JOURNAL OF EDUCATIONAL
AND INSTRUCTIONAL STUDIES IN THE WORLD
May 2019, Volume: 9 Issue: 2 ISSN: 2146-7463

CONDUCTING MULTIVARIATE ANALYSES BASED ON DATA SUMMARIES OF PUBLISHED RESEARCH

Prof. Dr. Mehmet Türegün
Mathematics Education
Barry University
Adrian Dominican School of Education
Curriculum Pedagogy & Research Unit
11300 NE Second Ave. Miami Shores, FL 33161- USA
MTuregun@barry.edu

Abstract

Most multivariate analyses use variance-covariance matrices and descriptive statistics, such as means and standard deviations, as their starting points. Inclusion of correlation matrices and descriptive statistics summaries in reporting results is recommended by journal editors and the American Psychological Association (APA) style guidelines. These descriptive data summaries in published research articles provide other researchers and graduate students in social science fields with opportunities to replicate or verify the results of the analyses without requiring access to the raw data. Developing transferable skills and increasing student-faculty collaboration make these types of analysis of published research ideal for use in classroom examples and research projects. The aim of this paper is to illustrate how to conduct various analyses, such as multiway frequency analysis, multivariate analysis of variance (MANOVA), multiple regression, and factor analysis via IBM SPSS syntax based on descriptive summary statistics reported in published research.

Keywords: Multivariate analysis, Quantitative research courses, Teaching, Replication, IBM SPSS syntax.

INTRODUCTION

The important role that statistics plays in graduate programs in social science fields, such as psychology and education, is evidenced by increasingly common requirements for students to complete quantitative research methods courses in which the list of topics invariably includes univariate and multivariate statistical analysis techniques (Davis, 2003; Perepiczka, Chandler, and Becerra, 2011; Sandals & Türegün, 2013). Generally, Analysis of Variance (ANOVA), MANOVA, discriminant analysis, multiple regression, path analysis, and factor analysis are commonly discussed topics in these graduate research methodology courses. The use of variance-covariance matrices, correlations and standard deviations may be considered as the computational starting points for these types of analyses. Editors for various peer-reviewed research journals in education, educational psychology, social sciences, and the American Psychological Association (APA) style guidelines recommend the inclusion of correlation matrices, means, and standard deviations as descriptive data summary tables in the published articles. Additionally, as stated by Zientek & Thomson (2009), reporting correlation/covariance matrices, standard deviations, and means in published research articles allows researchers opportunities to conduct secondary analyses. Hence, this paper is focused on illustrating how to conduct various analyses by using these descriptive summary statistics tables in the published research articles as the starting point.

Conducting analyses based on the descriptive data summaries from published research in the fields of social sciences and education can be a very useful tool in teaching multivariate analysis for graduate students. From a pedagogical point of view, the opportunities offered by the secondary analyses, such as replication of published results, conducting secondary analyses leading to publishable outcomes, and developing transferable skills in coding syntax make this types of analyses ideal for use in classroom examples, semester projects, capstone research, and supplemental studies (Rossi, 1987; Sautter, 2014). Additionally, the reviews and editors of peer-reviewed journals may benefit from such analyses in their efforts to replicate and verify the reported results of statistical analyses conducted in submitted manuscripts.

The aim of this paper is to illustrate how to conduct multivariate analyses from the data summaries in published research, and discuss the possible issues and potential benefits of this practice for novice researchers and graduate students. In the next section, I present a brief description of the setting where I used these types of analysis based on data summaries.

MODEL AND PROCEDURE

The college of education at the university where I teach requires students to complete quantitative research methods courses, and I teach various research methods courses, including a multivariate analysis course. The multivariate course, titled Advanced Quantitative Inquiry, is a semester-long, face-to-face, doctoral level, multivariate analysis course taught by a single instructor for students within a college of education. The topics discussed in the course include inferential statistics and data analysis techniques for educational research and practice. Among the specific statistical techniques listed in the course objectives are factorial ANOVA, one-way and factorial MANOVA, multivariate analysis of covariance (MANCOVA), multiple linear regression, logistic regression, path analysis, and factor analysis. Additionally, students are provided with a comprehensive knowledge of the IBM SPSS Statistics for Windows and other statistical aids in order to complete various course assignments and projects. In addition to being included in a master's level research methods course, the univariate and bi-variate topics, such as descriptive statistics, inferential statistics, t-tests, one-way ANOVA, and bi-variate correlation, are also discussed further in a doctoral level pre-requisite course for the multivariate course.

In order to provide students with hands-on methodological experiences and applications of various multivariate statistical concepts and techniques, the assessment of the students' performances in the course is based on several components consisting of a dissertation critique, a peer-reviewed article critique, a number of IBM SPSS assignments, and a final research project report. For the peer-reviewed article critique assignment, students conduct literature searches in order to locate published articles using the multivariate techniques discussed in the course. The articles using multiple regression analysis, factorial ANOVA, MANOVA, MANCOVA, factor analysis, and path analysis techniques are the most frequently chosen types of articles by the students.

Following the APA guidelines and the American Educational Research Association (AERA) standards for reporting statistical analyses have been considered by many as essential practices for the replication of results (MacCallum & Browne, 1993; Maxwell & Cole, 1995; Onwuegbuzie & Combs, 2009; Onwuegbuzie, Combs, Slate, and Frels, 2009; Sandals & Türegün, 2013; Thompson, 2007; Zientek & Thomson 2009). Furthermore, reporting the sample variance-covariance matrix, and the descriptive summary statistics, such as means and standard deviations, along with the sample sizes provides sufficient information to allow readers to replicate the authors' results for certain types of correlational analyses, such as multiple regression analysis, factor analysis, path analysis, and structural equation modeling (Cohen, 1968; Rossi, 1987; Zientek & Thomson, 2009).

The journal articles chosen for the article critique assignment are examined closely to verify that sufficient amount of information is reported to permit the replication of the published results. Raw data are rarely included in published research articles. However, following the APA guidelines and AERA standards for publications, the authors of the articles usually report appropriate descriptive summary data for other researchers to replicate the results via secondary analysis.

Illustrated Examples

In this section, I give four examples selected by students to critique for a course assignment. The sources from which the examples were taken ranged from peer-reviewed journal articles to dissertations. The examples include multiway frequency analysis, multiple regression analysis, MANOVA, and factor analysis.

Multiway Frequency Analysis (MFA) Example

Even though MFA is neither based on continuous variables nor uses a variance-covariance matrix as a starting point, it is an important technique for examining multiway relationships among variables. The General Loglinear Analysis (GLA) procedure, as an extension or part of MFA, is used to study the relationships among three or more categorical variables, and can be viewed as a multivariate extension of the Chi-square test. The GLA uses cell counts of the cross tabulation table formed by the cross-classification of the variables of interest in order to determine the least complex model that best explains the variance in the observed counts or frequencies. According to Tabachnick & Fidell (2013), the loglinear model building process starts with all the possible one-, two-, three-, and higher-way associations and eliminates as many of these associations as possible to arrive at a parsimonious model while still maintaining an adequate fit between expected and observed counts.

Foster, Barkus, and Yavorsky (2006) use data from a study examining the relationships among crime, substance use, and age. The data in aggregate form as a multiway contingency table are presented in Table 1. It was hypothesized that violent crime was related more to age than to substance use. The data consisted of three categorical variables, age, nature of crime, and substance use, for 178 participants from various treatment centers.

Table 1:Frequency distribution of crime type and substance use by age category, N=178

Age Category	Nature of crime	Observed Count Substance Yes	 No	Total	Expected Count Substance Yes	 No
Under 25	Violent crime	45	42	87	41.4	45.6
	Non-violent crime	15	24	39	8.4	13.5
Total		60	66	126		
Over 25	Violent crime	14	8	22	16.1	5.9
	Non-violent crime	24	6	30	21.9	8.1
Total		38	14	52		

Since MFA is a multivariate nonparametric statistical technique, there are no assumptions about the population distributions. However, there are several assumptions regarding the independence, adequate sample size, and the size of the expected counts. As depicted in Table 1, because the total number of observations, N, is equal to the number of cases, the independence is assured. According to Tabachnick & Fidell (2013), the adequate sample size should be at least five times as large as the number of cells in the design. The third limitation requires that all expected counts are greater than one, and no more than 20% of the expected counts are less than five. As illustrated in Table 1, all expected counts are greater than five.

With the three variables, age category, nature of crime, and substance use, there are seven possible associations. The only three-way association is *Age Category*Nature of Crime*Substance Use*. Additionally, there are three two-way associations: *Age Category*Nature of Crime, Age Category*Substance Use, and Nature of Crime*Substance Use*. The three one-way associations or effects are *Age Category, Nature of Crime, and Substance*. Based on Table 2, the test for the combined three two-way and one three-way associations showed statistical significance ($X^2(4)=24.523$, $p<.001$). Since the test for the single three-way association, provided in Table 2 when K=3, was not statistically significant ($X^2(1)=3.384$, $p>.05$), at least one of the three two-way associations was statistically significant. Hence, this particular three-way association could be eliminated towards obtaining a parsimonious model.

Table 2: MFA with K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson	
			Chi- Square	*p*	Chi-Square	*p*
K-way and Higher Order Effects	1	7	67.131	.000	67.483	.000
	2	4	24.523	.000	26.673	.000
	3	1	3.384	.066	3.420	.064
K-way Effects	1	3	42.607	.000	40.811	.000
	2	3	21.139	.000	23.253	.000
	3	1	3.384	.066	3.420	.064

Based on Table 3, the two-way association *Nature of Crime*Substance Use* was not statistically significant with *p* > .05, and could be eliminated. The two-way associations, *Age Category*Nature of Crime* and *Age Category*Substance Use*, were retained as statistically significant interactions with *p* < .05. The results, as presented in Table 3, replicated the results of the study by Foster, Barkus, and Yavorsky (2006), who demonstrated that age was significantly associated with violent crime and substance use.

Table 3: MFA with partial associations

Effect	df	Partial Chi-Square	*p*	Number of Iterations
AgeCat*ViolentCrime	1	11.071	.001	2
AgeCat*Substance	1	10.116	.001	2
ViolentCrime*Substance	1	.243	.622	2
AgeCat	1	31.718	.000	2
ViolentCrime	1	9.066	.003	2
Substance	1	1.823	.177	2

An example of IBM SPSS syntax command lines with the aggregate data from Table 1 as an input data list for conducting a MFA via SPSS HILOGLINEAR procedure is presented in Figure 1. The commands FREQ RESID produce the table of cell counts and residuals. The commands ASSOCIATION ESTIM produce test of all individual effects, combined individually effects with each other, and combined effects with each other and higher orders. The results produced by the execution of these command lines are presented in Table 2 and Table 3.

```
DATA LIST LIST / AgeCat CrimeNat Substance count.
BEGIN DATA
1 1 1 45
1 1 2 42
1 2 1 15
1 2 2 24
2 1 1 14
2 1 2 8
2 2 1 24
2 2 2 6
END DATA.
WEIGHT BY count.
HILOGLINEAR AgeCat(1 2) CrimeNat(1 2) Substance(1 2)
  /METHOD=BACKWARD
  /CRITERIA MAXSTEPS(10) P(.05) ITERATION(20) DELTA(.5)
  /PRINT=FREQ RESID ASSOCIATION ESTIM
  /DESIGN.
```

Figure 1. The IBM SPSS syntax command lines for conducting loglinear analysis by using SPSS HILOGLINEAR procedure

Multiple Regression Analysis (MRA) Example

In this example, a study focusing on the important and integral role that statistics play in research courses offered in graduate programs at schools of education was used to illustrate how MRA is implemented based on the summary statistics from published research studies. Perepiczka, Chandler, and Becerra (2011) investigated the nature and extent of the relationship among graduate students' *statistics self-efficacy*, *statistics anxiety*, *attitude towards statistics*, and *social support*. The data for these four variables were collected from a sample of 166 participants from various colleges of education across the United States. The descriptive statistics summary consisting of the means, standard deviations, and the Pearson product-moment correlations for the four variables is presented in Table 4.

Table 4: Means, standard deviations, and correlations for the variables

N=166	M	SD	Self-efficacy	Statistics Anxiety	Attitude towards Statistics	Social Support
Self-efficacy	49.73	18.97	1	-	-	-
Statistics Anxiety	119.95	35.83	-.679	1	-	-
Attitude towards Statistics	106.73	18.91	.708	-.832	1	-
Social Support	5.69	1.04	-.023	.006	.017	1

The data presented in Table 4 constituted the starting point in creating the IBM SPSS syntax commands given in Figure 2 to conduct a simultaneous multiple linear regression analysis. As can be seen from Table 4, *statistics anxiety* and *attitude towards statistics* were highly correlated (r=-.832). To combat multicollinearity, the problematic variable is either omitted from the analysis, or combined with the other variable(s) causing multicollinearity in order to create a single variable, especially for variables with correlation coefficients of .80 or higher (Stevens, 1992). However, since this strategy was not implemented by Perepiczka, Chandler, and Becerra (2011), it was not used here either to be able to compare the results.

Table 5: Model summary and ANOVA results

R	R-Square		Sum Squares	of df	Mean Square	*F*	*p*
.727	.528	Regression	31374.84	3	10458.28	60.5	.000
		Residual	28002.21	162	172.85		
		Total	59377.05	165			

A simultaneous multiple regression using the IBM SPSS syntax commands given in Figure 2 produced a statistically significant model to predict *self-efficacy* to learn statistics from *statistics anxiety*, *attitude towards statistics*, and *social support* (F(3, 162) = 60.5, p < .001, R^2=.528). As presented in Table 5, the three predictor variables combined together accounted for 52.8% of the variance in the *self-efficacy* to learn statistics.

As illustrated in Table 6, *statistics anxiety* was a statistically significant predictor of *self-efficacy* to learn statistics (t(162)= -2.983, p<.01). *Attitude towards statistics* was also a statistically significant predictor (t(162)= 4.797, p<.001). However, *social support* was not a statistically significant predictor of *self-efficacy* to learn statistics (t(162)= -.541, p>.05).

Table 6: MRA results with Self-efficacy to learn statistics as the dependent variable

	Unstandardized Coefficients		Standardized Coefficients	t	p
	B	Std. Error	Beta		
(Constant)	21.203	16.712		1.269	.206
StatAnxiety	-.154	.052	-.290	-2.983	.003
AttStat	.468	.098	.467	4.797	.000
SocSupp	-.533	.985	-.029	-.541	.589

The analysis based on the data summaries reported by Perepiczka, Chandler, and Becerra (2011) replicated their results and conclusion that *statistics anxiety* and *attitude towards statistics* were statistically significant predictors of *self-efficacy* to learn statistics, but not *social support*. The results of the analysis reported here using only summary statistics were aligned well with their results.

```
MATRIX DATA VARIABLES = ROWTYPE_ Selfeff StatAnxiety AttStat SocSupp
 /format = lower diagonal.
BEGIN DATA.
N 166 166 166 166
MEAN 49.73 119.95 106.73 5.69
STDDEV 18.97 35.83 18.91 1.04
CORR 1.00
CORR -0.679 1.00
CORR 0.708 -0.832 1.00
CORR -0.023 0.006 0.017 1.00
END DATA.
REGRESSION MATRIX = IN(*)
  /DEP = Selfeff
  /METHOD = ENTER StatAnxiety AttStat SocSupp.
```

Figure 2: The IBM SPSS syntax command lines for conducting multiple regression by using REGRESSION ENTER procedure

MANOVA Example

In this example, a dissertation examining the effects of various student characteristics, such as ethnicity, family educational history, and native language, on academic self-efficacy, faculty-student interactions, and students' self-reported cumulative grade point average was used to illustrate how to conduct a one-way MANOVA based on data summaries. Among various multivariate analyses, Gosnell (2013) conducted a one-way MANOVA to examine the effects of enrollment status, as full-time versus part-time, on students' self-reported grade point average, academic self-efficacy, and faculty-student interactions. Table 7 illustrates the descriptive statistics and the Pearson correlations for the variables.

Table 7: Means, standard deviations by full-time (FT) and part-time (PT) enrollment status, and correlations for Self-reported Grade Point Average (GPA), academic self-efficacy, and faculty-student interactions.

Variable	Enrolment Status FT M(SD)	Enrolment Status PT M(SD)	Self-reported GPA	Academic Self-efficacy	Faculty-student interactions
Self-report GPA	3.36(,436) n=92	3.18(.409) n=39	1	-	-
Academic Self-efficacy	7.50(1.60) n=75	7.47(1.63) n=31	.028	1	-

Faculty-student interactions	3.38(.54) n=88	3.07(.55) n=36	.123	.240	1

The descriptive summary statistics presented in Table 7 and the pooled standard deviations were used as a starting point in creating the IBM SPSS syntax given in Figure 3 to conduct a one-way MANOVA. The results, given in Table 8, were aligned with to the results reported by Gosnell (2013).

Table 8: Multivariate Tests of Significance

Test Name	Values	*F*	Hypoth. df	Error df	*p*
Pillai's Trace	.090	3.36	3.00	102.00	.022
Hotellings	.099	3.36	3.00	102.00	.022
Wilks' Lambda	.910	3.36	3.00	102.00	.022

There was a statistically significant effect of students' enrollment status on the linear combination of the variables, self-reported grade point average, academic self-efficacy, and faculty-student interactions, ($F(3,102)=3.36$, $p=.022$, $\eta^2=.090$). As is illustrated in Table 9, subsequently conducted univariate ANOVAs revealed that the only statistically significant effect was for faculty-student interactions ($F(1,104)=7.23$, $p=.008$, $\eta^2=.065$) at the Bonferroni adjusted alpha significance level of .0167.

Table 9: Univariate F-tests

Variable	Hypoth. SS	Error SS	Hypoth. MS	Error MS	*F*	*p*	ETA Square
Srgpa	.711	19.05	.711	.183	3.88	.052	.03596
AcadSelf	.020	296.58	.020	2.592	.01	.931	.00007
FacStInt	2.108	30.33	2.108	.292	7.23	.008	.06499

As reported by Gosnell (2013), the full-time students (M=3.38, SD=.54) had statistically significantly higher faculty-student interactions than the part-time students (M=3.07, SD=.55).

```
MATRIX DATA VARIABLES = EnrStatus ROWTYPE_ Srgpa AcadSeffeff FacStInt
   /factor EnrStatus.
BEGIN DATA.
1 MEAN 3.36 7.50 3.38
1 STDDEV 0.436 1.60 0.54
1 N 75 75 75
2 MEAN 3.18 7.47 3.06
2 STDDEV 0.409 1.63 0.55
2 N 31 31 31
. STDDEV 0.428 1.61 0.54
. CORR 1.00
. CORR 0.028 1.00
. CORR 0.123 0.240 1.00
END DATA.
MANOVA Srgpa AcadSeffeff FacStInt
    by EnrStatus(1,2) / MATRIX = in(*)
    /method=unique
    /print = descriptive cellinfo(all) signif(univ efsize)
    /design EnrStatus(1,2).
```

Figure 3: The IBM SPSS syntax command lines for conducting MANOVA by using descriptive data summaries as input.

Factor Analysis Example

In this example, a study focusing on the predictive ability of a set of institutional and student characteristics for retention rates of full-time, degree-seeking, first-time freshmen was used to illustrate the use of secondary analysis from data summaries. Scott, Velazquez, Türegün, and Wolman (2016) examined the relations among a set of predictor variables based on a sample obtained from the Integrated Post Secondary Education Data Systems (IPEDS) Data Center. Based on previous research studies, Scott et al. (2016) considered a total of sixteen institutional and student characteristics as variables from 233 institutions for a factor analysis in order to summarize the patterns of correlations among the observed variables.

The variables used to describe student characteristics for each institution were first semester average GPA, SAT 25th percentile score, percent of full-time students, percent of full-time, first-time, undergraduates receiving federal financial aid, percent of full-time, first-time undergraduates receiving Pell grants, percent of undergraduates over the age of 24, and percent of first-time freshmen receiving financial aid and living on campus. The variables used to describe institutional characteristics were grand total enrollment, student-to-faculty ratio, highest degree offered, degree of urbanization, academic support per full-time enrollment (FTE), net Instruction per FTE, selectivity, average net price for students receiving grants or scholarship aid, and percent of undergraduate FTE. The variables selectivity, degree of urbanization, and highest degree are based on the percentage of freshman applications that are accepted, the area where the institution operates and/or the geographical region where the institution is based, and the highest degree offered by the institution, respectively.

Scott et al. (2016) reported the descriptive statistics and the correlation coefficients, Pearson for continuous and Spearman for ordinal variables. The reported correlations were the starting point in creating the IBM SPSS syntax given in Figure 4 to conduct a factor analysis using Principal Axis Factoring (PAF). The analysis conducted by using the syntax given in Figure 4 replicated the results reported by Scott et al. (2016). The sampling adequacy of items was determined via the Kaiser-Meyer-Olkin (KMO) measure, and the Bartlett's sphericity test was used for appropriateness of conducting a factor analysis. The KMO statistic is a summary of how small the partial correlations are for each pair of the variables. If the variables share common factors then the partial correlations should be small and the KMO should be close to 1.0, indicating extraction of distinct and reliable factors. A value close to 0 indicates that the sum of the partial correlations is large relative to the sum of the correlations, which would result in factor analysis to be inappropriate.

```
MATRIX DATA VARIABLES=SAT GPA Select PerFT TotEnr AcadSup NetInst PerFed PerPell
PerO24 S2FRat MPSwAid HDeg PerUgrad Urb PerFinCam
 /N=233
 /CONTENTS=CORR.
BEGIN DATA.
1
0.646 1
-0.156 -0.133 1
0.360 0.174 -0.154 1
0.509 0.323 -0.067 0.164 1
0.347 0.160 -0.170 0.225 0.364 1
0.396 0.190 -0.178 0.260 0.285 0.733 1
-0.265 -0.134 0.111 -0.282 -0.319 -0.144 -0.129 1
-0.765 -0.567 -0.168 -0.316 -0.317 -0.187 -0.248 0.300 1
-0.542 -0.268 0.090 -0.767 -0.225 -0.204 -0.237 0.223 0.529 1
-0.120 -0.099 0.020 0.140 0.389 -0.221 -0.374 -0.306 0.153 -0.045 1
0.316 0.322 0.076 0.134 -0.135 0.132 0.246 0.236 -0.472 -0.296 -0.455 1
0.296 0.145 0.080 -0.077 0.509 0.305 0.246 0.044 -0.211 -0.094 0.004 0.056 1
-0.263 -0.244 0.062 0.298 -0.218 -0.396 -0.421 -0.242 0.149 -0.110 0.404 -0.309 -0.496 1
```

```
0.161 0.153 -0.028 -0.048 0.241 0.182 0.214 -0.032 0.011 0.060 -0.026 -0.052 0.138 -0.278 1
0.378 0.185 0.034 0.468 -0.004 0.208 0.212 0.034 -0.425 -0.519 -0.304 0.404 0.108 -0.087 -0.144 1
END DATA.
EXECUTE.
FACTOR
 /MATRIX IN(COR=*)
 /PRINT INITIAL CORRELATION SIG KMO EXTRACTION
 /FORMAT BLANK (.30)
 /PLOT EIGEN ROTATION
 /CRITERIA MINEIGEN(1) ITERATE(25)
 /EXTRACTION PAF
 /CRITERIA ITERATE(25)
 /ROTATION VARIMAX
 /METHOD=CORRELATION.
```

Figure 4: The IBM SPSS syntax command lines for conducting a factor analysis by using correlation matrix as input.

According to Kaiser (1974) and Cerny & Kaiser (1977), values greater than 0.5 are considered acceptable. Furthermore, values between 0.5 and 0.7 are mediocre, values between 0.7 and 0.8 are good, values between 0.8 and 0.9 are great, and values above 0.9 are considered superb. As presented in Table 10, the results of the extraction yielded a KMO statistic of .729, thus indicating the sampling adequacy of items was satisfied via the KMO measure.

Table 10: Kaiser-Meyer-Olkin measure and Bartlett's test of sphericity

Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy		.729
Bartlett's Test of Sphericity	Approx. Chi-Square	1935.06
	df	120
	p	.000

Bartlett's test of sphericity tests the null hypothesis that the original correlation matrix is an identity matrix, which indicates that all correlation coefficients are zero. Bartlett's test of sphericity produced a significant result ($\chi^2(120)=1935.06$, $p < .001$), verifying the appropriateness of factor analysis.

In factor analysis, communalities can be thought of as the squared multiple correlations for each of the variables that have been included in the analysis using the factors as independent variables and the variable as a dependent variable. It represents the proportion of variance of each variable that is explained by the factors. Initial communalities are the squared multiple correlation between a given variable and all other variables. The initial and extracted communalities for the variables are presented in Table 11.

Table 11: Commonalities for the variables

Variable	Initial	Extraction	Variable Name	Initial	Extraction
1 SAT	.790	.851	9 PerPell	.786	.806
2 GPA	.499	.519	10 PerO24	.717	.757
3 Select	.325	.131	11 S2FRat	.633	.712
4 PerFT	.726	.826	12 MPSwAid	.519	.552
5 TotEnr	.677	.819	13 HDeg	.467	.627
6 AcadSup	.588	.654	14 PerUgrad	.615	.718
7 NetInst	.664	.781	15 Urb	.201	.154
8 PerFed	.425	.381	16 PerFinCam	.485	.540

The communality values in the extraction column of Table 11 represent the proportion of the variance by the extracted factors. These values ranged from .851 to .131, suggesting that most of the variables are moderately, in some cases strongly, related to the set of factors, with the exception of the variables Selectivity and Degree of Urbanization. These two variables did not seem to be well represented in the common factor space.

Table 12 presents the eigenvalues obtained with a PAF, and the total variance explained. In determining the number of factors to be retained, traditionally and most commonly used practice is to use either Kaiser's eigenvalue rule or Cattell's scree test. Kaiser's eigenvalue rule is the default option in most statistics packages.

Table 12: Eigenvalues and total amount of variance explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.460	27.877	27.877	4.181	26.132	26.132
2	2.427	15.171	43.048	2.112	13.203	39.335
3	2.151	13.447	56.495	1.784	11.151	50.485
4	1.457	9.108	65.603	1.111	6.943	57.428
5	1.118	6.988	72.591	.641	4.006	61.435
6	.888	5.550	78.142			
7	.827	5.170	83.312			
8	.537	3.359	86.671			
9	.453	2.832	89.503			
10	.373	2.329	91.832			
11	.327	2.042	93.874			
12	.280	1.750	95.624			
13	.248	1.550	97.174			
14	.178	1.113	98.287			
15	.151	.943	99.231			
16	.123	.769	100.000			

Even though Kaiser's rule may be the most widely used decision rule for determining the number of factors to retain, Kaiser's rule has been shown to almost always severely overestimate the number of factors to retain (Zwick & Velicer, 1986). Despite its subjective nature in interpretation, Cattell's scree test has been shown to be much more accurate, but also tended to overestimate the number of factors. Applying Kaiser's rule to the eigenvalues presented in Table 12 suggests the presence of five factors, as the eigenvalues for these factors are greater than 1, with factors 5 and 6 being equally close to 1 from above and below, respectively. Factors 1, 2, 3, 4, and 5 explained 27.88%, 15.17%, 13.45%, 9.11%, and 6.99% of the variance, respectively, with a cumulative total variance of 72.59%.

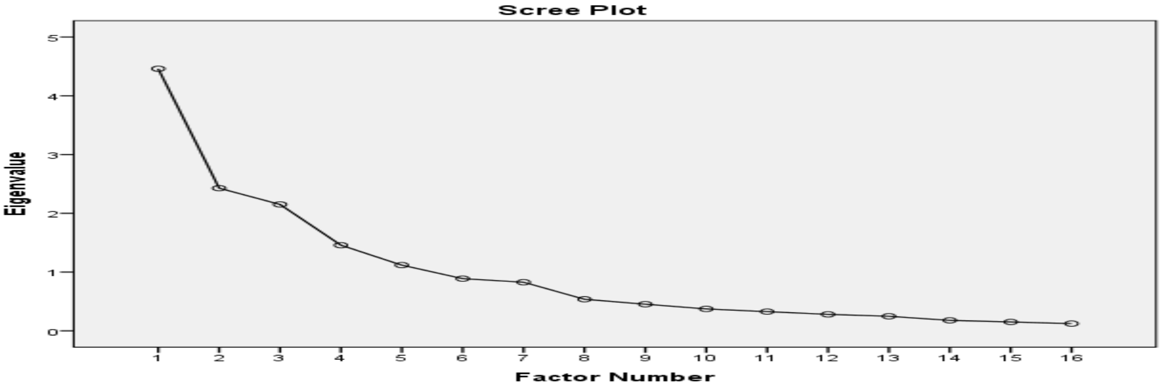

Figure 5: Scree plot

In addition to Kaiser's rule, the scree plot shown in Figure 5 was examined as a second criterion. The scree plot appears to indicate the presence of at least five factors and possibly up to six factors, as there is a slight drop after the fifth factor. As indicated previously, the use of scree plot in deciding on the number of factors to retain can be somewhat subjective, and tend to overestimate the number of factors to retain. Additionally, as illustrated in Table 13, the factor matrix revealed that there was only one item loading on Factor 5. Hence, a four-factor solution seemed reasonable, as suggested by Scott et al. (2016).

Table 13: Factor matrix

	Factor				
	1	2	3	4	5
1 SAT	.860				
2 GPA	.573			-.325	
3 Select					
4 PerFT	.512	-.619		.368	
5 TotEnr	.512		.721		
6 AcadSup	.562			.464	
7 NetInst	.623	.340		.505	
8 PerFed		.412	-.308		
9 PerPell	-.758			.416	
10 PerO24	-.650	.489			
11 S2FRat		-.570	.553		
12 MPSwAid	.444		-.523		
13 HDeg	.386	.322	.378		.454
14 PerUgrad	-.378	-.740			
15 Urb					
16 PerFinCam	.513		-.441		

Factor 1, with the highest loadings of SAT, PerPell, and PerO24, seemed to focus on student characteristics. With high loadings of PerUgrad, and PerFT, Factor 2 reflected traditional enrollment. Factor 3, with the highest loadings of TotEnr, S2FRat, and MPSwAid was interpreted as institutional affluence by Scott et al. (2016). Factor 4, with the highest loadings of NetInst and AcadSup can be interpreted as Institutional academic support.

CONCLUSIONS

There are many different purposes and benefits of replicating the results of various multivariate analyses via analyses based on published summary statistics, without relying on raw data. It may be an effective way for editors or reviewers to replicate the results of various multivariate analyses of submitted manuscripts to ensure and verify accuracy.

These types of analysis of published research provide graduate students with opportunities and experiences that may lead to developing transferable skills in coding syntax and statistical analysis. Integration of these pedagogical aspects into one's teaching approach makes these types of data summary-based analysess ideal for use in classroom examples, semester projects, and capstone research studies (Rossi, 1987; Sautter, 2014). Although IBM SPSS syntax was exclusively used in the examples given here, there are other data analysis packages to use, such as SAS, Stata, OpenStat, LISREL, and R, which provides an open-source option.

There are a several imitations of using data summary-based analyses. For example, insufficient data summaries and disregard for APA guidelines in reporting results prevent other researchers from replicating the results of published research. Descriptive data summaries reported in the published research can potentially be the starting point for the subsequent analyses to replicate the results. Therefore, making decisions about how to summarize the raw data becomes very important, and can further serve as a valuable teaching point.

Although most multivariate analyses can be performed based on descriptive data summaries and correlation matrices, the information contained in the raw data is hardly ever recovered from the descriptive data summaries reported in the published research. Additionally, since the analyses based on published research use the data summaries, instead of the original raw data sets, as the starting point, novice researchers or graduate students do not have an opportunity to gain experience in or practice data screening or preparation techniques, such as handling of missing data, or verifying assumptions.

REFERENCES

Cerny, C. A., & Kaiser, H. F. (1977). A study of a measure of sampling adequacy for factor-analytic correlation matrices. *Multivariate Behavioral Research*, *12*(1), 43-47.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426–433.

Davis, S. (2003). Statistics anxiety among female African American graduate-level social work students, *Journal of Teaching in Social Work*, *23*, 143-158.

Foster, J., Barkus, E., & Yavorsky, C. (2006). *Understanding and using advanced statistics*. Thousand Oaks, CA. Sage Publications, Inc.

Gosnell, J. (2013). *Academic self-efficacy, faculty-student interactions, and student characteristics as predictors of grade point average.* Unpublished doctoral dissertation. Barry University.

Kaiser, H. F. (1974). An index of factorial simplicity. Pscychometrika, 39, 31-36.

MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin, 114*(3), 533-541.

Maxwell, S. E. & Cole, D. A. (1995). Tips for writing (and reading) methodological articles. *Psychological Bulletin, 118*(2), 193-198.

Onwuegbuzie,A. J., & Combs, J. P. (2009). Writing with discipline: A call for avoiding APA style guide errors in manuscript preparation. *School Leadership Review, 4*, 116-149.

Onwuegbuzie, A. J., Combs, J. P., Slate, J. R., & Frels, R. K. (2009). Evidence-based guidelines for avoiding the most common APA errors in journal article submissions. *Research in the Schools, 16*(2), ix-xxxvi.

Perepiczka, M., Chandler, N., & Becerra, M. (2011). Relationship between graduate students' statistics self-efficacy, statistics anxiety, attitude toward statistics, and social support. *The Professional Counselor, 1*(2), 99–108.

Rossi, J. (1987). How often are our statistics wrong? A statistics class exercise. *Teaching of Psychology, 14*(2), 98-101.

Sandals, L. & Türegün, M. (2013). *A model for teaching research methodology for graduate students in social sciences and education.* Paper presented at the 20th International Conference on Learning. University of the Aegean, Rhodes, Greece.

Sautter, J. (2014). Secondary analysis of existing data in social science capstone research. *Council on Undergraduate Research*, *34*(4), 24-30.

Scott, V., Velazquez, M., Türegün, M., & Wolman, C. (2016). *Predictors of college and university retention rates: A path analytic model*. Paper presented at Phi Delta Kappa (PDK) International Biennial Research Symposium. Barry University, Miami Shores, FL.

Stevens, J. P. (1992). *Applied multivariate statistics for the social sciences* (2nd edition). Hillsdale, NJ: Erlbaum.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th edition). Pearson Education, Inc.

Thomson, B. (2007). *Standards in conducting and publishing research in education.* Paper presented at annual meeting of the Mid-Western Educational Research Association (MWERA). St. Louis, MO.

Zientek, L. R., & Thompson, B. (2009). Matrix summaries improve research reports: Secondary analyses using published literature. *Educational Researcher, 38*(5), 343–352.

Zwick, W., & Velicer, W. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 452-442.