DOCUMENT RESUME

ED 359 209 TM 019 835

AUTHOR Olejnik, Stephen; Huberty, Carl J.
TITLE Preliminary Statistical Tests.
PUB DATE Apr 93
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Covariance; Analysis of Variance; *Goodness of Fit; Literature Reviews; *Models; Regression (Statistics); *Research Methodology; Scholarly Journals; *Statistics; *Test Use
IDENTIFIERS Exploratory Data Analysis; F Test; *Omnibus Tests; Power (Statistics); *Preliminary Tests; Repeated Measures Design; Sphericity Tests; Variance (Statistical); Violation of Assumptions

ABSTRACT

Applied researchers frequently precede analyses of interest with one or more preliminary tests, used to help researchers determine which variables to examine more closely, or whether there are anomalies in the data set. These texts can be classified into three categories: omnibus tests, tests for model fit, and exploratory tests. Fifty-four applied journal articles, 31 statistical textbooks, and the statistical literature are reviewed to discuss some limitations associated with uses of some preliminary tests. The focus is on the following topics: (1) the omnibus F-test in analysis of variance; (2) tests for variance equality; (3) tests for equality of regression slopes in analysis of covariance; and (4) tests for sphericity in repeated measures designs. In general, it is concluded that many preliminary statistical tests are not useful. In many contexts, omnibus tests do not answer questions of substantive interest. Preliminary analyses in tests for model fit are unnecessary because alternative less restrictive models can be used, and because many tests for violations of data assumptions lack adequate statistical power, or are overly sensitive to another assumption violation. More focused analyses are advocated, using less restrictive analytical models, and an increased use of exploratory analyses is recommended. One table illustrates the discussion. Lists of the journal articles and textbooks reviewed are appended. (Contains 43 references.) (SLD)

Preliminary Statistical Tests

Stephen Olejnik

Carl J Huberty

University of Georgia

Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA April 1993.

ABSTRACT

Applied researchers frequently precede analyses of interest with one or more preliminary tests. These tests can be classified into three categories: 1) omnibus tests; 2) tests for model fit; and 3) exploratory tests. The present paper reviews a sample of applied journal articles, statistical textbooks, and the statistical literature to discuss some limitations associated with uses of some preliminary tests. In general it is concluded that many preliminary tests are not useful. We advocate more focused analyses using less restrictive statistical models, and recommend an increased use of exploratory analyses.

Preliminary Statistical Tests

Researchers interested in answering substantive questions with specific analyses often precede their analyses with one or more preliminary statistical tests. These preliminary tests can be classified into one of three categories: a) omnibus tests; b) tests for model fit; and c) exploratory analyses.

An omnibus test, that is the simultaneous test of several hypotheses in a single analysis, is frequently examined before individual hypotheses are tested. A purpose of this test often given is to limit the risk of making a Type I error across multiple follow-up tests. Application of this preliminary test is very popular and examples can be found throughout the applied research literature; this test dominates statistical methods textbooks. The simultaneous comparison of several population means through the analysis of variance F-test is a good example of an omnibus test that is often conducted before specific hypotheses are tested through contrast analyses. Other examples include the test of the full model in multiple regression analysis before specific coefficients are tested, and a multivariate test for the simultaneous equality of several outcome measures before a series of univariate tests is conducted.

Each statistical test is based on a mathematical model that has been formulated assuming a specific data structure. Preliminary statistical tests can be completed in an effort to

determine whether there is sufficient evidence to support the conclusion that the observed data do not fit the assumed model. For example, the t-test for comparing two independent population means is based on the assumptions that the outcome variable measures are independent and the population distributions are normal with equal variance. A violation of any of these assumptions can invalidate the probability statements made in drawing conclusions from the analysis (Glass, Peckham, & Sanders, 1972). Heterogeneous population variances can, for example, result in a risk of a Type I error greater than the nominal level. The Hartley F-max and the Bartlett chi-squared tests are well known procedures that might be used to determine whether populations have equal variances. Other examples of model fit include a preliminary test on higher order interaction effects to justify a main effects model, and a test for linearity before accepting a linear regression model.

As an exploratory analytic technique, preliminary tests are used to help researchers determine which variables to examine more closely or to determine whether there are anomalies in the data set. Examples might include the use of variable selection procedures to identify subsets of variables to include in a model, or the use of factor or component analysis to reduce the number of predictor variables to be considered in a regression model. The examination of data sets for outliers by using the Cook distance statistic is still another example where a preliminary analysis is conducted before specific hypotheses are

tested.

The routine application of preliminary tests is often recommended by instructors and textbook writers. The purpose of the present paper is to discuss the limitations associated with uses of some preliminary tests. Our discussion is based on a review of a convenience sample of statistical methods textbooks, a review of some published educational and psychological research articles, and a review of the statistical literature found by searching through the Current Index to Statistics (CIS) from 1986 to the present. In the statistical literature, preliminary tests are occasionally referred to as "tests under conditional specifications."

The review of journal articles was limited to either articles published in the Journal of Experimental Education Volumes 58 (1990-91) and 59 (1991-92), or empirical research published since 1979 on the effectiveness of study strategies with post secondary students. The Journal of Experimental Education was chosen because we felt that articles published in this journal are fairly representative of competent quantitative research inquiries conducted by behavioral science researchers. Studies on the effectiveness of study strategies were chosen because of a research interest of the first author and because investigations in this area can be found in a wide variety of behavioral science journals. Bibliographic information on 54 articles reviewed is given in Appendix A.

Twenty-five introductory and six intermediate statistical

methods textbooks were selected for review - - all but six of the 31 books were published in 1990 or later. We chose these books because we are either currently using them, have used them recently, or have considered them as a primary text for the statistical methods classes that we teach. A complete listing of the books is presented in Appendix B.

While there are many preliminary tests that can be and are conducted by researchers, the focus of the present paper is on four tests: the omnibus F-test in analysis of variance, test for variance equality, test for equality of regression slopes in analysis of covariance, and tests for sphericity in repeated measures designs. These tests were chosen because we felt that they are well known by applied researchers, because they are frequently included in statistical methods textbooks, and because these tests are often considered in conjunction with the research designs frequently used by applied researchers in the behavioral sciences.

Omnibus Tests

ANOVA F-test

Part of the rationale for the development of the analysis of variance (ANOVA) F-test was to allow researchers to compare several population means simultaneously. Multiple two-group t-tests for pairwise comparisons have been discouraged by some (e.g., Stevens, 1990, p. 32) because the overall probability of at least one Type I error can be quite large depending on the number of tests. [The probability of at least one Type I error

among k tests is no more than $1-(1-\alpha)^k$, where α is the probability of committing a Type I error with each of the k tests.] The most important limitation of the omnibus F-test is that it is so general that it typically does not address an interesting substantive question. The rejection of the null hypothesis simply means that there is sufficient evidence to conclude that the populations from which the samples were selected do not have identical means. This is an answer to a question that is rarely, if ever, of interest to the applied researcher. Answers to substantitive questions of typical interest to applied researchers require specific contrast analyses. In our review of 54 published articles, 31 of the studies involved an explanatory variable with more than two levels, and the analysis for each of these 31 studies began with an omnibus F-test. Several of these articles had explicitly stated research questions that would be answered appropriately with a specific pairwise or complex contrast. For example, in one study the researchers wrote "the first question asked whether students in mapping treatments, A and B, would score significantly higher on holistic scores.... than in the nonmapping, group C treatment". The researchers incorrectly based the answer to this question on the omnibus F-test.

Concern for an inflated Type I error rate may be over-emphasized by instructors and textbook authors. This concern is not shared by all statisticians (e.g., Saville, 1990). The overall risk of a Type I error can be controlled by using one of

the many Bonferroni-type adjustments (Li, Olejnik, & Huberty, 1992). In our review of the 31 journal articles, none of the researchers stopped their analyses following the rejection of the ANOVA F-test. Each researcher continued by either discussing the specific group means or by further hypothesis testing with a contrast procedure. The most popular contrast tests were the Scheffe test and the Newman-Keuls test. We did not find a single reference to a Bonferroni-type adjustment.

Textbooks generally emphasize the use of the ANOVA F-test followed by one of several contrast analyses. Many discussions of contrast procedures referred to post hoc techniques. Fourteen of the introductory texts took this approach. Authors often mislead readers to believe that contrast procedures can only be used after the omnibus hypothesis test has been rejected. While some procedures do require the omnibus test, most do not. We did not find a single incident where the procedure developed by Tukey took precedence ofer testing the omnibus hypothesis.

Among the six intermediate statistical methods textbooks reviewed, only the Maxwell and Delaney (1990) text suggests that contrasts can be tested instead of conducting the omnibus test. The five remaining texts do state that if planned contrasts are examined the omnibus test is unnecessary, but they also indicate that if post hoc procedures are of interest the omnibus test must be conducted first.

We recognize, however, that there are situations where an omnibus F-test can be useful. One such situation involves the

test for an interaction in a factorial design. A test for the interaction may precede contrast analyses by guiding the data analyst to construct the contrasts using cell rather than marginal means. It might be noted that Tukey (1991) has suggested contrasts that involve cell effects after removing main and interaction effects may be of interest; thus precluding the test for an interaction.

Another context where the preliminary omnibus F-test may be useful is when all pairwise contrasts are of interest to the researcher. Hayter (1986) demonstrated that the omnibus F-test can be used in conjunction with a contrast analysis procedure to enhance the statistical power. Alternatively, Shaffer (1986) recommends the omnibus F-test as a preliminary test to a sequential multiple range contrast procedure. Seaman, Levin, and Serlin (1991) studied these approaches and concluded that they both can be useful. If all pairwise contrasts are not of interest, neither the Hayter nor the Shaffer procedures would be used.

Other Omnibus Tests

While some writers (e.g., Maxwell & Delaney, 1990, p. 200; Toothaker, 1991, p. 55; Tukey, 1992) have encouraged researchers to ignore the overall test of equal means if it does not pertain to a substantive question of interest such a recommendation is virtually unheard of when it comes to testing the equality of proportions. Three situations in which an omnibus test might be bypassed are: (1) two groups, polytomous outcome; (2) multiple

groups, dichotomous outcome; and (3) multiple groups, polytomous outcome. In each of these situations, questions more specific than omnibus questions would typically be of greater substantive interest. Notationwise, let p_{ij} denote the true proportion of experimental units in Group j who were expected to respond according to Category i with respect to the categorical outcome variable. For situation (1), a specific null hypothesis might be H_o: $p_{31} - p_{32} = 0$; for situation (2), H_o: $p_{11} - p_{14} = 0$; and for situation (3), H_o: $2p_{32} - p_{33} - p_{34}=0$. The omnibus test for any of these three situations typically pertains to the independence of the grouping variable and the (categorical) outcome variable. [The statistic used is the so-called "Pearson chi-square(d)" statistic.] Rejection of the null hypothesis is not seen as being too substantively informative in most situations; tests of more specific hypotheses would, in many cases, be more informative.

Another omnibus test that yields very limited substantive information is that for a multivariate analysis of variance (MANOVA) conducted prior to multiple ANOVAs. It is often implicitly or explicitly argued that a MANOVA rejection gives the researcher a "license" to proceed to the use of multiple ANOVAs. This rationale has been rebuked by Huberty and Morris (1989).

There is another preliminary null hypothesis in the multiple response variable arena that has been advocated by some writers. This test involves a true correlation matrix, $\tilde{R}$. [In the SPSS MANOVA procedure, this test is called "Bartlett's test of

sphericity."] Testing H_o: $\tilde{R} = I$ (the identity matrix) makes sense to us in one context, but not in another. The sensible context is that of factor analyzing the sample correlation matrix, R. As McDonald (1984, p. 24) points out, "It is the obvious test to use as a general protection against foolish optimism when hunting for relations in a mass of data." Considering H_o: $\tilde{R} = I$ as an hypothesis prior to examining individual bivariate correlations for statistical significance (e.g., Steiger, 1980) is not judged to be defensible. To us, this is analogous to employing MANOVA prior to multiple ANOVAs.

Finally, an omnibus preliminary test that is suggested by some methodologists (e.g. Cliff, 1987, p. 431) is to conduct a canonical correlation analysis (CCA) and, if "significant results' are obtained, conduct multiple multiple correlation/regression analyses. If the discovery cf canonical variates is not of substantive interest, then conducting a CCA is judged irrelevant.

Test for Model Fit

Variance Equality

Statistical methodologists have studied extensively the effect of variance heterogeneity on the validity of the ANOVA F-test and the two-group t-test for means. The results of these studies consistently show that the violation of the equal variance assumption can result in an increased risk of a Type I error when population variances are negatively related to sample sizes (Glass, Peckham, & Sanders, 1972; Milligan, Wong, &

Thompson, 1987; Tomarken & Serlin, 1986). Even when there is no relationship between sample size and variance, an increased risk of a Type I error can occur if there are substantial differences in the variances (Wilcox, Charlin, & Thompson, 1986). Because of these results, researchers might be expected to examine the sample variances to try to determine whether there is evidence to indicate that the assumption has been violated. Several statistical tests have been developed to compare variances, including the Hartley F-max test, the Cochran test, and the Bartlett test. Since the violation of the assumption can threaten the statistical validity of the test of means, it might seem to be an important consideration to be addressed by textbook authors of introductory and intermediate texts. In our review, only the textbook by Popham and Sirotnik (1992) recommends the preliminary test in the two group case and only Heiman (1992) recommends the test in the multiple group situation. Most of the texts mention the assumption but generally ignore the problem with respect to testing hypotheses on the means. Among the intermediate textbooks only Stevens (1990) and Keppel (1991) suggest formal testing for variance equality.

In our review of the journal articles only two articles commented on the apparent inequality of the population variances and neither used a statistical procedure in a formal test of the assumption.

Although variance inequality is a serious threat to the statistical validity of tests for mean equality, we do not

believe that a test for the violation of the assumption is warranted or advised. Several tests have been developed to test the equality of population variances but most of these tests are sensitive to non-normality (Conover, Johnson, & Johnson, 1981). Tests for variance homogeneity that are robust to non-normality (Brown & Forsythe, 1974; O'Brien, 1978; Ramsey & Brailsford, 1990) are not sufficiently sensitive to a violation of the assumption (Olejnik, 1987; Wilcox et al 1986). More importantly, alternatives to the traditional ANOVA F-test and two-group t-test are available. Specifically, the Welch solution to the Behrens-Fisher problem is available on the SAS T-Test procedure and the James second-order test is available for analysis of variance (Oshima & Algina, 1992). Moser and Stevens (1992) also recommend that the Welch alternative in the two group case when variances are unknown. Although these tests are approximate tests, they do limit the risk of a Type I error below the nominal level. In terms of statistical power they are only slightly less powerful than the independent samples t-test or the ANOVA F-test. The textbook by Moore and McCabe (1989) introduces the Welch procedure in discussing the test for two independent means; but fails to continue with this position in the multiple-group situation. Contrast procedures are also available which do not require equal variances (Dunnett, 1980; Games & Howell, 1976; Hsuing & Olejnik, 1991).

Test for Equal Slopes

A fairly common statistical procedure in the behavioral

sciences is analysis of covariance. The primary purpose of this technique is to reduce error variance in an experimental study and to attempt to equate comparison groups in a non-experimental study (Porter & Roundenbush, 1987). Both purposes are achieved by considering the relationship between the covariate and the outcome variable, which is assumed to be the same for all populations being compared. When the relationship between outcome variable and the covariate differs as a function of levels of the grouping variable, there is an interaction between the covariate and the grouping variable. If there is an interaction, then the interpretation of the main effect for the grouping variable can be ambiguous. Consequently, as with the factorial ANOVA, the test of the interaction generally precedes the test of the grouping variable main effect.

Three of the introductory statistical methods texts that we reviewed commented on this test and only two of them (Glass & Hopkins, 1984; Hays, 1988) provide sufficient information to compute the test statistic. All six of the intermediate texts we examined recommend that the preliminary test be conducted.

Nine of the journal articles we reviewed used analysis of covariance. Only two of the articles commented on examining the within-groups regression slopes, and neither of these studies rejected the equal slopes hypothesis.

We believe that the test for the equality of regression slopes is useful but cannot be relied upon except for situations where the violation of the assumption is extreme. Rogosa (1980)

pointed out that the preliminary test for slope equality had both statistical and logical limitations. If the sample size is large trivial differences in slopes will lead to the rejection of the model while small sample sizes can result in accepting an inappropriate model. Furthermore, small differences in the slopes may not invalidate the conclusion regarding group differences on the response measure.

The issue of statistical power can be evaluated by estimating the necessary sample size needed to detect the interaction. Using the Cohen (1988) power analysis tables, the sample sizes needed to test the equality of two regression slopes were determined for three levels of statistical power. Table 1 summarizes our results using the Cohen definition for a small, medium, and large difference between two population standardized regression coefficients. Examples of a pair of standardized regression coefficients reflecting a small difference is .60, .66 or .10, -.10; a pair of coefficients of .60, .76 or .10, -.20 reflects a medium difference; and a large difference is reflected by pairs of coefficients equalling .60, .83 or .10, -.40.

Insert Table 1 here

Because a Type II error would be more serious in this case, the statistical power should be set no less than .9 and the Type I error rate may be relaxed to equal .20. From Table 1 a sample size of at least 56 units from each population would be needed to

detect a large difference between two regression slopes. Smaller differences in slopes would require many more units.

In our review of the nine studies using analysis of covariance only one study had a sample size meeting this requirement. Typically, the sample size per treatment condition approximately equalled 30 units. This is not very surprising. If a researcher had planned the research study to test the hypothesis that two population means were equal and set the Type I error rate to equal to .05 and the Type II error rate to equal .20, then using the Cohen definition of a medium effect size as the criterion of practical significance and assuming a correlation of .70 between the covariate and the outcome, the researcher would find (using the Cohen power tables) that a sample size of 32 was sufficient to test the hypothesis the two populations have equal means on an outcome of interest. However with this sample size, the power to test the equality of slopes would equal .73 for a large difference in slopes and less than .50 for medium and small differences in slopes when the test of equal slopes has a Type I error rate of .20.

These examples demonstrate a serious problem with the preliminary test for the equality of regression slopes in analysis of covariance. That is, the sample size necessary to test the equality of means is considerably less than the sample size needed to test the equality of the within group regression slopes.

We believe that researchers should not rely on the

preliminary test for guidance on the correct statistical model. As an alternative, the data should be analyzed allowing the slopes to differ as Rogosa (1980) suggested (also see Maxwell & Delaney, 1990, pp. 406-420). Specific hypotheses can then be tested through simultaneous contrast analyses for all relevant levels of the covariate. If the slopes are in fact equal and this analysis is conducted, some statistical power will be lost. Chou (1991) showed, however, that the reduction in statistical power was small when the slopes were equal.

Test for Sphericity

Educational research studies often involve the repeated measurement of experimental units. In our review of the journal articles we found 19 studies that used a repeated measures design. Seven of these studies included at least three measures on each subject and 12 studies included only two measures. The statistical model for the design, with more than two measures assumes that the variance of the difference scores between measures are equal. This assumption is known as the sphericity assumption. When the assumption is violated, the univariate hypothesis test will have a Type I error rate that exceeds the nominal significance level. If the assumption is violated, a multivariate test, which does not assume sphericity, can be used. When the sphericity assumption is met, the univariate test is more powerful than the multivariate test. Consequently, a test for sphericity might seem as appropriate to determine whether a univariate or a multivariate test should be conducted. Several

alternatives have been suggested.

In our review of the textbooks none of the introductory textbooks discussed the repeated measures design. All six of the intermediate texts discuss repeated measures designs and only the Lomax (1992) text did not discuss the sphericity issue at all. Winer et al (1991) and Kirk (1982) discuss a test for sphericity but do not recommend it, and the remaining texts do not recommend the tests but suggest using the univariate approach with an adjusted degrees-of-freedom value for the computed test statistic to create a conservative test.

Seven of the 19 studies in our sample of journal articles involved a within subjects factor having three of more levels. None of these studies commented on the sphericity assumption and none of them used an adjusted degrees of freedom test or a multivariate test.

The test for sphericity has received considerable attention in the statistical literature. Robey and Barcikowski (1987) discuss and review five tests for sphericity, and recommend that researchers forego any of the tests since they are all sensitive to nonnormality. Looney and Stanley (1989) discourage the test for sphericity in favor of using both the univariate test and the multivariate test each using a reduced statistical criterion ($\alpha/2$). On the other hand, Cornell, Young, Seaman, and Kirk (1992) have recently examined the statistical power for eight tests for sphericity and concluded that these tests are sensitive to violations of the sphericity assumption when population

distributions are normal. They recommend a preliminary test for normality in addition to the preliminary test for sphericity.

We agree with the textbook authors who discourage the use of preliminary tests for sphericity. Our position is based on two considerations. First, the tests for sphericity are sensitive to the assumption of multivariate normality. Micceri (1990) has provided convincing evidence to indicate that many variables studied by behavioral researchers are not normally distributed. There is ample reason therefore to doubt the validity of these tests for sphericity with data from the behavioral sciences.

Our second reason for rejecting the sphericity test is our belief that omnibus tests are generally not needed and that contrast analyses are more appropriate. The sphericity assumption is necessary only for the omnibus test involving the repeated measures factor. As we pointed out earlier the omnibus test is too general to be of interest to most serious researchers. Consequently, multiple related samples t-tests as suggested by Maxwell (1980) seem more appropriate to us (also see Toothaker, 1991, p. 134). Control for an inflated Type I error rate can be provided through a Bonferroni-type adjustment (Holland, 1991; Keselman, Keselman, & Shaffer, 1991).

Other Tests for Model Fit

The recommendation that measures on response variables be examined for fit with theoretical normal distributions is sometimes suggested by textbook authors. Although there are formal statistical tests for normality, the suggestion often

given is the use of an "eyeball test" that would be made via a quantile-quantile plot (Moore & McCabe, 1989, p. 65) or a residual plot (Neter, Wasserman, & Whitmore, 1988, p. 734). Residual plots may also be employed in assessing linearity in correlation/regression analyses. In the special bivariate situation, scatter-plots should be routinely generated prior to the calculation and interpretation of a Pearson correlation coefficient.

Finally, there is the problematic test of the equality of population correlation matrices, a test often considered prior to a MANOVA. This may also be considered a "test for linearity" in the context of predictive discriminant analysis. This test is problematic because of its extensive statistical power for samples of respectable sizes and because of its reliance on the troublesome condition of multivariate normality.

Exploratory Tests

A third type of preliminary test might be classified as an exploratory test. Such a test arises in two situations. First, when researchers do not have a strong theoretical model to drive their data collection and analysis; and second, when the population being studied is not well understood or clearly defined. When new inquiries are made into some phenomenon that does not have a theoretical base, the constructs or the indicators of the constructs under investigation are often not well understood. Consequently, researchers sometimes take a "shotgun" approach to data collection. Rather than a focused

inquiry using a limited number of construct indicators, multiple indicators are often used and preliminary tests are conducted to better understand the interrelationships among the indicators and possibly to reduce the number of indicators used in the primary analysis to answer the research question. So, for example, a preliminary analysis using principal components or factor analysis might be used to reduce the number of indicators used in a primary analysis. Also, measures for multicolinearity might be examined prior to a multiple regression analysis to reduce the redundancy among the explanatory indicators.

When the population being studied is not clearly defined or understood, preliminary analyses might be carried out to gain some insight into the characteristics of the subjects studied. Cluster analysis might be carried out to group the units into more homogeneous subgroups. The examination of outliers using the Cook (1977) distance statistic can be used to identify experimental units which do not belong with the others in the data set (e.g. Bollen & Jackman, 1990).

Such preliminary tests can at times be very useful to a data analyst, and researchers should be encouraged to use them. In our review of the journal articles we did not come across a single example where exploratory tests were conducted. We were somewhat surprised and disappointed that none of the authors commented on any effort to identify outliers. We can only assume that such analyses were not conducted. The textbooks we examined generally do not encourage researchers to conduct exploratory

analyses. The intermediate textbooks do comment on tests for outliers but we feel that these tests should be given greater and earlier emphasis.

Conclusion

In general, it is our position that many preliminary statistical tests are not necessary. As we discussed above, in many contexts omnibus tests do not answer questions of substantive interest to the applied researcher. In terms of tests for model fit, preliminary analyses are unnecessary because alternative less restrictive models can be used, and because many tests for violations of data assumptions either lack adequate statistical power or are overly sensitive to another assumption violation. Consequently, these preliminary tests are often uninformative at best and can seriously mislead the data analyst. As an exploratory technique, however, preliminary analyses can be extremely useful; these analyses typically do not rely on statistical criteria.

In general we advocate that researchers <u>think</u> about their research problem to identify the specific questions to answer and to test only those hypotheses of interest. Most of the studies that we reviewed had specific questions in mind when they planned their research; unfortunately the researchers felt compelled (probably because of tradition and training) to use an analysis strategy that did not address the question at hand.

We think that researchers should become more knowledgeable about the population they are studying and about the constructs

under investigation. Based on this knowledge, researchers should identify the appropriate test statistic that is valid for the situation.

We also believe that alternative analysis techniques which minimize model assumptions should be encouraged and used. Appropriate less restrictive models are available and the only serious consequence of using these models when more restrictive models can be used may be a loss of statistical power. Additional research in this area is needed but there is some evidence to indicate that in some contexts the loss in statistical power is not great.

Finally, exploratory analyses which contribute to a greater understanding of the constructs and populations under investigation should be encouraged and expected as a routine segment of the data analysis process. Textbook authors should emphasize and demonstrate these analyses to a greater extent and researchers should comment on these analyses in their articles even if the results do not change the planned analysis strategy.

Table 1

Estimated sample sizes needed to detect a small, medium, and large difference in two regression slopes for three levels of power and two levels of Type I error rates

Effect Size	Probability of Type I Error	Power		
		.7	.8	.9
Small	.10	944	1240	1716
	.20	655	905	1317
Medium	.10	108	140	193
	.20	75	103	149
Large	.10	42	52	72
	.20	29	39	56

References

Bollen, K. A., & Jackman, R. W. (1990). Regression diagnostics: An expository treatment of outliers and influential cases. In J. Fox & S. Long (Eds.), Modern methods of data analysis ([p-257-291). Newbury Park, CA: Sage.

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. Journal of the American Statistical Association, 69, 364-367.

Chou, T. F. (1991). An investigation of some parametric alternatives to two-group fixed effect analysis of covariance. Unpublished doctoral dissertation, University of Georgia, Athens.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. 2nd Edition. Hillsdale, NJ: Erlbaum.

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics, 23, 351-361.

Cook, R. D. (1977). Detection of influential observation in linear regression. Technometrics, 19, 15-18.

Cornell, J. E., Young, D. M., Seaman, S. L., & Kirk, R. E. (1992). Power comparisons of eight tests or sphericity in repeated measures designs. Journal of Educational Statistics, 17, 233-249.

Dunnett, C. (1980). Pairwise multiple comparisons in the unequal variance case. Journal of the American Statistical Association, 75, 796-800.

Fletcher, H. J., Daw, H., & Young (1989). Controlling multiple F test errors with an overall F test. The Journal of Applied Behavioral Science, 25, 101-108.

Games, P., & Howell, J. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. Journal of Educational Statistics, 1, 113-125.

Glass, G., Peckham, P., & Sanders, J. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Review of Educational Research, 42, 237-288.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. Journal of the American Statistical Association, 81, 1000-1004.

Holland, B. (1991). On the application of three modified Bonferroni procedures to pairwise multiple comparisons in balanced repeated measures designs. Computational Statistics Quarterly, 3, 219-231.

Hsiung, T., & Olejnik, S. (1991, April). Power of pairwise multiple comparisons in the unequal variance case. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Huberty, C. J, & Morris, D. (1989). Multivariate analysis versus multiple univariate analyses. Psychological Bulletin, 105, 302-308.

Keselman, H. J., Keselman, J. C., & Shaffer, J. P. (1991). Multiple pairwise comparisons of repeated measures means under violation of multisample sphericity. Psychological Bulletin, 110, 162-170.

Li, J., Olejnik, S., & Huberty, C. J, (1992, April). Multiple testing with modified Bonferroni methods. Paper presented at the annual meeting of the American Educational Research Association, San Francisco CA.

Looney, S. W., & Stanley, W. B. (1989). Exploratory repeated measures analysis for two or more groups: Review and update. The American Statistician, 43, 220-225.

Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. Journal of Educational Statistics, 5, 269-287.

Maxwell, S. E. & Delaney, H. D. (1990). Designing experiments and analyzing data: A model comparison perspective. Belmont, CA: Wadsworth.

McDonald, R. P. (1984). Factor analysis and related methods. Hillsdale, NJ: Erlbaum.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105, 156-166.

Milligan, G. W., Wong, D. S., & Thompson, P. A. (1987). Robustness properties of nonorthogonal analysis of variance. Psychological Bulletin, 101, 464-470.

Moore, D. S. & McCabe, P. G. (1989). Introduction to the practice of statistics. New York: Freeman.

Moser, B. K., & Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test. American Statistician, 46, 19-21.

Neter, J., Wasserman, W., & Whitmore, G. A. (1988). Applied statistics. (3rd ed.) Boston: Allyn and Bacon.

O'Brien, R. G. (1978). Robust techniques for testing heterogeneity of variance effects in factorial designs. Psychometrika, 43, 327-342.

Olejnik, S. (1987). Conditional ANOVA for mean differences when population variances are unknown. Journal of Experimental Education, 55, 141-148.

Oshima, T. C., & Algina, J. (1992). A SAS program for testing the hypothesis of the equal means under heteroscedasticity: James's second-order test. Educational and Psychological Measurement, 52, 117-118.

Porter, A. C. & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. Journal of Counseling Psychology, 34, 383-392.

Ramsey P. H., & Brailsford, E. A. (1990). Robustness and power of tests of variability on two independent groups. British Journal of Mathematical and Statistical Psychology, 43, 113-130.

Robey, R., & Barcikowski, R. S. (1987, April). Sphericity tests and repeated measures data. Paper presented at the annual meeting of the American Educational Research Association, Washington D C.

Rogosa, D. R. (1980). Comparing non-parallel regression lines. Psychological Bulletin, 88, 307-321.

Saville, D. J. (1990). Multiple comparison procedures: The practical solution. The American Statistician, 44, 174-180.

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. Psychological Bulletin, 110, 577-586.

Shaffer, J. P. (1979). Comparison of means: An F-test followed by a modified multiple range procedure. Journal of Educational Statistics, 4, 14-23.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87, 245-251.

Stevens, J. (1990). Intermediate statistics: A modern approach. Hilsdale, NJ: Erlbaum.

Tomarken, A. J., & Serlin, R. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychological Bulletin, 99, 90-99.

Toothaker, L. E. (1991). Multiple comparisons for researchers. Newbury, CA: Sage.

Tukey, J. W. (1991). The philosophy of multiple comparisons. Statistical Science, 6, 100-116.

Tukey, J. W. (1993). Where should multiple comparisons go next? In F. M. Hoppe (Ed.) Multiple comparisons, selection, and applications in biometry (pp. 187-207). New York: Marcel Dekker.

Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W and F* statistics. Communications in Statistics - Simulation, 15, 933-943.

Appendix A

Journal Articles Reviewed

Andre, T. (1990). Type of inserted question and the study-posttest Delay. Journal of Experimental Education, 58, 77-86.

Andrews, P. E., Beal, C. E., & Corson, J. A. (1990). Talking on paper: Dialogue as a writing task for sixth graders. Journal of Experimental Education, 58, 87-94.

Blanchard, J., & Mikkelson, V. (1987). Underlining performance outcomes in expository text. Journal of Educational Research, 80, 197-201.

Casteel, C. A. (1991). Answer changing on multiple-choice test items among eighth-grade readers. Journal of Experimental Education, 59, 300-309.

Clifford, M. M., Chou, F. C., Mao, K-N, Lan, W. Y., & Kuo, S-Y. (1990). Academic risk taking, development, and external constraint. Journal of Experimental Education, 59, 45-66.

Clift, R. T., Ghatala, E. S., Naus, M. M., & Poole, J. (1990). Exploring teachers' knowledge of strategic study activity. Journal of Experimental Education, 58, 253-264.

Crano, W. D., & Johnson, C. D. (1991). Facilitating reading comprehension through spatial skills training. Journal of Experimental Education, 59, 113-128.

Cunningham, L. J., & Gall, M. D. (1990). The effects of expository and narrative prose on student achievement and attitudes toward textbooks. Journal of Experimental Education, 58, 165-176.

Dansereau, D. F., Brooks, L. W., Holley, C. D., & Collins, K. W. (1983). Learning strategies training: Effects of sequencing. Journal of Experimental Education, 51, 102-108.

Das, J. P., & Mishra, R. K. (1991). Relation between memory span, naming time, speech rate, and reading competence. Journal of Experimental Education, 59, 129-140.

Dee-Lucas, D., & Di Vesta, F. J. (1980). Learning-generated organizational aids: Effects on learning from text. Journal of Educational Psychology, 72, 304-311.

Dyer, J. W., Riley, J., & Yekovich, F. R. (1979). An analysis of three study skills: Notetaking, summarizing and rereading. Journal of Educational Research, 73, 3-7.

Erinosho, S. Y. (1990). The Effect of two remediation methods in high school physics classes in nigeria. Journal of Experimental Education, 58, 177-184.

Evans, R. D., & Evans, G. E. (1989). Cognitive mechanisms in learning from metaphors. Journal of Experimental Education, 58, 5-20.

Feldman, D., Gerstein, L. H., & Feldman, B. (1989). Teachers' beliefs about administrators and parents of handicapped and nonhandicapped students. Journal of Experimental Education, 58, 43-56.

Garner, R. (1982). Efficient text summarization: Costs and benefits. Journal of Educational Research, 75, 275-279.

Garner, R., & Gillingham, M. G. (1991). Topic knowledge, cognitive interest, and text recall: A microanalysis. Journal of Experimental Education, 59, 310-319.

Geva, E. (1992). Facilitating reading comprehension through flowcharting. Reading Research Quarterly, 18, 385-405.

Hamilton, R. (1990). The Effect of elaboration on the acquisition of conceptual problem-solving from prose. Journal of Experimental Education, 59, 5-18.

Hastie, P. A., & Saunders, J. E. (1991). Effects of class size and equipment availability on student involvement in physical education. Journal of Experimental Education, 59, 212-225.

Hayamizu, T., & Weiner, B. (1991). A test of Dweck's model of achievement goals as related to perceptions of ability. Journal of Experimental Education, 59, 226-234.

Heinze-Fry J. A., & Novak, J. D. (1990). Concept mapping brings long-term movement toward meaningful learning. Science Education, 74, 461-472.

Hynd, C. R., Simpson, M. L., & Chase, N. D. (1990). Studying narrative text: The effects of annotating vs. journal writing on test performance. Reading Research and Instruction, 29, 44-54.

Jacobowitz, T. (1990). AIM: A metacognitive strategy for constructing the main idea of text. Journal of Reading, 33, 620-624.

Johnson, G. J. (1990). Directional biases in children's perception of spatial order. Journal of Experimental Education, 59, 19-30.

Johnson, L. L. (1988). Effects of underlining textbook sentences on passage and sentence retention. Reading Research and Instruction, 28, 18-32.

King, J. R., Biggs, S., & Lipsky, S. (1984). Students' self-questioning and summarizing as reading study strategies. Journal of Reading Behavior, 16, 205-218.

Knudson, R. E. (1991). Effects of instructional strategies, grade, and sex on students' persuasive writing. Journal of Experimental Education, 59, 141-152.

Kosmoski, G. J., Gay, G., & Vockell, E. L. (1990). Cultural literacy and academic achievement. Journal of Experimental Education, 58, 265-272.

Larson, C. O., Dansereau, R. F., Hythecker, V. L., O'Donnell, A., Young, M. D., Lambrotte, J. G., & Rocklin, T. R. (1986). Technical training: An application of a strategy for learning structural and functional information. Contemporary Educational Psychology, 11, 217-228.

Martin, M. A. (1985). Students' applications of self-questioning study techniques: An investigation of their efficacy. Reading-Psychology, 6, 69-83.

McCagg, E. C., & Dansereau, D. F. (1991). A convergent paradigm for examining knowledge mapping as a learning strategy. Journal of Educational Research, 84, 317-324.

McMinn, M. M., Troyer, P. K., Hannum, L. E., & Foster, J. D. (1991). Teaching nonsexist language to college students. Journal of Experimental Education, 59, 153-164.

Meulenbroek, R. J., & Van Galen, G. P. (1990). Perceptual-motor complexity of printed and cursive letters. Journal of Experimental Education, 58, 95-110.

Miller, J. W., McKenna, M. C., & Kear, D. J. (1982). An examination of the efficiency of four reading/study techniques. Journal of Reading, 26, 239-242.

Mohanna, A. H., & Al-Heeti, K. N. (1989). Mathematical information processing skills and concept attainment. Journal of Experimental Education, 58, 21-28.

Newman, R. S., & Stevenson, H. W. (1990). Children's achievement and causal attributions in mathematics and reading. Journal of Experimental Education, 58, 197-212.

Nist, S. L., & Hogrebe, M. C. (1987). The role of underlining and annotating in remembering textual information. Reading Research and Instruction, 27, 12-25.

Okun, M. A., Weir, R. M., Richards, T. A., & Benin, M. H. (1990). Credit load as a moderator of the intent-turnover relation among community college students. Journal of Experimental Education, 58, 213-224.

Peterson, S. E., Ridenour, M. E., & Somers, S. L. (1990). Declarative, conceptual, and procedural knowledge in the understanding of fractions and acquisition of ruler measurement skills. Journal of Experimental Education, 58, 185-196.

Pigge, F. L., & Marso, R. N. (1990). A longitudinal assessment of the affective impact of preservice training on prospective teachers. Journal of Experimental Education, 58, 283-290.

Ranzijn, F. J. A. (1991). The number of video examples and the dispersion of examples as instructional design variables in teaching concepts. Journal of Experimental Education, 59, 320-330.

Rasinski, T. V. (1989). Adult readers' sensitivity to phrase boundaries in texts. Journal of Experimental Education, 58, 29-42.

Reynolds, S. B., & Hart, J. (1990). Cognitive mapping and word processing: Aids to story revision. Journal of Experimental Education, 58, 273-282.

Rogers, W. T., & Bateson, D. J. (1991). Verification of a model of test-taking behavior of high school seniors. Journal of Experimental Education, 59, 331-351.

Ruddell, R. B., & Boyle, O. F. (1989). A study of cognitive mapping as a means to improve summarization and comprehension of expository text. Reading Research and Instruction, 29, 12-22.

Simpson, M. L., & Nist, S. L. (1990). Textbook annotation: An effective and efficient study strategy for college students. Journal of Reading, 34, 122-129.

Simpson, M. L., Stahl, N., Hayes, C. G. (1989). PORPE: A research validation. Journal of Reading, 33, 22-28.

Smith, E. R., & Standel, T. C. (1981). Learning styles and study techniques. Journal of Reading, 24, 599-602.

Thompson, J. M., & May, R. B. (1987). Summary writing and text expectancy in cued recall of prose. Reading Canada Lecture, 5, 70-75.

Thornton, N. E., Bohlmeyer, E. M., Dickson, L. A., & Kulhavy, R. W. (1990). Spontaneous and imposed study tactics in learning prose. Journal of Experimental Education, 58, 111-126.

Troutt-Ervin, E. D. (1990). Application of keyword mnemonics to learning terminology in the college classroom. Journal of Experimental Education, 59, 31-44.

Tuckman, B. W. (1990). Performance of students differing in self-efficacy. Journal of Experimental Education, 58, 291-300.

Wicker, F. W., Brown, G., Hagen, A. S., Boring, W., & Weihe, J. A. (1991). Student expectations about affective correlates of academic goal setting. Journal of Experimental Education, 59, 235-249.

Appendix B

Textbooks Reviewed

Introductory

Arney, W. R. (1990) Understanding statistics in the sciences. New York: Freeman.

Christensen, L. B., & Stoup, C. M. (1991). Introduction to statistics for the social and behavioral sciences (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Freedman, D., Pisani, R. Purves, RR., & Adhikari, A. (1990). Statistics (2nd ed.). New York: Norton.

Glass, G. V. & Hopkins, K. D. (1984). Statistical methods in education and psychology. (2nd ed.) Englewood Cliffs, NJ: Prentice-Hall.

Gravetter, F. J., & Wallnau, L. B. (1992). Statistics for the behavioral sciences (3rd ed.). New York: West.

Groninger, L. D. (1990). Beginning statistics within a research context. New York: Harper & Row.

Hamilton, L. C. (1990). Modern data analysis. Pacific Grove, CA: Brooks/Cole.

Hays, W. L. (1988). Statistics. (4th ed.) New York: Holt, Rinehart and Winston.

Heiman, G. W. (1992). Basic statistics for the behavioral sciences. Boston: Houghton Mifflin.

Howell, D. C. (1992). Statistical methods for psychology. Boston: PWS-Kent.

Jaccard, J., & Becker, M. A. (1990). Statistics for the behavioral sciences (2nd ed.). Belmont, CA: Wadsworth.

Kirk, R. E. (1990). Statistics: An introduction (3rd ed.). Fort Worth: Holt, Rinehart and Winston.

Lehman, R. S. (1991). Statistics and research design in the behavioral sciences. Belmont, CA: Wadsworth.

Marasculio, L. A. & Serlin, R. C. (1988). Statistical methods for the social and behavioral sciences. New York: Freeman.

May, R. B., Masson, M. E. J., & Hunter, M. A. (1990). Application of statistics in behavioral research. New York: Harper & Row.

McCall, R. B. (1990). Fundamental statistics for behavioral sciences (5th ed.). New York: Harcourt Brace Jovanovich.

Moore, D. S. & McCabe, P. G. (1989). Introduction to the practice of statistics. New York: Freeman.

Neter, J., Wasserman, W., & Whitmore, G. A. (1988). Applied statistics. (3rd ed.) Boston: Allyn and Bacon.

Ott, L. (1988). An introduction to statistical methods and data analysis. (3rd ed.) Boston: PWS Kent.

Popham, W. J., & Sirotnik, K. A. (1992). Understanding statistics in education. Itasca, IL: Peacock.

Rosenberg, K. M. (1990). Statistics for behavioral sciences. Dubuque, IA: Brown.

Sprinthall, R. C. (1990). Basic statistical analysis (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.