Methods of Psychological Research Online 1996, Vol.1, No.4
Internet: http://www.pabst-publishers.de/mpr/

On testing a priori hypotheses about quantitative and qualitative trends[*]

Willi Hager

Abstract

Tests for quantitative trends are performed rather frequently in experimental psychology. Most often, a single test for the presumed trend or a few tests for some easily interpretable trends (of lower order) are carried out. It is argued that these procedures are susceptible to leading to inappropriate decisions on whether the data can be appropriately described by a particular trend if they are not accompanied by further tests concerning the deviations from this trend. The wide-spread F test for a particular trend is, in general, not sensitive to the two most important features of quantitative trends, that is a specific strict rank order among parameters and the strict dependency of the differences between the values of the dependent variable on the respective differences between the values of the independent variable. Some testing strategies are proposed which take these central features into account and which mainly demand further conventional tests in addition to the one referring to the presumed or predicted trend. These strategies demand the tests to be linked in a certain way to enable test-based inferences about the presence or absence of quantitative trends.

For qualitative trends, comparable problems are identified and proposals on how to solve them are presented. Because of its versatility, the method of planned contrasts is suggested in order to appropriately test a priori hypotheses on qualitative trends which have been derived from psychological hypotheses as predictions.

Keywords: Quantitative and qualitative trends, statistical testing strategies

Zusammenfassung

In der psychologischen Forschung werden Tests auf quantitative Trends vergleichsweise häufig angewendet. Dabei wird entweder nur auf einen (vermuteten oder vorhergesagten) Trend getestet oder auf einige wenige, aber selten auf alle möglichen Trends. Diese Verfahrensweise kann aber unter nicht einmal "pathologischen" Umständen leicht zu falschen Entscheidungen hinsichtlich des Vorliegens eines bestimmten Trends führen. Der Grund dafür liegt darin, daß der üblicherweise eingesetzte F-Test insensitiv für die beiden definierenden Aspekte quantitativer Trends ist, nämlich eine bestimmte Rangordnung der Parameter und genau angebbare Differenzen zwischen den Parametern, die eine Funktion der Werte der quantitativen unabhängigen Variablen darstellen. Im Falle der Prüfung vorgegebener Hypothesen kann dieses Problem dadurch vermieden werden, daß man sowohl auf den vorhergesagten Trend als auch auf das Fehlen der Abweichungen von diesem Trend testet.

Vergleichbare Probleme ergeben sich bei der Vorhersage und Prüfung qualitativer (etwa monotoner oder bitoner) Trends, wie sie sehr häufig aus psychologischen

[*] Two anonymous reviewers' valuable comments on an earlier draft of this paper helped me to clarify the issues and the presentation. Ms. Lisette Morris did her best to improve the style of presentation; the remaining faults are mine. I am indebted to these persons as well as to the managing editors of this journal, Prof. Dr. J. Rost and Prof. Dr. R. Steyer.

Hypothesen als Vorhersagen ableitbar sind. Auch hier wird eine Vorgehensweise auf der Basis der Methode der a priori geplanten Kontraste zur Prüfung entsprechender Trendhypothesen vorgeschlagen, die neben ihrer vielseitigen Verwendbarkeit noch weitere Vorteile auf sich vereinigt.

Schlüsselwörter: Quantitative und qualitative Trends, statistische Teststrategien

1 Introduction

Usually, it depends on one of two prerequisites whether tests for quantitative trends are applied or not. First, the independent variable is quantitative, and second, the independent variable is quantitative *and* a particular quantitative trend hypothesis is to be tested (see Keppel, 1973, p. 114). In the first case, the experimenter does not proceed from certain expectations, he or she just looks for the best functional description of his or her data. In the second case, however, the data are examined as to their compatibility with predictions derived from a certain theory or substantive (i.e., psychological) hypothesis. Under these circumstances, it is always possible to specify the exact relations between the independent variable and the values of the dependent variable in advance. Although this article deals exclusively with the case of *testing theories and (psychological) hypotheses* by means of predictions derived from them, the considerations presented here will prove their importance for other cases, too.

The distinction between substantive or psychological hypotheses and statistical hypotheses is often either blurred or not taken into account in empirical psychological literature as well as in some textbooks. Psychological hypotheses refer to psychological constructs such as 'aggression,' 'self-esteem,' or 'imagery' and they 'treat the phenomena of nature and man' (Clark, 1963, p. 457). In contrast, 'statistical hypotheses concern the behavior of observable random variables' (Clark, 1963, pp. 456-457) such as 'population variances,' 'population means,' 'population correlations,' and 'distribution functions.' Most often, psychological hypotheses are examined using statistical hypotheses which are only loosely related to them. This is the case, for instance, when the psychological hypothesis enables the prediction of a certain *rank order* of parameters across several experimental conditions and the well known F test is applied, testing against the hypothesis that not all parameters (population means μ_k) are equal or homogeneous.

Some authors call for a closer connection between the psychological hypothesis and the statistical hypothesis or hypotheses. They specifically demand that statistical hypotheses should be *derived* from the psychological one, 'even in a rather loose sense of derive' (cf. Hager, 1987, 1992; Meehl, 1967; Wampold, Davis & Good, 1992; Westermann & Hager, 1986). Hager (1992, pp. 54-68) has argued that this derivation should preserve the psychological hypothesis' empirical content as it is understood by Popper (1981, 1992). To this aim, he has proposed two additional criteria of derivation, namely *appropriateness* and *exhaustiveness.*

'Appropriateness' means that the derived statistical hypothesis has to conform with the *direction of the relation* claimed in the psychological hypothesis, and 'exhaustiveness' means that a prediction has to *encompass any relation or aspect of the psychological hypothesis* which can be expressed by statistical concepts (see Hager, 1987, 1992, and Hager & Hasselhorn, 1995, for further details). If a statistical hypothesis is connected to a psychological hypothesis by a derivation and if it meets with the two criteria just mentioned, it is called a *statistical prediction* (SP for short). This linkage between two kinds of hypotheses by a derivation together with two criteria seems necessary and sufficient to ensure an *unambiguous* separation of those results which are in complete accordance with the psychological hypothesis from those that contradict it. Such a partition of possible results conforms to demands formulated by Fisher (e.g., 1966) as well as by Popper (1980).

It is, however, very often the case in empirical psychological literature that this basic principle, advocated independently by a statistician and by a philosopher, is violated, as the analyses by Hager (1992), by Hager and Westermann (1983) and by Westermann and Hager (1986) show.

A statistical prediction is a special statistical hypothesis which is not necessarily equivalent to the null or the alternative hypothesis of a (wide-spread and/or single) statistical test. A null hypothesis (H_0) is any statistical hypothesis which comprises one of the signs '$=$', '$\leq$', or '$\geq$' and which is testable by a given statistical test. It's opposite is an alternative hypothesis (H_1), which usually is complementary to the H_0 and against which the test is performed. Furthermore an H_1 usually refers to the relations '$\neq$', '$>$' or '$<$'. This distinction is made in most textbooks for psychologists (see Hays, 1988; Howell, 1992; Kirk, 1982; Wilcox, 1987; Winer, Brown & Michels, 1991) and suffices for the purposes of this article. If the statistical prediction is not eqivalent to a single testable H_0 or H_1, there are basically two options: either to perform a less well suited test and interpret the 'apparent' empirical relations among the sample statistics, or to apply more than one test. The more tests that are performed the greater the cumulation of statistical error probabilities, but the greater information gained in general. Besides, the cumulation can be adjusted for, but the possible adjustments will not be considered in any detail (see, among many others, Hochberg & Tamhane, 1987; Kirk, 1982, 1994; Miller, 1981; Westermann & Hager, 1986).

Choosing the first option means that either one or both of the principles of appropriateness and exhaustiveness with respect to the particular statistical prediction is violated by the statistical hypotheses *actually* tested, and/or that the decisions made are mainly data-based. *Data-based* decisions rely on statistical tests *and* on subsequent differential interpretations of data patterns. If - for example - the significance of an overall F test is taken as the basis for interpreting the rank order of *sample* means as being the same as of the *population* means, this is a data-based decision not covered by the test performed. If the F test is performed on a comparison with more than one degree of freedom, it does not refer to distances $\Delta_{kk'}$ among the means, but to a quadratic function of these distances, which are squared, summed up, and averaged for the purposes of the F test. Besides, more individual decisions are made than are covered by the nominal significance level α of the F test (Ramsey, 1980), as the increase in the conditional probabilities α and/or β depends on the *number of decisions* actually made rather than on the *number of tests* performed. The 'correct' *test-based* interpretation of a significant F value only permits saying that there are at least two population means different from one another. The numerous techniques of multiple comparisons can be said to have been developed to replace mainly data-based statements with test-based propositions, controlling for the cumulation of the error probability α. In contrast, test-based decisions are based on tests only and they are not modified, 'corrected,' or augmented by additional interpretations of the data patterns. These considerations should not be taken as an argument against careful data inspections, which always should be done. The present article deals with some testing strategies, the application of which enable making test-based decisions and avoiding data-based decisions.

If, on the other hand, the statistical hypotheses actually tested turn out to be only loosely linked to the psychological hypothesis of interest or to the statistical prediction derived from it, the probability of false decisions concerning the psychological hypothesis can be enhanced substantially, or in more general terms: the probability of false 'truths' can be enhanced greatly. I will cite no examples from current empirical literature to demonstrate this, but rather deal with some textbook presentations; empirical researchers should not be expected to act in a more sophisticated manner than textbook authors. To lower the probability of false 'truths' it is important to apply the criteria of adequateness and exhaustiveness when deriving testable statistical hypotheses from the statistical prediction or when decomposing it into testable partial hypotheses. Several of the subsequent considerations will focus on this demand.

If psychologists describe relations among variables by means of mathematical functions they aim for a greater degree of exactness or precision than is possible when using less precise methods of description. This goal of exactness, however, may be rendered unattainable by choosing tests which are not exact enough: One is working with a very precise and seemingly exact scientific terminology and hypotheses, but because of inappropriate statistical procedures the hypotheses actually tested do not reflect the quantification or the functionality to a sufficient degree. This lack of correspondence between the *quantitative* hypothesis to be tested and the one actually tested will be examined subsequently. It will also be argued that certain tests of *qualitative* trend hypotheses can result in analoguous problems.

By calling a trend hypothesis a statistical prediction, it is meant that hypotheses of this kind can occur in empirical research, that is, may serve as the target hypothesis to be tested. I shall not deal with psychological hypotheses leading to a particular statistical trend prediction, but intend to discuss some trend hypotheses and their relation to some commonly administered tests. Thus, the main question I seek to answer is: *Given a particular (quantitative or qualitative) trend hypothesis, which of some well-known statistical tests is best suited to test it*, whereby 'best suited' does not refer to statistical assumptions, but to features of trends. This question will be discussed from the perspective of the *method of planned or focussed contrasts* (among expectations of normally distributed random variables or population means μ since '... it is to the experimenter's advantage to specify a select, limited number of contrasts in advance' (Kirk, 1982, p. 106). The usual parametric assumptions are taken for granted throughout, equal n's in a one-way layout are assumed, and the quantitative variable X (values x_1 through x_K) is equidistant. Despite these restrictions, the general considerations are applicable to other parameters, tests, and layouts than those addressed herein (see, for example, Marascuilo & Mc Sweeney, 1977). Furthermore, it is assumed that appropriate power analyses for controlling both conditional error probabilities (α and β) takes place (see Cohen, 1988; Hager, 1987, 1995). No reference will be made to more robust alternatives to the tests considered (see, e.g., Wilcox, 1987) and to the various procedures of ordering and selection which seem to be more appropriate for data analyses *after* data collection (see, e.g., Dykstra, Robertson & Wright, 1986; Lovie, 1986; Robertson, Wright & Dykstra, 1988, and Wilcox, 1987, chap. 12). These techniques, however, may be applied *in addition* to the tests considered here, but the examination of psychological hypotheses formulated in advance should be separated carefully from additional data analyses which could also be interesting. *Testing psychological hypotheses means that the kind of trend can and most importantly should be predicted prior to data collection.* Since the testing strategies proposed subsequently mainly consist in suggestions of how to link certain well-known tests no reference will be made to particular computer programs for data analyses.

2 Quantitative Trends, Trend Tests, And Testing Strategies

Since the identification of the main features of quantitative trends is lacking in standard textbooks on statistics and experimental design for psychologists (cf., for example, Cohen & Cohen, 1983; Edwards, 1985; Kirk, 1982; Maxwell & Delaney, 1990; Myers & Well, 1991; Wilcox, 1987; Winer et al., 1991), these will be delineated for linear trends as they are the kind of quantitative trends most often of (primary) interest. For brevity's sake, the generalizations to other trends will not be considered here; they are straightforward, although definitions and formulas are more complex than for a linear function.

2.1 Main features of linear trends and the standard test on linearity

An example for a quantitative psychological hypothesis is given by Myers and Well (1991, p. 204): 'The experimenters ... believe that the magnitude of conditioned responses should vary directly with the magnitude of the test stimulus; this implies that ... [the galvanic skin response] scores should increase as a function of the height of the rectangle of light.' If we translate the rather unprecise formulation 'increase as a function' into 'increase as a *linear* function', we get a quantitative psychological hypothesis from which the prediction of a quantitative linear trend can be derived. Speaking of quantitative trends means that the values of the dependent variable (Y) and the values of the quantitative independent variable (X) are related by a function which is linear, in the case of linear trends. Therefore, a (positive) linear trend has two characteristics: first, the complete specification of a strictly increasing rank order for all K population means of Y parallel to a corresponding increase in the values of X; and second, specification of distances or differences $\Delta_{kk'}$ between any two means μ_k and $\mu_{k'}$ ($k' = k-1$) in functional dependence on the corresponding distances between two values x_k and $x_{k'}$ of X (k denoting a particular experimental condition; the values of X increase with indices). These components lead to the following definition, expressed as a statistical hypothesis (SH) about population means μ_k:
SH-(strictly linear trend, positive) = *SH-lin:*

$$[(\mu_k < \mu_{k'}) \wedge (\Delta_{kk'} = g > 0)] \text{ for } \textit{all} \;\; k, k' \text{ with } k = k' - 1, k = 1, \ldots, K-1, \qquad (1)$$

with '$\wedge$' symbolizing a logical conjunction and 'g' being a real valued constant in functional dependence on the corresponding distances among values of the independent variable X. The minimum number of experimental conditions necessary to detect a linear (and a quadratic) trend is $K = 3$, since, in addition to the rank order of means, two differences or distances have to be compared as to their compatibility with the functional rule. Graphic representations of quantitative trends will not be given here, since they can be found in many textbooks (see, e.g., Edwards, 1985, p. 146; Keppel & Zedeck, 1989, p. 492; Maxwell & Delaney, 1990, p. 222; Meddis, 1973, p. 85; Myers & Well, 1991, p. 212; Winer et al., 1991, p. 199).

Although it seems possible to develop tests directly aiming at the two main features of linear or other quantitative trends, statistical hypotheses usually refer to these trends less directly. Most, if not all textbooks dealing with the analysis of quantitative trends address the *method of orthogonal polynomials* through which it is possible to split the sum of squares between the K experimental conditions (SS_{bet}) into $K-1$ orthogonal trend components, each associated with a single degree of freedom (*df* = 1) and with a single sum of squares. They then present two kinds of tests, one for the predicted trend or the trend of main interest and another one for the deviations from this trend. The sum of squares associated with the trend of interest, SS_{lin} in our example, is submitted to an F test. Usually, two possibilities for testing for deviations are discussed. One consists of adding the sums of squares of the deviations from the trend of interest to form SS_{dev} with *df* $= K-2$ and to perform a single F test on these deviations. The other consists of performing $K-2$ separate F tests on each trend component (*df* = 1).

If the author is mainly interested how the form of relationship between X and Y can best be approximated, she or he will employ the latter procedure in order to know which trend components are needed for an adequate description of the data (see, e.g., Hays, 1988, pp. 709-710; Myers & Well, 1991, pp. 211-216; Winer et al., 1991, pp. 203-204). If he or she is only interested in the the questions whether there are any deviations from predictions, usually only one F test is performed covering all deviations simultaneuously (cf. Edwards, 1985, p. 147).

Although these tests were addressed in all textbooks I reviewed their meaning was not explicitly covered. The impression is often given that testing for deviations is optional or less important than testing for the trend of interest. Keppel and Saufley (1980, p. 311), for example, focus on the expected trend and state: 'Since the obtained ... (F_{lin}) ratio exceeds ... [the] critical value, we reject the null hypothesis and conclude that linear trend is present in ... [the] data.' Myers and Well (1991, p. 207) state with respect to their example: '... a significant linear trend would support the hypothesis that the magnitude of a conditioned response tends to increase as the magnitude of the stimulus increases.' (This interpretation also holds true for a statistically significant *qualitative* trend, since the functional dependence is not addressed.) Other textbook authors perform both (or more) tests and interpret the results in a 'test-oriented' manner, that is, by stating that both the linear and the other trend components are 'statistically significant' (see, e.g., A.L. Edwards & L.K. Edwards, 1994, p. 51; Howell, 1992, p. 374; Myers, 1972, p. 388). What is typically disregarded, however, is the important *implication* of this statement: The significance of the quadratic component and/or of other higher-order components indicates a deviation from homogeneity in the *differences* among means, or even an inversion of ranks. Both of these patterns are not consistent with strict linearity; a point which is rarely unequivocally addressed (e.g., by Lee, 1975, p. 310, as an exception). To show this, let us consider the following example taken from Winer et al. (1991, pp. 203-205).

The authors deal with $K = 6$ experimental conditions, the treatment totals being T_k = 100; 110; 120; 180; 190; 210 (Case 1). The linear component is $C_{lin} = 850$, SS_{lin} = 1032.14, $MS_{error} = 18.52$, and $F = 55.73$, significant at $\alpha = .01$, as we would expect for the data given. Now let us consider set of fictitious data: T'_k = 100; 100; 100; 100; 100; 210 (Case 2), for which the linear contrast is $C_{lin} = 550$, $SS_{lin} = 432.14$, and F = 23.33, significant at $\alpha = .01$. The same results hold for another set of fictitious treatment totals: T''_k = 100; 0; 0; 0; 0; 210 (Case 3), though this would not necessarily be expected. In still another case we have: T'''_k = 100; 100; 100; 100; 250; 210 (Case 4), for which $C_{lin} = 1200$, $SS_{lin} = 2057.14$, and $F = 110.08$, significant at $\alpha = .01$. (Inspection of data in the latter three cases would certainly lead to some caution when interpreting the significances, but exactly this is the kind of *data-based* 'correction' of test results I seek to avoid.)

The examples illustrate two points: First, the significance of the linear trend component may not be as meaningful as one might expect. Second, the tests aims at the H_1: $\sum(c_{k,lin} \cdot \mu_k)^2 > 0$, which should be interpreted as the conventional H_1 of the F test: $H_1 : \mu_k \neq \mu_{k'}$ for *at least* one pair of means with $k \neq k'$ (instead of $k' = k - 1$), as especially Case 4 shows. This hypothesis is far less precise than is necessary when dealing with quantitative trends as defined above, because there usually is a good chance for the respective test to be significant if at least two means differ, the most likely candidates for this being the two means at the extremes (experimental conditions x_1 and x_K).

But data-based 'corrections' are not necessary, since the information required for the appropriate interpretation of the test results is contained in the test(s) for deviations from linearity. In Case 1, the example taken from Winer et al., the F test for deviations comes out insignificant ($F = 1.03$; Winer et al., 1991, p. 205), which means that the deviations from linearity are small as judged by conventional tests of significance. One can say that tests of significance introduce a probabilistic element into the exact definitions of quantitative trends. In Case 2, we compute $SS_{dev} = 1108.33$ and $F = 14.96$, significant at $\alpha = .01$. For Case 3, we get $SS_{dev} = 3376.19$ and $F = 45.57$, also significant at α = .01; and for Case 4: $SS_{dev} = 1133.33$ and $F = 15.30$, again significant at $\alpha = .01$. The additional examples show that there may be a (rather weak) linear component, but there also are substantial deviations from linearity.

As argued above, deviations from linearity mean that there are different distances $\Delta_{kk'}$ and/or rank inversions for the means and both kinds of deviations are not in accordance with the definition of linearity. But if there are no such deviations (Case 1), there are no rank inversions, although the distances $\Delta_{kk'}$ may vary within the limits of chance as defined by the tests applied and their probabilistic side conditions such as sample size

n, α, β, and so on, as is the case for the original data of Winer et al. (1991). Thus, it is vital for any analysis of quantitative trends to test *both* the trend component of interest *and* the deviations from it for significance. The necessity for considering the deviations becomes even more apparent when dealing with an odd number of experimental conditions since the linear trend coefficient for the middle position $[(K-1)/2+1]$ always is $c_{lin,[(K-1)/2+1]} = 0$. Thus the mean at this position may be of *any size without affecting the magnitude of the linear component.*

Returning to the case of strict hypothesis testing, one may consider the subsequent testing strategy as a possibility to avoid the interpretational difficulties addressed above.

2.2 A testing strategy for hypotheses on quantitative trends

Let us assume that a psychological theory or hypothesis or some prior experiments lead to the (statistical) prediction of an exclusively positive linear relation between two variables X and $Y(SP-lin)$. Translating the definition of strict positive linearity given in (1) into testable statistical (partial) hypotheses results in:
SP-lin:

$$\left[(H_{1,lin}(t) : \psi_{lin} = \sum c_{k,lin} \cdot \mu_k > 0) \wedge (H_{0,dev}(F) : \sum_{q'=2}^{K-1} \psi^2_{q'(dev)} = 0) \right] \qquad (2)$$

where 'ψ' denotes a population contrast, $c_{k,lin}$ are the orthogonal polynomials, 'q' stands for any non-linear contrast ($q' = 2, \ldots, K-1$), and t refers to a (one-sided) t test on a directional statistical hypothesis, expressing a *positive* linear relation. In the case of predicting a negative linear trend the $H_{1,lin}(t)$ should of course be: $\psi_{lin} < 0$, and if the researcher predicts more than one trend, further tests referring to the additional components must also be planned (see for an example Myers & Well, 1991, p. 216).

The presentation in (2), however, mixes two aspects. The *SP-lin* is decomposed into two statistical hypotheses, one about a contrast (with *df* $= 1$) and one about a comparison (*df* ≥ 1) consisting of one or more contrasts. But in order to make a decision concerning the *SP-lin* a *decision rule* has to be defined: A *strict* decision rule states that the statistical prediction *SP-lin* is only to be accepted if *both* partial hypotheses can be accepted.[1] According to this rule the two statististical partial hypotheses derived from it have to be connected *conjunctively* ($\wedge$) (see Morgenstern, 1980, for an analogous procedure). Using a *lenient* decision rule leads to linking the partial hypotheses derived from the *SP-lin* by the disjunctive operator ($\vee$), and it results in accepting the *SP-lin* if *at least one* of the derived hypotheses is accepted. Such a disjunctive linkage of both hypotheses would not fulfill the criterion of exhaustiveness. Although it is sometimes possible to apply the lenient decision rule even with quantitative predictions, this case will not be considered at present. But the choice of a decision rule has important consequences with respect to the psychological hypothesis.

[1] Despite many statisticians' strict refusal to even consider 'accepting' an H_0, this kind of decision must be possible or admitted when examining *substantive or psychological* hypotheses via the statistical hypotheses derived from them. Bredenkamp (1972, 1980) has repeatedly presented the reasons why null hypotheses *must* be 'acceptable:' Usually, but not always, a null hypothesis contradicts a substantive or psychological proposition. If null hypotheses cannot be accepted, no decision *against* this proposition is possible, as is necessary from a falsificationist point of view (but see below). If a null hypothesis is retained we act as if it holds (see Cook & Campbell, 1979, pp. 44-45), whereby 'acting as if' encompasses decisions on psychological hypotheses. And if power a priori is great enough that the probability of a wrong retention of a null hypothesis is controlled there seems to be no convincing reason why researchers cannot *decide* to accept or to retain a null hypothesis, although R.A. Fisher (e.g., 1966) always repudiated the notion of *deciding* about statistical hypotheses and especially of 'accepting' null hypotheses. Despite of this many null hypotheses have been retained or accepted in empirical psychological literature, and retention of a null hypothesis is the expected result for any test of model fitting. Decisions, however, are *not proofs*, neither of the null nor of the alternative hypothesis: both of them cannot be proved in the sense of showing them as being 'true' (see Gigerenzer, 1993; Hager, 1992, and Serlin & Lapsley, 1993, for further thoughts and literature on this problem).

Since psychological hypotheses do not contain any information concerning decision rules, as far as this author knows, these rules have to be chosen on other grounds, whenever more than one statistical test is necessary for an exhaustive examination of a particular psychological hypothesis. With respect to psychological hypotheses the following relations holds in general, other things being equal: The stricter the decision rule, the more severe the test of the psychological hypothesis, that is, the higher the probability that it is 'not confirmed' if it indeed is false; see Popper (1980, pp. 119-123; 1981, pp. 388-391) for a definition of 'severe'. The more lenient the decision rule, the less severe the test, that is, the higher the probability that the psychological hypothesis is called 'confirmed' if it indeed holds true. These and further relations will not be discussed here (see Hager, 1992; Westermann, 1988; Westermann & Hager, 1983, 1986). For brevity's sake the aspect of mere decomposition of a statistical prediction and choosing a decision rule to link the partial hypotheses will not be dealt with in any depth in the remainder of the text.

To perform an F test on the linear trend component violates the criterion of appropriateness, since a *directional* relation between the dependent and the independent variable is predicted. Failing to perform a test concerning the deviations from linearity, the predicted trend, violates the criterion of exhaustiveness, since a significant linear component does not exclude the existence of other trends, whereas the psychological hypothesis claims them to be absent. Another case, although rarely encountered, consists of predicting a linear trend without specifying its direction. The hypothesis to be tested in this instance is $H_{1,lin}: \psi_{lin} \neq 0$, which can be tested by a two-sided t or a (one-sided) F test. This does not affect the null hypothesis of no deviations.

Looking upon trend tests from the *regression analysis* point of view (see Bredenkamp, 1980; Cohen & Cohen, 1983; Keppel & Zedeck, 1989, pp. 496-499), the hypotheses just handled can also be expressed using correlations:

$$SP-lin: \left[(H_{1,lin}: \rho_{lin} > 0) \wedge \left(H_{0,dev}: \eta^2 - \rho^2_{lin} = 0\right)\right], \tag{3}$$

where ρ denotes the simple (population) correlation coefficient referring to the predicted trend and η^2 the squared multiple correlation (in the populations) containing all trend components.

There is another version of this procedure called Testing Strategy 1 (TS 1) that leads to the same results as the two versions just considered. It rests on the fact that a functional relation enables the *prediction of population means* μ^*_k using the general formula given, for example, by Keppel (1973, pp. 128-130) and by Myers and Well (1991, pp. 207-215). These predicted values are then compared to the actual values μ_k; the values μ^*_k and μ_k can, of course, be estimated from the sample data (this procedure is discussed by Myers & Well, 1991, pp. 205-208). The application of the prediction equation should however be preceded by testing if the predicted trend component is of sufficient magnitude, since even insignificant trend components can be used for prediction. The degree of deviation from the predicted (linear) trend is then determined by using the difference $\mu_k - \mu^*_k$. Subsequently the null hypothesis that these deviations simultaneously equal null is tested by an F test against the alternative that there is a deviation from null in at least one experimental condition. The statistical prediction of linearity, *SP-lin*, then, can be accepted if the following conjunction of partial hypotheses holds, which is equivalent to the two former expressions:

$$SP-lin: \left[(H_{1,lin}: \psi_{lin} > 0) \wedge (H_{0,dev}(F): \sum(\mu_k - \mu^*_k)^2\right]. \tag{4}$$

Before presenting two further testing strategies the possible patterns of results for Testing Strategy TS 1 and their interpretations will be addressed.

2.3 Possible patterns of results for Testing Strategy TS 1

Assuming a linear trend is predicted once again, Testing Strategy TS 1 can lead to the following patterns of decisions, each one supplemented with its appropriate test-based interpretation ('*AH*' means 'acceptance of hypothesis *H*... or retention in cases of null-hypotheses', necessary minimum power assumed):

1) $\mathbf{AH_{0,lin}} \wedge \mathbf{AH_{0,dev}}$: All population means are homogeneous; there is no trend. The *SP-lin* is rejected.

2) $\mathbf{AH_{1,lin}} \wedge \mathbf{AH_{0,dev}}$: Not all population means are homogeneous and the data are exhaustively describable by means of a strictly linear trend. The *SP-lin* is accepted.

3) $\mathbf{AH_{0,lin}} \wedge \mathbf{AH_{1,dev}}$: Not all population means are homogeneous and the data can be described exclusively by higher-order trends. The *SP-lin* again is rejected.

4) $\mathbf{AH_{1,lin}} \wedge \mathbf{AH_{1,dev}}$: Not all population means are equal, and in order to describe the data exhaustively, both linear and non-linear trend components have to be considered. Although this pattern of results does not lead to accepting the *SP-lin*, it is in accordance with the less precise (psychological) hypothesis and prediction that there is at least one trend in the data (*SP-trend*).

But it cannot even from the predicted pattern of results be inferred that the distances $\Delta_{kk'}$ are 'large'. The only interpretation possible is that they are homogeneous or (about) equal and that they have the predicted algebraic sign. Considered alone, the test for linearity, if significant, does not even allow for this inference: It only tells us that at least the *largest* distance is 'large', but it tells us nothing about (the necessary) *homogeneity of all distances* nor of their algebraic signs.

2.4 Two further testing strategies

From the hypothesis testing point of view the decomposition of the *SP-lin* given in (2), (3), and (4) (Testing Strategy TS1) is adequate and exhaustive, meaning that any (statistical) information needed to decide on the statistical prediction and the corresponding psychological hypothesis is contained in the two tests. But often questions arise which should and can, in addition to strict hypothesis testing, be taken into consideration when decomposing statistical predictions such as the *SP-lin*. Usually, these additional questions first refer to the type of trend possibly responsible for the deviations (if such occur) and second to the experimental conditions in which deviations occur. To answer these questions in a test-based manner, more than the two tests from Testing Strategy TS 1 are necessary, which means there will be a greater cumulation of the statistical error probabilities α and/or β. But this is the usual price you have to pay for more information. This possible disadvantage, however, can be compensated for by enlarging sample size (power analysis; see Cohen, 1988; Hager, 1992).

In the first case the global hypothesis of no deviations from the predictions should be decomposed in $K-2$ partial hypotheses, each concerning one trend component (maximum number of partial hypotheses: $K-1$), or, alternatively, in as many partial hypotheses as refer to meaningfully interpretable trend components plus a further partial hypothesis referring to all higher-order components (see, e.g., Keppel, 1973, pp. 127-128, and Myers & Well, 1991, p. 216, on this point).

Thus, given a particular prediction (linear in our case) *and* further questions concerning the type of possible deviations from linearity, the *SP-lin* should be decomposed into directly testable partial hypotheses, leading to Testing Strategy TS 2, which is closely related to Expression (2):

$$SP-lin: \left[\begin{array}{l} (H_{1,lin}(t): \psi_{lin} > 0) \wedge (H_{0,qua}: \psi_{qua} = 0) \wedge \\ (H_{0,cub}: \psi_{cub} = 0) \wedge \ldots \wedge (H_{0,trend(K-2)}: \psi_{trend(K-2)} = 0) \end{array} \right] \quad (5)$$

The tests concerning the derived null hypotheses can be performed as two-sided t or as (one-sided) F tests, in which case the trend contrasts have to be squared. If a study consists of more than three experimental conditions the tests for the strategies, TS 1 and TS 2, are based on different numerator degrees of freedom, that is, on different probabilistic testing conditions, and can entail different decisions (see, e.g., Kirk, 1982, p. 156; Maxwell & Delaney, 1990, p. 226). For this reason, these two procedures should be differentiated and are considered as different strategies.

In the case of deviations from the predictions, neither Testing Strategy TS 1 nor Testing Strategy TS 2 makes a *test-based* identification of the corresponding *experimental conditions* where these deviations arise possible. If the experimenter wants to obtain this information *in addition* to testing a priori hypotheses, Testing Strategy TS 3, which is more closely related to Expression (3), is preferable. According to this strategy, the test of the predicted trend component is followed by a separate statistical hypothesis for each experimental condition, postulating that there is no deviation from the prediction:

$$H_{0,k} : \mu_k - \mu_k^* = 0; k = 1, \ldots , K. \tag{6}$$

The rationale underlying the preceding decompositions calls for combining the tests with the strict *decision rule*: 'Only in case of the acceptance of $H_{1,lin}$ *and* of retention of *all* $KH_{0,k}$ (sufficient power provided) should strict linearity be inferred.' This results in the following decomposition of the *SP-lin*, leading to $K + 1$ partial hypotheses:

$$SP - lin : [(H_{1,lin}(t) : \psi_{lin} > 0) \wedge (H_{0,k} : \mu_k - \mu_k^* = 0 \text{ for all } k)] \tag{7}$$

As a consequence of this decision rule, a strictly (positive) linear trend should not be inferred if *one or more* of the alternatives $H_{1,k}$ is accepted. Although this finding contradicts the prediction, the particular tests planned allow for the identification of the experimental conditions in which the deviations from the predicted trend occur. The researcher should then try to find out possible reasons for the deviations in these experimental conditions. A possible application of this strategy is addressed by Keppel (1973, pp. 90-91). Testing Strategy TS 3 may lead to an overall decision concerning the *SP-lin* different from the overall decisions made using Testing Strategies TS 1 and TS 2, since the one-sample t tests rest on probabilistic testing conditions different from those of the other tests.

The testing strategies proposed in the preceding paragraphs can also be applied if two or more quantitative hypotheses which aim at the same phenomenon, but postulate different functional rules are to be tested in one experiment (see Hager, 1993). Moreover, they can be are generalized to other designs than the one chosen here (see Hager, 1992).

Further testing strategies which systematically aim at the two central features of quantitative trends can be constructed quite easily but will not be considered here, and nor will those methods of estimating parameters from the data which lead to F tests with reduced numerator degrees of freedom be dealt with (e.g. Kirk, 1982, pp. 159-161). In addition, Cohen and Cohen (1983, pp. 242-252), Lee (1975, pp. 307-313), Maxwell and Delaney (1990, chap. 6) and Winer et al. (1991, pp. 234-236) discuss various models for trend analysis, especially with respect to determining MS_{error}.

If a functional rule can be used to predict means, their rank order and the distances between them, it also enables the prediction of the magnitude of variances, correlations, and so on from a psychological hypothesis. Although these values, exactly predicted from theory or hypothesis, can be used as effect sizes in power analysis or sample size determination, they cannot eliminate the arbitrariness inherent in specifying effect sizes prior to experimentation, as is sometimes claimed (cf. Cohen, 1988). This arbitrariness is introduced once more when specifying the effect sizes for *deviations* from predictions. These effect sizes are also necessary, but cannot be predicted from psychological theory; instead, they must be chosen primarily, if not exclusively according to methodological or

Table 1: Decomposition of a statistical prediction concerning a linear trend (SP-lin) into testable a priori hypotheses

Additional question	Decomposition and decision rule
None	$H_{1,lin} \wedge H_{0,dev}$
Identifying trends deviating from prediction	$H_{1,lin} \wedge H_{0,qua} \wedge H_{0,cub} \wedge \ldots \wedge H_{0,trend(K-1)}$ [a]
Identifying experimental conditions deviating from prediction	$H_{1,lin} \wedge H_{0,1} \wedge H_{0,2} \wedge \ldots \wedge H_{0,K}$

Notes. [a] A more parsimonious decomposition is adressed in the text. '$\wedge$' symbolizes the conjunctive linkage of the partial hypotheses and implies the strict decision rule. $H_{1,lin}$ denotes the (directional or bidirectional) alternative hypothesis about the predicted trend (*lin* is for linear tend, but can be substituted by any other predicted trend); if more trends have been predicted, each of them should be associated with a separate statistical hypothesis of the type of the $H_{1,lin}$. $H_{0,dev}$ stands for all trend components except the predicted (linear) one; $H_{0,qua}, H_{0,cub}$ and so on refer to the $K-2$ separate null hypotheses comparing predicted and actual means for each of K experimental conditions. See text for further information.

other considerations, such as the availability of subjects. Arbitrariness cannot be banned from statistics (see Gigerenzer, 1993; Hager, 1992).

The hypotheses and the proposed decompositions of the statistical predictions have been summarized in Table 1. It should be stressed, by the way, that I only consider statistical hypotheses and tests that are necessary and adequate with respect to given psychological hypotheses formulated in advance (and with respect to certain additional questions). This focus should not prevent the researcher from additionally performing those tests which she or he thinks appropriate to gain additional information not (directly) connected to the psychological hypothesis that is to be examined (see above).
Some testing strategies for qualitative trends will be addressed next.

3 Qualitative Trends And Trend Tests

In contrast to quantitative trends, there is no mathematical function underlying qualitative trend hypotheses. These are usually derived from qualitiative psychological hypotheses addressing *qualitative* independent variables, whose levels are either categorical or can be rank ordered, but they can also be formulated or derived when the independent variable is a quantitative. In some instances it seems appropriate to disregard the quantitative nature of the dependent variable or the possibility of assigning numerical values to it, because it is doubtful, for example, if these are psychologically meaningful. When studying hypotheses about the effects of, say, imagery (Paivio, 1986) it is possible to describe word lists, for example, according to their mean imagery values, as computed from respective norms. But since these mean values usually are interpreted as ranks indicating different levels of imagery without specifying any distances, this variable is most often considered to be qualitative. The subsequent considerations refer to qualitative hypotheses about qualitative trends.

3.1 Monotonic trend as an example of a qualitative trend

As far as theory or hypothesis testing is concerned, it can be said that most of the theories and hypotheses in psychology refer to qualitative relations and lead to predictions concerning qualitative trends (see the overview by Hager, 1992). The type of qualitative trend most often encountered is the monotonic trend, which can be predicted when considering psychological hypotheses like 'The higher the degree of imagery the better the retrieval' (Paivio, 1986) and when considering hypotheses concerning the effectiveness of different cognitive programs or therapies to a, say, comparison group without any intervention. This type of trend will be the focus of interest in the following sections of this article, as it is generally not addressed by text book authors (but see Bortz, 1993, pp. 259-260). No graphic representations of qualitative trends will be given here, as these graphs can be misleading because the distances between any two levels of the qualitative independent variable cannot be defined. Choosing equal distances leads to graphs which do not differ from graphs of linear trends, while choosing arbitrary distances leads to arbitrary graphic representations. Both of these procedures are neither correct nor incorrect, but they are just arbitrary, and the impression the graphs give depends mainly upon these arbitrary choices. In some empirical papers which I deliberately do not cite here, equal distances have been chosen for the graphical representations leading to straight lines. This led the authors to claim that the trend was 'linear' although the independent variable was qualitative instead of quantitative. Myers and Well (1991, pp. 568-569) state: '... according to the Yerkes-Dodson law, we would expect a *quadratic* relation between measures of performance and motivation' (italics added). Since motivation usually is considered a qualitative variable the relation to be expected is qualitative (inverted U-shaped or *bitonic*).

The most important implication of the lack of a functional rule connecting independent and dependent variable is that the relative magnitude of differences or distances $\Delta_{jj'}$, [referred to as $\Delta_{kk'}$ in the *SP-lin* above] between population means can neither be expressed as a function of the values of the independent variable nor can they be predicted from theory. Thus, the only predictions possible refer to the *rank order of the parameters* chosen. To speak of strict monotonicity means that there is a strictly increasing order of all ranks assigned to the parameters. Expressing this definition in the format of a statistical hypothesis (SH) referring to population means μ_j, we get:

$$\textit{SH}\text{-(strictly increasing monotonic trend)} = \textit{SH-mon: } \mu_1 < \mu_2 < \ldots < \mu_J. \qquad (8)$$

The minimum number of experimental conditions is $J = 2$ in this case, since a minimum of two ranks have to be assessed or compared. Let's call this minimum number a 'testing instance,' since a psychological hypothesis is testable when this minimum number is accounted for in the experiment. As soon as the experimenter decides to study more than this minimum number, more than one testing instance can be defined, and these testing instances can be linked either conjunctively or disjunctively (whereby the tests referring to one testing instance should always be conjunctively connected to fulfill the criterion of exhaustiveness). As argued above, a conjunctive linkage represents a strict decision rule and leads to more severe tests of the psychological hypothesis, while the disjunctive connection implies (more) lenient decision rules and leads to more lenient or less severe tests.

Based on these considerations, the definition for monotonicity of a trend can be combined with several *decision rules.* The strictest decision rule demands that *all* empirical means M_j must follow the predicted order without exception *and* the differences between them must be statistically significant in order to accept the *SH-mon* in (8). The latter demand means that if, for example, the adjacent means $M_j = 20$ and $M_{j+1} = 21$, referring to the mean number of words correctly reproduced ($n = 16; MS_{error} = 25$), do not differ significantly, they are not considered different, and then should be assigned equal ranks. This demand of statistical significance between any two adjacent means is analoguous to the two-sample situation ($J = 2$), in which two means are only considered different if the difference is statistically significant; any significant difference, however, implies that different ranks are assigned to the means. Although this demand may not seem cogent, it will be taken here as an additional criterion. Combining the statistical hypothesis in (9) with this strict decision rule results in the following expression *SP-mon1*:

$$\textit{SP-mon1:}\ \mu_j < \mu_{j'} \text{ for } \textit{all}\ j, j' \text{ with } j = j' - 1; j = 1, \ldots, J - 1. \qquad (9)$$

The following examples refer to various more lenient or less strict decision rules which rely on the disjunctive connection and different numbers of testing instances (pairs of means) to be considered. One more lenient decision rule demands an increasing rank order not among *all* adjacent means ($j = j' - 1$), but for *at least two* adjacent means. Combining this decision rule with the *SH-mon* results in the *SP-mon2*, for which the number of pairs is $J - 1$:

$$\textit{SP-mon2:}\ \mu_j < \mu_{j'} \text{ for } \textit{at least} \text{ one pair } j, j' \text{ with } j = j' - 1; j = 1, \ldots, J - 1. \qquad (10)$$

The most lenient decision rule allows the consideration of all possible $\binom{J}{2}$ pairs of means to find at least one pair of means which conforms with the hypothesis, leading to the

$$\textit{SP-mon3:}\ \mu_j < \mu_{j'} \text{ for } \textit{at least} \text{ one pair } j, j' \text{ with } j \neq j'; j = 1, \ldots, J - 1. \qquad (11)$$

The lenient decision rules just applied do not address the question of how to handle rank inversions, which can occur with those pairs of means which are not in accordance with the predictions. If one or more, but not all pairs are judged to be different according to predictions, the remainder of the pairs can consist of homogeneous means ($\mu_j = \mu_{j'}$; equality of ranks) and/or of pairs of means with inverted rank orders ($\mu_j > \mu_{j'}$; inversion of ranks). This point will be discussed further below. For the time being it suffices to keep the possibility of rank inversions in mind.

The next question is, which of the hypotheses discussed so far are tested by using some tests proposed in the literature which, by the way, are only rarely addressed in standard textbooks (for an exception, see Bortz, 1993).

3.2 Some testing strategies aiming at monotonic trends

A variety of testing strategies come in question when considering statistical hypotheses about monotonic trends. The application of a global F test of analysis of variance, followed by a differential and *data-based* interpretation seems to be the most widespread. As has been argued above, this procedure is problematic, however, as the differential interpretation is not in accordance with the acceptance of the H_1 of the global F test and besides, an uncontrollable inflation of statistical error probabilities will occur. Thus the probability of a wrong decision in favor of strict monotonicity can be significantly increased, exceeding the pre-chosen α by a substantial amount. Furthermore, this manner of testing a particular qualitative trend hypothesis does not fulfill the criteria of appropriateness (directional differences have been predicted, but the test refers to nondirectional differences) and exhaustiveness (directional differences between all or at least some means, but the F test can also come out significant, if a large difference is associated with an inversion of rank order).

Another procedure, which is sometimes recommended, is to perform an F test for the hypothesis about the *quantitative* trend that is formulated to match the *qualitative* trend of interest: In place of the hypothesis of a strictly monotonic trend the test refers to the respective linear component through the orthogonal polynomials (see, e.g., Levin & Marascuilo, 1972, pp. 372-373). However, as has been stated above, the test can turn out statistically significant even if one or more rank inversions occur and since it may remain insignificant even if the rank order of means meets the prediction, but the differences among the means are inhomogeneous. The latter case is in accordance to the (qualitative) trend hypothesis of interest, the former is not. Thus, the interpretation of the results is ambiguous with respect to the hypothesis of strict monotonicity. But what would the consequences be if Testing Strategy TS 1, outlined above, is applied?

Applying Testing Strategy TS 1, the *SH-mon* is decomposed into the testable conjunction of hypotheses '$H_{1,lin} \wedge H_{0,dev}$' from Expression (2). If both hypotheses are accepted the trend is strictly monotonic by implication, although it is not possible to infer that all distances are large enough to reach statistical significance if tested separately as has been chosen as an additional criterion above. If, on the other hand, the two tests lead to accepting one of the conjunctions of hypotheses '$H_{0,lin} \wedge H_{0,dev}$' or '$H_{0,lin} \wedge H_{1,dev}$', respectively, it can be concluded that there is *no* monotonic trend. Yet if the pattern of decisions is '$AH_{1,lin} \wedge AH_{1,dev}$' the 'presence' or 'absence' of a strictly monotonic trend cannot be inferred unambiguously and test-based, since deviations from linearity ($AH_{1,dev}$) can be caused either by unequal distances among increasing or decreasing ranks or by rank inversions across the J means. Unequal distances again are compatible with strict monotonicity, whereas inverted ranks are not. Additionally, it remains unclear again whether the demand is fulfilled that *adjacent* means differ significanctly for each pair of means. Overall, the interpretation of the outcomes of the respective tests are ambiguous with respect to strict monotonicity. Thus, Smith and Macdonald (1983, p. 3) conclude with respect to the procedure just outlined that these tests may be 'optimal,' 'when the true state of the world is a linear trend. When the intervals between successive ... (μ_j) are not equal or are not known (and this is very commonly the case in psychology) the linear trend procedure is suspect and alternatives need to be examined.'

The *method of orthogonal contrasts*, whether to be used following a significant F test or instead of it, is covered in all textbooks and is in frequent use. Therefore, the question arises whether a 'satisfactory' testing strategy can be devised for hypotheses about orthogonal contrasts, enabling a *test-based* decision about a strictly monotonic trend. Without going into the details it can be stated that a strict rank order across J means cannot be established without supplementing test-based propositions to a large degree with data-based ones (see Hager, 1992, pp. 365-368, for the details). For this reason, further alternatives to the quantitative trend tests described up to point are in demand. Another procedure consists of applying modified (quantitative) trend tests. The modification mainly concerns the choice of a set of 'optimum' contrast coefficients according to the proposals made by Abelson and Tukey (1963; see for an application Bortz, 1993, pp. 259-260). For comprehensive and comparative surveys of these and further tests see Berenson (1982) and Smith and Macdonald (1983). According to Barlow, Bartholomew, Bremner and Brunk (1972, p. 118, p. 194), Berenson (1982, p. 270), and Le (1987, p. 173), the hypotheses (H_0 and H_1^*) tested against each other by these and related tests are:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \ldots = \mu_J \\ H_1^* &: \mu_1 \leq \mu_2 \leq \ldots \leq \mu_J \text{ with } \textit{at least} \text{ one strict inequality.} \end{aligned} \tag{12}$$

The alternative H_1^* refers to a weak ordering of parameters, that is, to a weakly monotonic trend. But because of the way in which the respective test statistics are defined the alternative hypothesis, as it is given in (12), may not have been stated appropriately and should be replaced by the following one:

$$H_1 : \mu_j < \mu_{j'} \text{ for } \textit{at least one} \text{ pair of population means with } j < j'. \tag{13}$$

This formulation of the hypothesis takes the fact that the test results may turn up significant when only a *single* rank order of all $\binom{J}{2}$ pairs of means is in agreement with the expectation ($j < j'$) more adequately into account. The alternative hypothesis in expression (12), on the other hand, may suggest that this rank order always concerns *adjacent* means ($j = j' - 1$). Furthermore, it appears that the tests mentioned above do not even test against the alternative of a *weakly* monotonic trend, but rather against a more general (and less specific) class of alternatives which *allow for rank inversions*, thus putting the usefulness of talking of 'monotonic' relations in question (see Berenson, 1982). This has been shown for the non-parametric trend test devised by Jonckheere (1954) in some detail by Hager (1995, pp. 171-174) and by Bortz (1993, p. 260) for the test proposed by Johnson and Mehrotra (1971). These tests, therefore, reflect a rather weak (implicit) decision rule which among other things allows for rank inversions. This fact causes many researchers to add data-based decisions to their test results.

The numerous procedures of ordering and selection (see above) and several multistage procedures combining the parametric F test with a rank correlation (Chassan, 1960; Green & Nimmo-Smith, 1982; Macdonald & Smith, 1983) seem to test analoguous statistical hypotheses. At least, the statistical hypothesis of a *strictly* monotonic trend is addressed in neither case, as Macdonald and Smith (1983, p. 25) have pointed out, raising the question of further alternative procedures once again.

Only one of presumably various possibilities is described here, in which the testing of hypotheses about qualitative trends is interpreted as a problem of testing hypotheses by means of *planned a priori* or *focussed contrasts*, the manner usually advocated when examining hypotheses formulated in advance (see, among others, Kirk, 1982; Marascuilo & Levin, 1983, p. 337; Thompson, 1994), but rarely applied in psychological research practice. As to my own experience, one of the reasons for this may be that many reviewers demand overall tests which may be followed by a multiple comparison procedure. But Winer et al. (1991, p. 146) clearly state: 'A procedure which is appropriate for a series of planned ... [contrasts] is simply to carry out a series of t tests, where t is appropriately defined for the experimental design used' ['comparisons' replaced by 'contrasts']. But this method seems to be suspect to many researchers who rely on overall tests even if particular contrast hypotheses can be formulated in advance. Maybe a series of t tests is too simple a procedure to be 'scientific', even if there is good reason to perform them?

3.3 Testing strategies for monotonic and other qualitative trends

The psychological hypothesis to be examined may state that 'the amount of retrieval (dependent variable) increases with increasing values of imagery (the independent variable).' To test this hypothesis $J > 2$ levels of imagery and an observable dependent variable such as 'number of words correctly remembered' are chosen. Omitting the psychological prediction referring to the observable dependent variable and the design chosen, the statistical prediction is derived from the hypothesis in an adequate and exhaustive manner. This prediction refers to statistical concepts exclusively, and since the discussion is restricted to (population) means the resulting statistical prediction states a strictly monotonic trend among the J means. This prediction has been called *SP-mon* in (9). Since this statistical prediction cannot be tested appropriately and exhaustively by a single test, it is then decomposed into testable partial hypotheses about focussed pair contrasts. These partial hypotheses can be tested in a way that enables *unambiguous* (as far as test results are concerned) and *test-based* decisions concerning the statistical prediction and that avoids any inconsistencies stemming mainly from data-based inferences. 'To avoid inconsistencies' simply means: Ranks are only called 'different,' if there are 'significant differences' among the means according to the usual statistical criteria and tests applied, whereby 'usual tests' refers to any two-sample test, whether it is a t test or a multiple comparison procedure on pair contrasts, each with only one degree of freedom. Such rankings are test-based. The *SP-mon* has already been presented above, but is given here again:

$$\textit{SP-mon1:}\ \mu_j < \mu_{j'} \text{ for } \textit{all}\ j, j' \text{ with } j = j' - 1; j = 1, \ldots, J - 1. \tag{9}$$

This formulation of the statistical prediction suggests a decomposition into $Q = J - 1$ partial hypotheses $H_{1,q}$ for adjacent means, conjunctively combined:

SP-mon1:

$$\bigcap_{q=1}^{Q=J-1} H_{1,q} : (\mu_j < \mu_{j'})_q \text{ for } q = 1, \ldots, Q = J-1 \text{ and for } j = j'-1. \tag{14}$$

The $Q = J - 1$ directional partial hypotheses can be tested by one-sided t tests. If all Q tests come out according to predictions then a strict monotonic trend (without rank equalities and without rank inversions) can be inferred, at least within the limits of statistical error: The predicted ranks can unequivocally be assigned to the (empirical) means, symbolizing a 'significant' distance for each pair of means without, however, knowing the sizes of the distances. But this particular information is not necessary in respect to the psychological hypothesis, although it should be computed from the data at hand. Under the strict decision rule applied it is not necessary to consider all possible $\binom{J}{2}$ pair contrasts: If at least one partial prediction, each referring to one testing instance, does not show up, the *SP-mon1* should be rejected, and this decision cannot be modified by showing that there are significant differences between non-adjacent means.

The more experimental conditions that have been chosen, the greater the cumulation of error probabilities, but also the more severe the test of the psychological hypothesis, all other things being equal. The cumulation can be compensated for by an adequate adjustment, for example, by the Dunn-Bonferroni method or an improved version of it (see, e.g., Kirk, 1982, pp. 106-111; Westermann & Hager, 1986; Winer et al., 1991, pp. 158-166).

If the *SH-mon* in (8) is connected with a lenient decision rule the acceptance of *at least one* partial alternative out of $J-1$ partial hypotheses $(H_{1,q})$ suffices to accept the respective *SP-mon2* $(j = j'-1; j = 1, \ldots, J-1)$. The decomposition is the same as before, but the decision rule is different. This means that the *SP-mon2* is more easily accepted than the *SP-mon1*, but the test of the respective psychological hypothesis is less strict than with the *SP-mon1*. An even more lenient decision rule gives leave to 'look for' the one necessary conforming result among all $\binom{J}{2}$ possible pairs of means which leads to the *SP-mon3*. If tested according to this prediction the psychological hypothesis has an even less severe test to survive than if tested by the *SP-mon2*[2].

In the derivation and testing of the *SP-mon2* or the *SP-mon3* the problem of possible rank inversions has not been discussed. There are basically two options concerning rank inversions. First, they are accepted if they occur when testing the *SP-mon2* or the *SP-mon3*. Second, they are or at least a maximum number of them is exluded a priori by a corresponding extension of the decision rule. In this instance additional tests should be planned referring to these inversions. Let us return to the *SP-mon2* and extend its decision rule to handle possible rank inversions; this extension leads to the *SP-mon4*, which deliberately allows for a maximum of R_{max} rank inversions *a priori*, suggesting the following decomposition:

$$\left[\begin{array}{l} \left(\bigcup_{q=1}^{Q=J-1} H_{1,q} : (\mu_j < \mu_{j'})_q \text{ for } q = 1, \ldots, Q \text{ and } j = j'-1\right) \wedge \\ \left(\bigcup_{r=1}^{R_{min}} H_{1,r} : (\mu_j < \mu_{j'})_r \text{ for } j = j'-1 \text{ and } R_{min} = Q - R_{max}\right) \end{array}\right] \tag{15}$$

The total number of tests to be considered and planned is $T = Q + R_{min}$. The retention of one or more of the $H_{0,q} : (\mu_j \geq \mu_{j'})_q$ implies that either '$\mu_j = \mu_{j'}$' is true (indicating equality of ranks), or that '$\mu_j > \mu_{j'}$' holds, indicating a rank inversion. In order to enable

[2] Various lenient decision rules can also be considered when testing hypotheses about quantitative trends if more experimental conditions than the minimum number are examined, that is, if more than one testing instance can be defined. These testing instances can again be linked either conjunctively or disjunctively. The details cannot be considered here because of limited space (see Hager, 1992).

Table 2: Testing a statistical prediction concerning a monotonic trend (SP-mon) which is decomposed into testable a priori hypotheses based on the method of planned contrasts

Prediction	Decomposition	Number of pairs (tests)	Kind of decision rule
SP-mon1	$\mu_j < \mu_{j'}$ for *all* $j, j'; j = j' - 1$	$J - 1$ $(J - 1)$	strictest possible; no rank equalities or inversions accepted
SP-mon2	$\mu_j < \mu_{j'}$ for *at least one* pair $j, j'; j = j' - 1$	$J - 1$ $(J - 1)$	more lenient; rank inversions permitted
SP-mon3	$\mu_j < \mu_{j'}$ for *at least one* pair $j, j'; j < j'$	$J(J + 1)/2$ $[J(J + 1)/2]$	most lenient; rank inversions permitted
SP-mon4^a	$\mu_j < \mu_{j'}$ for *at least one* pair $j, j'; j < j'$ and further hypotheses to exclude rank inversions (see text)	$J(J + 1)/2$ (see text)	lenient; maximum number of rank inversions tolerated

Notes.[a] Various similar kinds of predictions differ with respect to the number of pairs or testing instances considered. See text for further details.

a *test-based* differentiation of the two possibilities, the tests on the corresponding partial null hypotheses $H_{0,r} : (\mu_j \leq \mu_{j'})_r$ should be performed for these pairs of means. If more than R_{min} of these tests come out significant, the *SP-mon4* should be rejected. This testing strategy is in accordance to the proposals made by Shaffer (1972, 1974).

The decision rule applied in the *SP-mon4* is more lenient than the one in the *SP-mon1*, but stricter than the one in the *SP-mon3*. Because of different numbers of pairs it is difficult to say whether the decision rule of the *SP-mon4* is stricter than the rule of the *SP-mon2*, but since the *SP-mon2* allows for $J - 2$ rank inversions at most, the *SP-mon4* will most probably lead to a more severe test of the psychological hypothesis. Further decision rules or criteria concerning the maximum number of rank inversions and/or the number of pairs of means to be considered can be additionally defined, but will not discussed here. The recommendations are summarized in Table 2.

3.4 Some further thoughts on qualitative trends

Analogous procedures can be devised for comparable qualitative psychological hypotheses. The details need not be specified here (see Hager, 1992; Hager & Hasselhorn, 1995). Nor will details on testing statistical predictions concerning bitonic or tritonic trends be presented here. A hypothesis like the Yerkes-Dodson law addressed above postulates to a bitonic trend. If one chooses $J = 5$ experimental conditions (degrees of motivation) the respective prediction may take the following form, where the μ's refer to some measure of performance:

$$\textit{SP-biton: } \mu_1 < \mu_2 < \mu_3 > \mu_4 > \mu_5. \tag{16}$$

These and other qualitative trends can also be tested using the method of planned contrasts, since the predictions always refer to certain rank orders, but not to any sizes of distances among means. Moreover, the method can additionally be used to achieve appropriate and exhaustive tests of other statistical predictions encountered in research practice (see the examples in Hager, 1995) such as:

$$\textit{SP: } \mu_1 = \mu_2 < \mu_3 = \mu_4 < \mu_5, \tag{17}$$

or for a two-factorial design with equal n's per cell:

$$\textit{SP: } \mu_{11} > \mu_{12} > \mu_{21} > \mu_{22}. \tag{18}$$

The examples considered here should suffice to demonstrate the *versatility of the method of planned or focussed contrasts* which can (and should) be applied in a multitude of different empirical situations where researchers wish to test psychological hypotheses by means of statistical ones. It leads to easy interpretations, which are exclusively test-based and need not be 'corrected' by data-based inferences. The cumulation of error probabilities can be compensated for by adjustments referring to the tests one has *actually* planned to perform or actually performed; adjustments do not refer to a fixed set of potentially relevant partial hypotheses as called for in numerous techniques of multiple comparisons. Furthermore, power analysis for them can be based on wide-spread tables like those presented by Cohen (1988). The versatility of the method offers the additional advantage that many empirical tests of psychological hypotheses can be handled successfully by applying a single statistical method, if the researcher derives her or his predictions carefully, appropriately, and exhaustively. In addition, this derivation should take into account the general postulate in choosing statistical tests: 'We generally prefer to carry out the minimum number of significance tests required to evaluate our theory' (Myers & Well, 1991, p. 216). With respect to this demand the criterion of exhaustiveness has been defined to assure that the tests indeed 'required' or necessary with respect to the statements (or empirical content) of the theory or hypotheses are carried out as well. In the parametric case considered here the tests refer to the same definition of a 'difference or distance among means,' whereas the more widely used F tests use a squared function of all distances among the means, this detail being responsible for the fact that several well-known techniques of multiple comparisons when applied after a significant F, do not necessarily lead to decisions in agreement with the overall result (see, e.g., Betz & Levin, 1982; Gabriel, 1969). As a consequence, the method of focussed contrasts employed without a preceding overall test is often (but not always; see the overview by Thompson, 1994) recommended, since the respective tests 'usually result in increased power and greater clarity of substantive interpretation,' as Rosnow and Rosenthal (1989, p. 1281) state.

Another consideration refers to power analysis for planned contrasts to test statistical predictions concerning monotonic (or other qualitative) trends. Power analysis enables the determination of the sample size necessary to detect population effect sizes with pre-chosen error probabilities. In the t test situation, the effect size δ is the standardized difference among two population means, the values of which have to be selected for each hypothesis on a pair contrast. Some authors (e.g., Bredenkamp, 1984) argue that specifying these values for each adjacent pair of population means implicitly leads to an upgrading of a monotonic trend to a strictly linear trend. This belief may lead to the recommendation that the monotonic trend may and should be statistically handled as linear trend (see above). Bredenkamp's argument, however, overlooks the fact that predicted effect sizes $\delta_{kk'}$ for a quantitative trend refer to *exact* values which are functionally dependent on the values of the quantitative independent variable. If at least one of these exact values is (substantially) larger than predicted the strict definition of linearity is violated. On the other hand, when dealing with qualitative trends effect sizes such as $\delta_{jj'}$ are *minimum* values, which cannot be predicted, but are chosen according to methodological or economic reasons (see above). If one or more population values $\delta_{jj'}$ are larger than prespecified, this would *not* disagree with the prediction of a particular qualitative trend as long as the other values are still large enough. Referring to the samples, the empirical effects $\delta_{jj'}$ for pairs of means must be large enough to reach statistical significance which, in turn, allows assignment of different ranks to the means. Thus, the choice of some minimum values for $\delta_{jj'}$ should not be interpreted as upgrading a monotonic trend, especially as this interpretation would violate the criterion of exhaustiveness: The statistical hypothesis then actually tested comprises more information than can be derived from the original psychological hypothesis referring to a qualitative variable and trend (see above).

Considering $J-1$ or $J(J-1)/2$ pair contrasts will always lead to contrasts which, considered as a whole, are not orthogonal to one another. Although Hays (1988, p. 415) demands that planned contrasts or comparisons have to be orthogonal, many other textbook authors do not share this opinion as 'after all, contrasts are tested because they are of psychological import, not because they are independent of each other. ... in many and perhaps most cases the contrasts of interest will not be orthogonal' (Myers, 1972, p. 362; see also Thompson, 1994; Winer et al., 1991). In addition, pair contrasts do not contain the complete information inherent in the sums of squares between in an analysis of variance on the same means. But as long as the only (statistical) information needed for examining a psychological hypothesis in a valid manner consists in knowing whether the means are in the predicted order or not, there is no need for further information. But if there is any interest in further information not (directly) related to the examination of the psychological hypothesis each additional test may be performed which is thought to deliver insightful information. But these additional tests should be separated from those which directly refer to the psychological hypothesis of interest.

4 Recommendations Summarized

Since basically there are only two situations which differ with respect to the psychological hypotheses considered, the recommendations can easily be summarized.

First: If the independent variable is quantitative and if the psychological hypothesis refers to a quantitative relation, one of the testing strategies proposed for quantitative trends should be applied, which usually consist of one test for the predicted trend and one or more further tests for the deviations from predictions. Both tests have to come out according to predictions to be able to accept the statistical prediction of a certain quantitative trend. More than two tests result, if there are more than one predicted trend and/or if there are additional questions. The testing strategies should be chosen or adapted with respect to these (or other) additional questions (see Table 1).

Second: If the quantitative nature of the independent variable is disregarded of or if the independent variable is qualitative, its levels can either be assigned ranks (as is the case with 'imagery') or different codes (categorical variable, as is the case with gender). In both instances, however, psychological hypotheses can refer to various kinds of qualitative trends. In this instance one of the testing strategies discussed for qualitative trends should be applied. The choice between the testing strategies should also the various decision rules take into account. The decision rule bears consequences with respect to different conceptualizations of qualitative trends: strict ordering, weak ordering, ordering with or without rank inversions, and so on. The more lenient the decision rule, the less strict the demands on the hypothesized trend will be (see Table 2).

Each testing strategy proposed in this article consists in decompositions of statistical predictions into more than one testable hypothesis. These decompositions, however, are neither necessary nor advisable if a statistical prediction is equivalent to a testable null or alternative hypothesis of an available test, but this case seems to occur only rarely when examining psychological hypotheses (Hager, 1992; Westermann & Hager, 1986).

5 Some Final Remarks

All other things being equal predictions of a, say, linear trend have a greater empirical content (as understood by Popper, 1980) than predictions of a trend that is 'only' monotonic, because linearity implies not only a particular rank order for parameter values but also precisely specifiable distances among them, whereas (strict) monotonicity only refers to the rank order without specifying any distances. From this perspective, the preference for quantitative trends is compatible with Popper's demand to focus on the theories and hypotheses with the greatest empirical content possible (Popper, 1980). Whether this demand is adequate for the domains of psychology has already been questioned elsewhere (Hager & Westermann, 1986). And Popper himself has pointed out, too, that conceptualizing theories which are more and more precise and more likely to be falsified empirically are not conducive to scientific progress (Popper, 1981, p. 244).

The repeated reference to Popper as a philosopher whose methodology is continuing a matter of debate should *not* be interpreted as meaning that the testing strategies discussed in the present paper serve the better *falsification* of substantive statements. My focus is on the examination of psychological hypotheses and whether the empirical data fit them or not. If a hypothesis is valid or 'true' the probability that it is 'confirmed' should be high, and if it is not valid or 'false' the probability that it is 'disconfirmed' should be high. The lower the validity of the study, in the sense used by Cook and Campbell (1979), and the poorer the correspondence between predictions derived from the hypothesis and the statistical methods and tests applied, i.e., the lower the hypothesis validity (cf. Wampold et al., 1991), the lower the probabilities of correct decisions concerning the psychological hypothesis will be, all other things being equal. Furthermore, the probabilities of correct decisions will be lowered if the derivation of (psychological and statistical) predictions do not take the empirical content of the hypothesis into full account, that is, if the predictions and statistical partial hypotheses are not derived appropriately and exhaustively. Since adequate explanations and descriptions of psychological phenomena can only be achieved through hypotheses and theories which have passed valid empirical tests successfully, there is no good reason to continually try to falsify these hypotheses, as strict falsificationists would demand. *The better and more general rule demands to plan and execute experiments in a way that gives hypotheses a good chance to be 'confirmed' if they are 'true' and that leads to a high probability of 'disconfirming' them if they are 'false'.* Overall, it can be said that correct decisions are more likely the higher the validity of the experiment (see also Westermann, 1988). Considerations like these also seem valid in the realm of 'applied' psychology: More is gained if one knows that an intervention program is effective than if one knows that it is not. Since intervention research, as one possible example, can also be designed as examining psychological hypotheses referring to effectiveness (see Hager, 1995), there is no great difference between testing hypothesis about phenomena in basic psychology and testing hypotheses in 'applied' or technological psychology, though hypotheses serve different aims in both realms and their theoretical background may be quite different. In both instances, however, predictions can and should be derived from them which refer to the same statistical constructs and which can be submitted to the same statistical testing strategies and tests. If the psychological hypotheses are precise, the same holds for the predictions, and if they are imprecise, also the predictions are less precise.

References

[1] Abelson, R.P. & Tukey, J.W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General analysis and the case of simple order. *Annals of Mathematical Statistics, 34*, 1347-1369.

[2] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. & Brunk, H.D. (1972). *Statistical inference under order restriction.* New York: Wiley.

[3] Berenson, M.L. (1982). A comparison of several k sample tests for ordered alternatives in completely randomized designs. *Psychometrika, 47*, 265-280.

[4] Betz, M.A. & Levin, J.R. (1982). Coherent analysis of variance hypotheses testing strategies: a general model. *Journal of Educational Statistics, 3*, 121-143.

[5] Bortz, J. (1993). *Statistik f"ur Sozialwissenschaftler* (4th ed.) [Statistics for the social sciences.]. Berlin: Springer

[6] Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung* [The test of significance in psychological research]. Wiesbaden: Akademische Verlagsgesellschaft.

[7] Bredenkamp, J. (1980). Theorie und Planung psychologischer Experimente [Theory and design of psychological experiments]. Darmstadt: Steinkopff.

[8] Bredenkamp, J. (1984). Anmerkungen und Korrekturen zu Hager & Westermann: Entscheidung "uber wissenschaftliche und statistische Hypothesen: Probleme bei mehrfachen Signifikanztests zur Pr"ufung *einer* wissenschaftlichen Hypothese [Comments and corrections to Hager & Westermann: Decisions on psychological and statistical hypotheses: Problems with multiple tests of significance when examining a single psychological hypothesis]. *Zeitschrift f"ur Sozialpsychologie, 15*, 224-229.

[9] Chassan, J.B. (1960). On a test for order. *Biometrics, 16*, 119-121.

[10] Clark, C.A. (1963). Hypothesis testing in relation to statistical methodology. *Review of Educational Research, 33*, 455-473.

[11] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

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

[13] Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation. Design and analysis issues for field settings.* Boston, MA: Houghton Mifflin.

[14] Dykstra, R.L., Robertson, T & Wright, F.T. (Eds.). (1986). *Advances in order restricted statistical inference.* New York: Springer.

[15] Edwards, A.L. (1985). *Experimental design in psychological research* (5th ed.). New York: Harper & Row.

[16] Edwards, A.L. & Edwards, L.K. (1994). Analysis of variance and the general linear model. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 3. pp. 29-75). Greenwich, CT: JAI.

[17] Fisher, R.A. (1966). *The design of experiments* (8th ed.). Edinburgh: Oliver & Boyd.

[18] Gabriel, K.R. (1969). Simultaneous test procedures - some theory of multiple comparisons. *Annals of Mathematical Statistics, 40*, 224-250.

[19] Gigerenzer, G. (1993). The super-ego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.

[20] Green, T.R.G. & Nimmo-Smith, I. (1982). 2). 'Outcome-counting' - significance tests from incomplete predictions of order. *British Journal of Psychology, 73*, 41-49.

[21] Hager, W. (1987). Grundlagen einer Versuchsplanung zur Pr"ufung empirischer Hypothesen in der Psychologie [Essentials of experimental design for examining psychological hypotheses]. In G. L"uer (Ed.), *Allgemeine experimentelle Psychologie* (pp. 43-264). Stuttgart: Fischer.

[22] Hager, W. (1992). *Jenseits von Experiment und Quasi-Experiment. Zur Struktur psychologischer Versuche und zur Ableitung von Vorhersagen* [Beyond experiment and quasi-experiment: The structure of psychological experiments and the derivation of predictions]. G"ottingen: Hogrefe.

[23] Hager, W. (1993). Teststrategien bei der Pr"ufung von quantitativen psychologischen Hypothesen: Das Beispiel der Gesamtlernzeit beim Lernen von Texten [Testing strategies when examining quantitative psychological hypotheses: The case of total learning time with texts]. *Zeitschrift f"ur experimentelle und angewandte Psychologie, 40*, 509-547.

[24] Hager, W. (1995). Planung und Durchf"uhrung der Evaluation von kognitiven F"orderprogrammen [Designing and executing evaluations of cognitive programs]. In W. Hager (Ed.), *Programme zur F"orderung des Denkens bei Kindern. Konzeption, Evaluation und Metaevaluation* (pp. 100-206). G"ottingen: Hogrefe.

[25] Hager, W. & Hasselhorn, M. (1995). Testing psychological hypotheses addressing two independent and one dependent variables. *Perceptual and Motor Skills, 81*, 1171-1182.

[26] Hager, W. & Westermann, R. (1983). Zur Wahl und Pr"ufung statistischer Hypothesen in psychologischen Untersuchungen [On choosing and testing statistical hypotheses in psychological experiments]. *Zeitschrift f"ur experimentelle und angewandte Psychologie, 30*, 67-94

[27] Hager, W. & Westermann, R. (1986). Zur Wirkungsweise von Zielvorgaben beim Lernen aus Texten. Experimentelle Pr"ufung zweier konkurrierender Hypothesen [The effects of advance organizers when learning from texts: Experimental tests of two rival hypotheses]. *Psychologie in Erziehung und Unterricht, 33,* 17-25.

[28] Hays, W.L. (1988). *Statistics* (4th ed.). Fort Worth: Holt, Rinehart & Winston.

[29] Hochberg, Y. & Tamhane, A.C. (1987). *Multiple comparison procedures.* New York: Wiley.

[30] Howell, D.C. (1992). *Statistical methods for psychology* (3rd ed.). Belmont, CA: Duxbury.

[31] Johnson, R.A. & Mehrotra, K.G. (1971). Some *c*-sample nonparametric tests for ordered alternatives. *Journal of the Indian Statistical Association, 9*, 8-23.

[32] Jonckheere, A.R. (1954). A distribution-free *k*-sample test against ordered alternatives. *Biometrika, 41*, 133-145.

[33] Keppel, G. (1973). *Design and analysis.* Englewood Cliffs, NJ: Prentice-Hall.

[34] Keppel, G. & Saufley, W.H., jr. (1980). *Introduction to design and analysis.* San Francisco, CA: W.H. Freeman.

[35] Keppel, G. & Zedeck, S. (1989). *Data analysis for research designs.* New York: W.H. Freeman.

[36] Kirk, R.E. (1982). *Experimental design* (2nd ed.). Belmont, CA: Brooks/Cole.

[37] Kirk, R.E. (1994). Choosing a multiple-comparison procedure. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 3. pp. 77-121). Greenwich, CT: JAI.

[38] Le, C.T. (1987). On testing a trend in means in oneway layout. *Biometrical Journal, 29*, 173-180.

[39] Lee, W. (1975). *Experimental design and analysis.* San Francisco, CA: W.H. Freeman.

[40] Levin, J.R. & Marascuilo, L.A. (1972). Type IV errors and interactions. *Psychological Bulletin, 78*, 368-374.

[41] Lovie, A.D. (1986). Ranking and selection of populations. In A.D. Lovie (Ed.), *New developments in statistics for psychology and the social sciences* (pp. 143-162). London: British Psychological Society and Methuen.

[42] Macdonald, R.R. & Smith, P.T. (1983). Testing for differences between means with ordered hypotheses. *British Journal of Mathematical and Statistical Psychology, 36*, 22-35.

[43] Marascuilo, L.A. & Levin, J.R. (1983). *Multivariate statistics in the social sciences.* Monterey, CA: Brooks/Cole.

[44] Marascuilo, L.A. & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences.* Monterey, CA: Brooks/Cole.

[45] Maxwell, S.E. & Delaney, H.D. (1990). *Designing experiments and analyzing data.* Belmont, CA: Wadsworth.

[46] Meddis, R. (1973). *Elementary analysis of variance for the behavioral sciences.* London: McGraw-Hill.

[47] Meehl, P.E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

[48] Miller, R.G., jr. (1981). *Simultaneous statistical inference* (2nd ed.). New York: Springer.

[49] Morgenstern, D. (1980). Berechnung des maximalen Signifikanzniveaus des Tests "Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung f"uhren" [Determining the maximum level of significance for the test 'Reject H_0, if k of n tests lead to rejection']. *Metrika, 27*, 285-286.

[50] Myers, J.L. (1972). *Fundamentals in experimental design* (2nd ed.). Boston: Allyn & Bacon.

[51] Myers, J.L. & Well, A.D. (1991). *Research design and statistical analysis.* New York: HarperCollins.

[52] Paivio, A. (1986). *Mental representations. A dual coding approach.* New York: Oxford University Press.

[53] Popper, K.R. (1980). *The logic of scientific discovery* (10th ed.). London: Hutchinson.

[54] Popper, K.R. (1981). *Conjectures and refutations* (4th ed.). London: Routledge and Kegan Paul.

[55] Ramsey, P.H. (1980). Choosing the most powerful pairwise multiple comparison procedure in multivariate analysis of variance. *Journal of Applied Psychology, 65*, 317-326.

[56] Robertson, T., Wright, F.T. & Dykstra, R.L. (1988). *Order restricted statistical inference.* Chichester: Wiley

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

[58] Serlin, R.C. & Lapsley, D.K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199-228). Hillsdale, NJ: Erlbaum.

[59] Shaffer, J.P. (1972). Directional statistical hypotheses and comparisons among means. *Psychological Bulletin, 77*, 195-197.

[60] Shaffer, J.P. (1974). Bidirectional unbiased procedures. *Journal of the American Statistical Association, 69*, 437-439.

[61] Smith, P.T. & Macdonald, R.R. (1983). Methods for incorporating ordinal information into analysis of variance: Generalizations of one-tailed tests. *British Journal of Mathematical and Statistical Psychology, 36*, 1-21.

[62] Thompson, B. (1994). Planned versus unplanned and orthogonal versus nonorthogonal contrasts: the neo-classical perspective. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 3, pp. 3-27). Greenwich, CT: JAI.

[63] Wampold, B.E., Davis, B. & Good, R.H. III. (1990). Methodological contributions to clinical research: Hypothesis validity of clinical research. *Journal of Consulting and Clinical Psychology, 58*, 360-367.

[64] Westermann, R. (1988). Structuralist reconstruction of psychological research: Cognitive dissonance. *The German Journal of Psychology, 12*, 218-231.

[65] Westermann, R. & Hager, W. (1983). On severe tests of trend hypotheses in psychology. *The Psychological Record, 33*, 201-211.

[66] Westermann, R. & Hager, W. (1986). Error probabilities in educational and psychological research. *Journal of Educational Statistics, 11*, 117-146.

[67] Wilcox, R.R. (1987). *New statistical procedures for the social sciences.* Hillsdale, NJ: Erlbaum.

[68] Winer, B.J., Brown, D.R. & Michels, K.M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

Submitted April 23, 1996
Revised July 24, 1996
Accepted August 2, 1996