttest difference. The researcher would identify the propensity score group or groups for which the differences in the posttest means were statistically significant, and describe the differences of the characteristics of the students in those groups as compared to the characteristics of the students in the propensity score groups for which the differences in the posttest scores were not statistically significantly different.

An ANCOVA Analysis of the Posttest Scores

To better understand how the results of propensity score analysis differ from results generated from an ANCOVA analysis, it is helpful to compare results produced by both analytic techniques. The next section presents an ANCOVA analysis for the data used in the propensity score analysis.

The initial ANCOVA analysis of the posttest criterion variable, which was conducted with the use of a multiple linear regression model (MLR Model 1), included seven independent variables: (a) treatment, (b) pretest, (c) OPT, (d) ability, (e) pre-by-OPT, (f) pre-by-ability, and (g) OPT-by-ability. Since none of the multiple linear coefficients was statistically significant at the .05 level, a second model (MLR Model 2) was constructed and analyzed that did not include the three two-way interaction variables. That is, the amount of variation in the posttest scores accounted for by the three two-way interaction variables in

MLR Model 1 was pooled into the error term in MLR Model 2. The results of MLR Model 2, which included the treatment independent variable and the pretest, OPT, and ability variables as covariates, are listed in Table 5.

The regression coefficient for the treatment variable (1.49) in the MLR Model 2 estimated the difference between the *adjusted* posttest means of the experimental and control groups. The treatment coefficient indicated that the estimated adjusted posttest mean of the experimental group was 1.49 points higher than the estimated adjusted posttest mean of the control group. The *t* test value (2.63) for this coefficient produced a corresponding one-tailed *p* value that was less than .01. Since this one-tailed *p* value was less than the alpha level of .05, the difference between the estimated adjusted posttest means of the experimental and control groups was statistically significant.

To better understand what the *t* test of the treatment variable coefficient produced by the ANCOVA is testing with respect to the posttest criterion variable, it is helpful to note that this test is equivalent to the statistical test (i.e., the *F* test) of the amount of variation in the posttest variable accounted for by the variation in the treatment variable *over and above* the amount of variation accounted for by the covariates. That is, the *t* test of the treatment coefficient is comparable to testing the amount of *unique* variation in the posttest scores accounted for by the treatment variable. It should be noted that statistically testing the unique *proportion* of variation in the posttest scores accounted for by the treatment variable is comparable to testing the unique *amount* of variation accounted for by the treatment variable.

The formula used to convert the *t* test of the treatment coefficient to the proportion of unique variation in the criterion variable accounted for by the treatment variable is as follows:

$$\Delta R^2 = \frac{t^2\left(1-R^2\right)}{df_d} \tag{3}$$

where: ΔR^2 is the proportion of unique variation accounted for by the treatment variable; t^2 is the square of the *t* test value of the treatment coefficient; and df_d is the denominator degrees of freedom value for the multiple linear regression model.

Substituting the *t* test value of the treatment coefficient (2.63), the R^2 value(.692), and the df_d value (247), which were generated by MLR Model 2, into Equation 3 revealed that the proportion of unique variation in the posttest variable accounted for by the treatment variable was .009. This proportion of unique variation accounted for by the treatment variable can be statistical tested with an *F* test. The statistical test of the .009 value produced an *F* test value of 6.92, which is equivalent to the square of the *t* test value for the treatment coefficient. Since the directional *p* value for this *F* test value ($p < .01$) was less than the .05 alpha level, the proportion of *unique* variation in the posttest scores accounted for by the treatment variable was statistically significant, which must be the case when the difference between the adjusted means is statistically significant.

Comparison of Results and Implications

It is interesting to note that this conclusion differs from the results produced by the propensity score analysis. The *t* test results produced by the propensity analysis indicated that no significant differences existed between the mean posttest scores of the experimental and control groups within each propensity score group and across all propensity score groups. The ANCOVA results, however, revealed that the difference between the *adjusted* posttest scores of the experimental and control groups was statistically significant. Why the difference?

One possible reason for the difference in the results of the two analytic methods is the difference in what is being analyzed by the two methods. In the propensity score analysis, the posttest scores of the control and experimental groups were compared within the propensity groups, that is, within groups of students with similar predicted probabilities of being members of the control or experimental groups based on the covariate variables. Thus, the propensity score analysis did not statistically test *adjusted posttest means*, which is to say, the propensity analysis did not test the amount of unique variation in the criterion variable accounted for by the treatment. In contrast, the difference between the *adjusted* posttest means of the control and experimental groups was analyzed in the ANCOVA, which is synonymous with testing the amount of unique variation in the posttest scores (the criterion variable) accounted for by the treatments.

Which method is appropriate? To address this question, two issues should be considered by the researcher. First, if the researcher is concerned that the construct represented by the criterion variable may be substantially altered by the analysis of adjusted means, which is the issue discussed by Tracz et al. (2005), propensity score analysis may provide an appropriate alternative analytic technique to ANCOVA. Tracz et al. note that if ANCOVA is used, researchers should consider establishing reliability and validity estimates for the adjusted scores, which is no small task and, we suggest, unlikely to be done by most researchers. The use of propensity score analysis provides a less time consuming alternative by not analyzing the amount of unique variation in the criterion variable accounted for by the treatments.

Second, if the researcher is interested in testing the difference between the *posttest means* and not the *adjusted posttest means*, propensity score analysis would be the recommended procedure. If, however, the researcher is interested in testing the difference in the *adjusted means*, ANCOVA will provide that analysis. The key point regarding this issue is that the researcher should strive to match the analytic technique to the research question. That is, the researcher should select the research technique that will not lead to a Type VI error.

Summary

The purpose of this article is to suggest the use of propensity score analysis as an appropriate analytical tool for addressing two concerns expressed in the literature. One of these concerns deals with the issue that the construct represented by the criterion variable may change (Tracz et al., 2005) when the analytic tool used by the researcher analyzes *adjusted* means, that is, the amount of unique variation in the criterion variable accounted for by the treatment variable. If ANCOVA is used, Tracz et al. (p. 20) suggest that the "residualized or adjusted dependent variable is no longer the same as the original dependent variable." In such a case, Tracz et al. recommend that the researcher establish the reliability and validity of the residualized or adjusted scores. Since this would be no small task, researchers may find it more practical to utilize propensity analysis to address selection bias issues because it involves the analysis of *means in propensity score groups* rather than the analysis of *adjusted means*, as is the case in ANCOVA.

The other concern deals with the use of an analytic technique that appropriately matches the research question, that is, the researcher avoids committing a Type VI error (Newman, Deitchman, et al. 1976; Newman, Fraas, et al., 2002). Specifically, if a researcher is concerned with selection bias and the research question involves *unadjusted* posttest scores, propensity analysis will produce results that deal with unadjusted posttest scores, while ANCOVA will not.

References

Campbell, D. T. & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally College Publishing Company.

Cochran, W. G. (1983). *Planning and analysis of observational studies* (L. E. Moses & F. Mosteller). New York: Wiley.

D'Agostino, R. B. (1918). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17,* 2265-2281.

Halperin, S. & Jorgensen, R. (1994, April). *The use of control in non-randomized design.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA: (ED 369 815).

Huitema, B. E. (1980). *The analysis of covariance and alternatives.* New York: Wiley.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences.* (2nd ed.). Belmont, CA: Brooks/Cole.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods (9)*4, 403-425.

McNeil, K., Newman, I., & Kelly F. J. (1996). *Testing research hypotheses with the general linear model.* Carbondale, IL: Southern Illinois University Press.

Newman, I., Deitchman, R., Burkholder, & J., Sanders, R. (1976). Type VI error: Inconsistency between the statistical procedure and the research question. *Multiple Linear Regression Viewpoints, 6*(4), 1-19.

Newman, I., Fraas, J. W., Newman, C., & Brown, R. (2002). Research practices that produce Type VI errors. *Journal of Research in Education, 12*(1), 138-145.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 516-524.

Tracz, S. M., Nelson, L.L., Newman, I., & Beltran, A. (2005). The misuse of ANCOVA: The academic and political implications of Type VI errors in studies of achievement and socioeconomic status. *Multiple Linear Regression Viewpoints, 31*(1), 19-24.

Yanovitzky, I., Zanutto, E., & Hornik, R. (2005). Estimating causal effects of public health education campaigns using propensity score methodology. *Evaluation and Program Planning (28)*, 209-220.

Zanutto, E. L., Lu, B., & Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national anti-drug media campaign. *Journal of Educational and Behavioral Statistics (30)*1, 59-73.

Send correspondence to: John W. Fraas
Ashland University
Email: jfraas@ashland.edu

Estimation Methods for Cross-Validation Prediction Accuracy: A Comparison of Proportional Bias

David A. Walker
Northern Illinois University

Using empirical data, the performance of the predictive effectiveness of four algorithms and a bootstrap method for cross-validation of a multiple regression equation were examined. Results indicated that the Browne algorithm was the most accurate in 8 of the 9 data situations. The Rozeboom algorithm, in a majority of conditions, had the second least amount of proportional bias. The Nicholson and Lord and Stein-Darlington formulas demonstrated a consistent pattern of low relative accuracy in many situations, with the most amount of proportional bias in 4 of 9 and in 6 of 9 data sets, respectively. The bootstrap method showed no discernable pattern of relative accuracy with results ranging from the most accurate in a situation to the least accurate in three different data situations.

For prediction studies derived via multiple regression that estimate how well a sample equation generalizes to other samples, or as Rozeboom (1978, p. 1350) called this "a sample regression's generalized validity," the use of empirical methods such as cross-validation or double cross-validation have been studied in the past (Lord & Novick, 1968; Mosier, 1951). However, there are cautions with these techniques. For example, when applied to small sample sizes, where the sample is split into two sub-samples, this can lead to less precise estimates of prediction (Browne, 1975; Cattin, 1980a; Cotter & Raju, 1982). To resolve this major limitation, the use of estimation algorithms has been noted in the scholarly literature, where the sample size is not split, but left whole, exacting a more accurate estimate of the criterion score estimation for cross-validation of a multiple regression equation (Allen, 1971; Claudy, 1978; Cotter & Raju, 1982; Gollob, 1967; Huberty & Mourad, 1980; Morris, 1984; 1986).

From a review of the literature in the area of proposed cross-validation techniques, the vast majority of the research conducted on the predictive effectiveness of multiple regression equations derived via an algorithm has been conducted as Monte Carlo studies. That is, very few studies reported results from cross-validation algorithms when empirical data were used. Of the small number of studies that did implement empirical data, most used data from large samples derived from business, government, or educational institutions (cf. Cotter & Raju, 1982; Huberty & Mourad, 1980; Kromrey & Hines, 1995). A few exceptions found in the literature that applied empirical data sets from smaller samples were Krus and Fuller (1982) who used cross-validation algorithms with data from a textbook, and Morris (1986) who employed Allen's (1971) PRESS (Predicted Error Sum of Square) technique with data from journal articles, textbooks, and professional conference papers.

Various formulas have been proposed for use as estimation algorithms for cross-validating a regression equation. From the literature, four formulas emerge as viable estimators. Many of the subsequent formulas are decidedly related algebraically and/or are hybrids of one another. Formulas 1, 2, and 4 are found in Huberty & Mourad (1980) and formula 3 is from Cattin (1980b).

Nicholson (1960) and Lord (1950) proposed R^2_{NL}, where:

$$R^2_{NL} = 1 - \left[\frac{N+p+1}{N-p-1}\right]\left[\frac{N-1}{N}\right]\left[1-R^2\right] \quad (1)$$

Stein (1960) and Darlington (1968) proposed R^2_{SD}, where:

$$R^2_{SD} = 1 - \left[\frac{N-1}{N-p-1}\right]\left[\frac{N-2}{N-p-2}\right]\left[\frac{N+1}{N}\right]\left[1-R^2\right] \quad (2)$$

Browne (1975) proposed R^2_B, rearranged by Cattin (1980b), where:

$$R^2_B = \frac{[(N-p-3)(R^2)^2]+R^2}{[(N-2p-2)R^2]+p} \quad (3)$$

Rozeboom (1978) restructured Browne's (1975) algorithm and proposed a simpler version, R^2_R, where:

$$R^2_R = 1-\left[\frac{N+p}{N-p}\right]\left[1-R^2\right] \quad (4)$$

where, N = Sample size, p = Number of X variables, and R^2 = Squared multiple correlation coefficient.

Finally, in prediction studies derived via multiple regression that have the intention of closely approximating a sample prediction equation, the bootstrap method, as well as the leave-one-out method and the jackknife method, have been found to be similar to cross-validation techniques (Efron, 1983; Gong, 2003; Huberty & Mourad, 1980; Kromrey & Hines, 1995; Lachenbruch, 1967). The bootstrap is a resampling method where the sampling properties of a statistic, in this instance R^2, are derived by recomputing its value for artificial samples. Thus, the sample data from this study will serve as pseudo-populations and 1,000 random samples with replacement will be drawn from these full samples. One thousand iterations will be used as an established threshold where all of the nine empirical data sets will have had convergence. Once the bootstrap method is repeated 1,000 times on each empirical data set, a distribution of bootstrapped estimates for R^2 will emerge, where the mean value (i.e., R^2_{BOOT}) of each bootstrapped distribution is the estimate for R^2.

Purpose

Using empirical data, the intention of the current research is to determine the stability of predictive effectiveness of the criterion score estimation of four algorithms and the bootstrap method for cross-validation of a multiple regression equation. The stability of predicative effectiveness is defined as the performance of the techniques in terms of relative accuracy as determined from bias and proportional bias (cf. Aaron, Kromrey, & Ferron, 1998; Morris, 1986). As bias multiples, the distance between the R^2_{CV} value and the R^2 value increases, which leads to diminished stability and a lesser proportion of the criterion score variance accounted for in the predicted Y value than in the original sample's Y value. The measures of bias are:

$$\text{Bias} = R^2 - R^2_{CV} \quad (5)$$

$$\text{Proportional Bias} = \text{Bias} / R^2 = 1 - (R^2_{CV} / R^2) \quad (6)$$

where, R^2_{CV} is defined by either R^2_{NL}, R^2_{SD}, R^2_R, R^2_B, or R^2_{BOOT}.

Thus, this study will compare the performance of four algorithms and the bootstrap method in a three-tier situation: (1) in the first set of empirical data, each will contain two regressor variates (p=2), variable sample sizes (N) = 12, 20, 30, and variable R^2 values = .799, .255, .358; (2) in the second set of data, each will have p =3, N = 20, 30, 93, and R^2 = .943, .617, .261; (3) in the third set of data, each will have p = 4, N = 13, 30, 36, and R^2 = .982, .640, .355.

Methods

The data sets for this research came from Agresti and Finlay (1986), Cohen and Cohen (1983), Hald (1965), Kerlinger and Pedhazur (1973), Rulon, Tiedeman, Tatsuoka, and Langmuir (1967), Sprinthall (2000), and Thurstone (1947). These data are well-known and found in textbooks utilized in graduate-level research design and statistics courses. Also, they exemplify the types of data often applied in social science research, with varying distributional characteristics, multicollinearity, sample sizes, regressor variates, and criterion variables such as predicting grade point average, psychiatric impairment, or job success.

The previously listed N, p, and R^2 values from the data sets will be entered into the four formulas for R^2_{NL}, R^2_{SD}, R^2_R, and R^2_B, which are part of a program written in SPSS (Statistical Package for the Social

Sciences) v. 14.0 by the author (see Appendix A). The data sets also will be bootstrapped in the program AMOS v. 5.0 (Analysis of Moment Structures) for R^2_{BOOT}.

Results and Discussion

Table 1 shows that in terms of the relative accuracy of prediction, overall, the Browne algorithm, R^2_B, was superior in every data situation, except one. The R^2_B showed a distinct pattern of low bias and high stability and appeared to have the most relative accuracy for predictive effectiveness of criterion score estimation. Furthermore, following the standard set by Kromrey and Hines (1995), estimates within .01 of the sample R^2 value can be thought of as statistically unbiased, which occurred in two situations with R^2_B (i.e., the Thurstone and Hald data sets).

For the other three algorithm estimation techniques, none were noticeably superior in all nine of the data sets. That is, based on empirical data with varying sample sizes, regressor variates, and R^2 values, no generalizable rules can be constructed concerning which of the remaining three algorithm-based cross-validation methods were the "best" to use in a particular condition. However, patterns from the results do emerge to allow for some suggestions. For example, in the majority of data sets (i.e., 6 of 9), the Rozeboom algorithm (R^2_R) had the second least amount of proportional bias of the remaining formulas. In the other three data situations, this formula's prediction accuracy was the next most precise. Though research by Huberty and Mourad (1980) and Cotter and Raju (1982) studied the same three cross-validation formulas (e.g., R^2_{NL}; R^2_{SD}; R^2_R) in different ways, and came to some differing conclusions pertaining to predictive accuracy, they both concluded that $R_R{}^2$ was a precise estimator in most empirical data circumstances. Another apparent pattern from the current study's results was that after R^2_B and R^2_R, the Nicholson and Lord (R^2_{NL}) and the Stein-Darlington (R^2_{SD}) formulas demonstrated consistent patterns of low relative accuracy in many situations, with the most amount of proportional bias in 4 of 9 and in 6 of 9 data sets, respectively.

For the bootstrap method, R^2_{BOOT} showed no discernable pattern of relative accuracy with results ranging from the most accurate in a situation to the least accurate in three different data situations. Of interest is that the R^2_{BOOT} method had the least amount of proportional bias, in fact it was less than 0.01, when used with the Hald data set, which was the only data set with multicollinearity (e.g., variance inflation factor > 38 and tolerance < 0.03). This situation was checked with a data set independent from the others used in this study, which also manifested multicollinearity (e.g., variance inflation factor > 30 and tolerance < .02). In the scholarly literature, results from a study conducted by Ayabe (1985) using a technique similar to the bootstrap method, the jackknife procedure, found inferior estimates as well. Kromrey and Hines (1995) found mixed results, similar to the current study's findings, with use of the bootstrap method with small sample sizes. However, when sample sizes were $N \geq 100$, they found more unbiased estimates when using the bootstrap.

Conclusion

Given the very unique characteristics of each data set in this study in the areas of dissimilar N, p, and R^2 values, the R^2_B algorithm was the most accurate in 8 of the 9 data situations. None of the remaining three proposed cross-validation algorithms, or the bootstrap method, were exceedingly superior or inferior to each other when compared based on proportional bias. Although it may be convenient to run all four cross-validation methods from the program in Appendix A to determine which one has the least amount of bias given a specific data situation, the definite preference is toward R^2_B in nearly every

Table 1. Bias Affiliated with Cross-Validation Estimation Methods

Data Set	R^2	Biases	R^2_{NL}	R^2_{SD}	R^2_R	R^2_B	R^2_{BOOT}
Kerlinger (1973) $N = 12$ $p = 2$ DV = Attitude Score	0.799	R^2_{CV} Bias Proportional Bias	0.693 0.106 **0.133**	0.667 0.132 **0.165**	0.719 0.080 **0.100**	0.775 0.024 **0.030**	0.675 0.124 **0.155**
Sprinthall (2000) $N = 20$ $p = 2$ DV = Anger Score	0.255	R^2_{CV} Bias Proportional Bias	0.042 0.213 **0.835**	0.016 0.239 **0.937**	0.089 0.166 **0.651**	0.221 0.034 **0.133**	0.073 0.182 **0.714**
Agresti (1986) $N = 30$ $p = 2$ DV = Psychiatric Impairment	0.358	R^2_{CV} Bias Proportional Bias	0.241 0.117 **0.327**	0.233 0.125 **0.349**	0.266 0.092 **0.257**	0.336 0.022 **0.061**	0.215 0.143 **0.399**
Thurstone (1947) $N = 20$ $p = 3$ DV = Volume of a Box	0.943	R^2_{CV} Bias Proportional Bias	0.919 0.024 **0.025**	0.915 0.028 **0.03**	0.923 0.02 **0.021**	0.935 0.008 **0.008**	0.929 0.014 **0.015**
Kerlinger (1973) $N = 30$ $p = 3$ DV = GPA	0.617	R^2_{CV} Bias Proportional Bias	0.516 0.101 **0.164**	0.506 0.111 **0.18**	0.532 0.085 **0.138**	0.588 0.029 **0.047**	0.478 0.139 **0.225**
Rulon (1967) $N = 93$ $p = 3$ DV = Success Score	0.261	R^2_{CV} Bias Proportional Bias	0.203 0.058 **0.222**	0.202 0.059 **0.226**	0.212 0.049 **0.188**	0.246 0.015 **0.057**	0.167 0.094 **0.360**
Hald (1965) $N = 13$ $p = 4$ DV = Heat Evolved (Cement)	0.982	R^2_{CV} Bias Proportional Bias	0.963 0.019 **0.019**	0.954 0.028 **0.029**	0.966 0.016 **0.016**	0.974 0.008 **0.008**	0.975 0.007 **0.007**
Kerlinger (1973) $N = 30$ $p = 4$ DV = GPA	0.640	R^2_{CV} Bias Proportional Bias	0.513 0.127 **0.198**	0.497 0.143 **0.223**	0.529 0.111 **0.173**	0.599 0.041 **0.064**	0.508 0.132 **0.206**
Cohen (1983) $N = 36$ $p = 4$ DV = Religious Attitude	0.355	R^2_{CV} Bias Proportional Bias	0.171 0.184 **0.518**	0.152 0.203 **0.572**	0.194 0.161 **0.454**	0.303 0.052 **0.146**	0.222 0.133 **0.375**

empirical data cross-validation circumstance. Thus, it is probably prudent to apply R^2_B first while regarding the proportional bias derived from R^2_R as a comparison. The remaining two algorithms, R^2_{NL} and R^2_{SD}, did not perform well in virtually any data situation. Though use of the AMOS bootstrap technique it not difficult (cf. Fan, 2003 for application instructions), the mixed results derived from R^2_{BOOT} should afford caution when used with small sample sizes (i.e., $N < 100$), except in situations of multicollinearity where the R^2_{BOOT} method showed the least amount of proportional bias.

Appendix A. Cross-Validation Algorithms Program

```
**************************************************************************
Copyright David A. Walker, 2006
Contact dawalker@niu.edu
Northern Illinois University, 101J Gabel, DeKalb, IL 60115
        **APA 5th Edition Citation**
Walker, D. A. (2006). Four estimators for sample cross-validation [Computer
    program]. DeKalb, IL: Author.
**************************************************************************.


**************************************************************************
NOTE: Between BEGIN DATA and END DATA, insert the multiple correlation
coefficient (R2), the sample size (N), and the number of regressor variates (p)
derived from your data
**************************************************************************.

DATA LIST LIST / R2(F9.3) p(F8.0) N(F8.0).
BEGIN DATA
.799 2 12
.255 2 20
.358 2 30
.943 3 20
.617 3 30
.261 3 93
.982 4 13
.640 4 30
.355 4 36
END DATA.
COMPUTE RNICHOL = (N+p+1)/ (N-p-1).
COMPUTE RLORD = (N-1)/(N).
COMPUTE RNICLORD = (1-(RNICHOL*RLORD)* (1-R2)).
COMPUTE RSTEIN1 = (N-1)/ (N-p-1).
COMPUTE RSTEIN2 = (N-2)/ (N-p-2).
COMPUTE RDARLING = (N+1)/(N).
COMPUTE RSTDARL = (1-(RSTEIN1*RSTEIN2*RDARLING)* (1-R2)).
COMPUTE RROZE = (1- (N+p)/ (N-p) * (1-R2)).
COMPUTE RBROWNE1 = ((N-p-3) * (R2)**2)+R2.
COMPUTE RBROWNE2 = ((N-2*p-2) * R2)+p.
COMPUTE RBROWNE = RBROWNE1 / RBROWNE2.
EXECUTE.
FORMAT RNICHOL TO RBROWNE (F9.3).
VARIABLE LABELS R2 'Multiple Correlation Coefficient'/p 'Number of Predictor Variables'/ N 'Sample
Size'/RNICLORD 'Nicholson-Lord'/ RBROWNE 'Browne'/RSTDARL 'Stein-Darlington'/ RROZE
'Rozeboom'/.
REPORT FORMAT=LIST AUTOMATIC ALIGN (CENTER)
MARGINS (*,110)
  /VARIABLES=N p R2 RNICLORD RSTDARL RROZE RBROWNE
  /TITLE "Estimation of the Sample Cross-Validity Expectancy".
```

References

Aaron, B., Kromrey, J. D., & Ferron, J. M. (1998, November). *Equating r-based and d-based effect size indices: Problems with a commonly recommended formula*. Paper presented at the annual meeting of the Florida Educational Research Association, Orlando, FL.

Agresti, A., & Finlay, B. (1986). *Statistical methods for the social sciences* (2nd ed.). SanFrancisco: Dellan Publishing Company.

Allen, D. M. (1971). *The prediction sum of squares as a criterion for selecting predictor variables* (Tech. Rep. No. 23). Lexington, KY: University of Kentucky, Department of Statistics.

Ayabe, C. R. (1985). Multicrossvalidation and the jackknife in the estimation of shrinkage of the multiple coefficient of correlation. *Educational and Psychological Measurement, 45*, 445-451.

Browne, M. W. (1975). Predictive validity of a linear regression equation. *British Journal of Mathematical and Statistical Psychology, 28*, 79-87.

Cattin, P. (1980a). Estimation of the predictive power of a regression model. *Journal of Applied Psychology, 65*, 407-414.

Cattin, P. (1980b). Note on the estimation of the squared cross-validated multiple correlation of the regression model. *Psychological Bulletin, 87*, 63-65.

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement, 2*, 595-607.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cotter, K. L., & Raju, N. S. (1982). An evaluation of formula-based population squared cross-validity estimates and factor score estimates in prediction. *Educational and Psychological Measurement, 42*, 493-519.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin, 69*, 161-182.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvements on cross-validation. *Journal of the American Statistical Association, 78*, 316-331.

Fan, X. (2003). Using commonly available software for bootstrapping in both substantive and measurement analyses. *Educational and Psychological Measurement, 63*, 24-50.

Gollob, H. F. (1967, September). *Cross-validation using samples of size one*. Paper presented at the meeting of the American Psychological Association. Washington, D.C.

Gong, G. (2003). Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. In P. I. Good & J. W. Hardin (Eds.), *Common errors in statistics (and how to avoid them)* (pp. 173-186). Hoboken, NJ: John Wiley & Sons, Inc.

Hald, A. (1965). *Statisitcal theory with engineering applications* (6th ed.). New York: John Wiley & Sons, Inc.

Huberty, C. J., & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. *Educational and Psychological Measurement, 40*, 101-112.

Kerlinger, F. N., & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston, Inc.

Kromrey, J. D., & Hines, C. V. (1995). Use of empirical estimates of shrinkage in multiple regression: A caution. *Educational and Psychological Measurement, 55*, 901-925.

Krus, D. J., & Fuller, E. A. (1982). Computer assisted multicrossvalidation in regression analysis. *Educational and Psychological Measurement, 42*, 187-193.

Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics, 23*, 639-645.

Lord, F. M. (1950). *Efficiency of prediction when a regression equation from one sample is used in a new sample* (Research Bulletin No. 50-40). Princeton, NJ: Educational Testing Service.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental tests*. Reading, MA: Addison-Wesley.

Morris, J. D. (1984). Cross-validation with Gollob's estimator: A computational simplification. *Educational and Psychological Measurement, 44*, 151-154.

Morris, J. D. (1986). Microcomputer selection of a predictor weighting algorithm. *Multiple Linear Regression Viewpoints, 1*, 53-68.

Mosier, C. I. (1951). Problems and designs of crossvalidation. *Educational and Psychological Measurement, 11*, 5-11.

Nicholson, G. E. (1960). Prediction in future samples. In I. Olkin et al. (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlation: A clarification. *Psychological Bulletin, 85*, 1348-1351.

Rulon, P. J., Tiedeman, D. V., Tatsuoka, M. M., & Langmuir, C. R. (1967). *Multivariate statistics for personnel classification*. New York: John Wiley & Sons, Inc.

Sprinthall, R. C. (2000). *Basic statistical analysis* (6th ed.). Needham Heights, MA: Allyn and Bacon.

Stein, C. (1960). Multiple regression. In I. Olkin et al. (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: The University of Chicago Press.

Send correspondence to: David A. Walker
ETRA
Northern Illinois University
DeKab, IL 60115
Email: dawalker@niu.edu