The examination of psychological hypotheses by planned contrasts referring to two-factor interactions in fixed-effects ANOVA

Willi Hager

Abstract

When interested in examining psychological hypotheses preceding data gathering especially in designs with more than two groups the standard ANOVA approach does not represent a very satisfying testing strategy. In these cases the statistical hypotheses actually tested are linked only loosely to the psychological hypothesis under scrutiny.

In order to link both kinds of hypotheses more closely, the statistical hypotheses should be derived from the psychological one. These derived statistical hypotheses refer to contrasts \textit{a priori}, which are constructed in order to conform with the predictions. Accordingly, the method of planned contrasts as a versatile alternative avoiding the shortcomings of the common ANOVA approach is chosen, and its advantages for testing \textit{a priori} hypotheses are demonstrated. Examples are provided to show how to examine psychological hypotheses by means of the proposed testing strategy where these hypotheses refer to two independent (e.g. age and an experimental variable) and one (or more) dependent variable(s) and enabling predictions within and across age groups.

**Keywords:** Examination of psychological hypotheses, two-factor interaction in fixed effects ANOVA, planned contrasts.

---

1 Willi Hager is professor at Georg-Elias-Mueller-Institute for Psychology, University of Goettingen, Gosslerstrasse 14, D-37073 Goettingen, Germany. He specializes in methodology and program evaluation.
1. **Introduction**

The problem of how to deal adequately with two-factor and/or higher order interactions in behavioral, psychological, and in educational research over years has been attacked from an plentitude of viewpoints and more or less successfully. Recently, Boik (1993) has analyzed two-factor interactions in fixed effects analysis of variance (ANOVA) and covariance (ANCOVA) and recommends the use of maximum $F$ statistics for analyzing product or interaction contrasts (see also Rosnow & Rosenthal, 1996).

Boik as well as nearly all other authors dealing with interactions in ANOVA and ANCOVA argue from the standpoint of analyzing given data, and their answers to the problems posed by statistical interactions can - strictly speaking - only refer to this situation. This "analysis point of view," although it may be called a successful research strategy which is used by many, if not by most researchers, will not be adopted here. For it is not the only research strategy, another one being the "hypothesis testing approach." This name, however, does not stand for a unified research strategy but instead it denotes a bundle of approaches which will not be discussed here.

Subsequently, mainly one problem will be attacked, namely, "How to test statistical predictions concerning certain patterns of data adequately and exhaustively in the two-factor case?"

2. **Examination of psychological and other behavioral hypotheses by means of predictions derived from them**

Although some of our colleagues will deny, virtually any psychological examination begins with an expectation or an hypothesis: "In planning an experiment, a researcher usually has in mind a specific set of hypotheses that the experiment is designed to test" (Kirk, 1994, p. 81); or as Howell (1992, p. 208) asks: "Do we really not know, at least vaguely, what will happen in our experiments; if not, why are we running them?" Or:

In terms of scientific progress, any statistical analysis whose purpose is not determined by theory, whose hypotheses and methods are not theoretically specified, or whose results are not related back to theory must be considered, like atheoretical fishing and model building, to be hobbies. (Serlin, 1987, p. 371)
The hypothesis may or may not be linked to a theory, and the expectation may be nothing more than an educated guess, or a "wild" presupposition, a "conjecture," "or what you will" (Popper, 1992, p. 32). The researchers' expectations very often are not made explicit, but in either case they concern a possible result of an investigation. In the realm of behavioral or educational psychology, these conjectures may be called behavioral or educational hypotheses. I'll use the name “psychological hypotheses” (PH) to stand for all kinds of substantive hypotheses. These refer to behavioral or psychological constructs, i.e. non-observable variables (such as "motivation," "amount of learning," or "academic achievement"). In order to test these hypotheses predictions have to be derived from them that refer to observable variables (such as "scores in a motivation questionnaire", "number of points in a learning test," and "scores in an academic achievement test") and to a particular layout or (experimental) design (e.g. "a two-factorial layout"). The psychological hypothesis then tells the researcher how many observable variables (e.g., factors in ANOVA) and how many experimental conditions per factor he or she has to choose at least in order to set up the minimum requirements for a valid test of the hypothesis.

These minimum requirements, that depend on the number of factors, the number of independent variables and the type of relationship claimed to hold between the independent and the dependent variables and that form necessary (though not sufficient) conditions for a valid test of the psychological hypothesis, may be called "testing instance" for a psychological hypothesis (Hager, 1992a, 1992b). A testing instance consists of four cells in ANOVA interaction.

Another type of prediction then refers to statistical constructs such as "expected means (of normally distributed random variables)," "correlations," or whatever parameters and statistical distributions one may prefer to deal with. Statements concerning values of these parameters are called statistical hypotheses. This kind of hypothesis should be sharply distinguished from the psychological (or substantive) hypotheses mentioned above. This distinction is known in education and psychology (see Bolles, 1962; Clark, 1963; Cohen, 1990; Meehl, 1967, 1978) and in statistics (Kendall & Stuart, 1963, p. 161), but its consequences are most often neglected. For the present purposes the two kinds of hypotheses are separated according to Clark (1963, p. 456-457): "Statistical hypotheses concern the behavior of observable random variables, whereas ... [psychological] hypotheses treat the phenomena of nature and man" ("scientific" replaced by "psychological").
As has been said, statistical hypotheses can serve as predictions from psychological hypotheses. Among others, especially Meehl (1967) has discussed the question how to connect the two kinds of hypotheses. He favors to relate them by means of *derivation*:

The statistical hypothesis is to be derived from the psychological one, "even in a rather loose sense of 'derive'" (p. 106-107). Allowing for "a rather loose sense of 'derive'," however, will lead to statistical hypotheses that are not connected "optimally" to the psychological ones. This means that they do not express the "empirical content" of the psychological hypothesis (Lakatos, 1970; Popper, 1992) "adequately" and/or "exhaustively." Although these criteria can only be evaluated with respect to the psychological and the statistical operationalizations of the psychological constructs and to a pre-chosen experimental design, it can be stated in general that "adequateness" refers to formulating a directional or non-directional statistical hypothesis with respect to the psychological one, which either is directional or non-directional, too. "Exhaustivity," on the other hand, refers to the number and the type of relationship(s) addressed by the psychological hypothesis and represented by the statistical one (see below). Both criteria can be interpreted to guarantee "hypothesis validity" in the sense given by Wampold, Davis, and Good (1990).

These criteria have been connected with Meehl's idea in order to match a demand which has been advocated by the statistician R. A. Fisher (1966, p. 15-16) as well as by the philosopher K. R. Popper (1992, p. 86), namely to *unequivocally* identify all those results and patterns of results that agree with a (psychological) hypothesis and those that disagree. In order for a statistical analysis to be relevant to a psychological hypothesis this unambiguous identification of two sets of results with opposite meanings must not be sacrificed by using inappropriate statistical tests or hypotheses, respectively.

In short, if a directional psychological hypothesis is tested by means of a non-directional statistical hypothesis the derivation is not *adequate*. If the psychological hypothesis leads to predicting a monotonic trend across \( K \) expected means and if this prediction is tested by the ANOVA \( F \) test the "implicit" derivation is not adequate, since the statistical alternative hypothesis \( H_1 \) actually tested by this test is (Capital \( K \) symbolizes the number of treatment conditions, whereas lowercase \( k \) represents one of these conditions):

\[
H_1(F): \mu_k \neq \mu_{k'}, \text{ for at least one pair of means } (k \neq k', k = 1, \ldots, K), \tag{1}
\]
whereas the hypothesis or statistical prediction (SP) of a strictly monotonic trend (SMT) may be formulated as follows:

\[ \text{SP(SMT): } \mu_{k-1} < \mu_k \text{ for all pairs of means, } k = 1, \ldots, K. \] (2)

The criterion of exhaustivity will be addressed soon. The particular hypotheses called null \((H_0)\) and alternative hypothesis \((H_1)\) are used for those hypotheses which (usually) are complementary to form a pair of hypotheses that is tested by a certain test. Thus, the \(H_0(F)\) complementary to the \(H_1(F)\) above states:

\[ H_0(F): \mu_k = \mu_{k'} \text{ for each pair of means } (k \neq k', k = 1, \ldots, K). \] (3)

It will not alter any of the subsequent considerations if your prefer another terminology; but I will base my argument on the foregoing distinction (see also Westermann & Hager, 1986).

Focussing on univariate psychological hypotheses leads to using \(F\) tests of univariate ANOVA, and these tests can be considered the most widely used statistical tests in psychology. They are sometimes attacked from two points of view: First, they are assumed not be appropriate with respect to the "underlying multivariate nature" of psychology which clearly asks for multivariate analysis strategies (see, among others, Bock, 1975, p. xii; Borgen & Seling, 1978; Hubble, 1984). Second, they are assumed not to be appropriate in case of more than two factor levels, for example, when it should be replaced by the procedure of planned comparisons (Hale, 1977; Huberty & Morris, 1988).

In this article, I am not going to deal with the first objection against univariate ANOVA \(F\) tests since its evaluation depends on whether you believe that there is a "real" nature of psychology or if you think the nature of psychology depends on how scientists conceive of their discipline. This may lead either to a multivariate or to a univariate conceptualization (cf. Marascuilo & Levin, 1983, p. 2). To take psychology's alleged multivariate nature into account would mean to formulate and examine multivariate psychological hypotheses that refer to more than one dependent variable and the relationships between these. This type of hypothesis seems extremely rare at the time being.

For this reason I focus on univariate psychological hypotheses and univariate statistical hypotheses that have been derived from the psychological hypothesis and that are tested by means of statistical tests which are appropriate to the hypotheses stated \textit{a priori}. In particular, I consider a type of psychological hypothesis which can be
said to be of ubiquitous interest to researchers, that is, hypotheses addressing two independent variables and their combined effect on one dependent variable.

I shall not consider the details when deriving a psychological prediction from the psychological hypothesis or when deriving a statistical prediction from the psychological one, although both steps are very important. For brevity's sake, I focus attention on deriving from a statistical prediction testable statistical hypotheses which refer to expected means.

Stressing the primacy of psychological hypotheses and their examination means that the analytical method of planned contrasts should be used instead of the ANOVA $F$ test (Hale, 1977; Hays, 1988; Hertzog & Rovine, 1985; Myers, 1972; Thompson, 1994; Wilcox, 1986) and irrespective of the question whether the contrasts are orthogonal to each other or not (Thompson, 1994; Winer, 1971). My recommendations concerning formulation of adequate and exhaustive hypotheses and tests are based on this particular analytical method.

This method, however, is not new. What is hoped to be new are some consequences of this approach when it is applied to more complicated cases than usually dealt with, that nonetheless are very often encountered in empirical research. In addition, the method of planned contrast is associated with a simultaneous testing strategy as opposed to a sequential one. The results achieved by simultaneous testing strategies are the same irrespective of the order of tests, whereas the results of sequential strategies may depend on a certain ordering of tests. Thus, a simultaneous testing strategy is preferable to sequential ones when testing psychological hypotheses (see Hager, 1992a, for a comprehensive review of these and other testing strategies).

3. Two-factorial psychological hypotheses.

As the first example I choose an experiment by Bjorklund and Zeman (1982). They asked children of three age groups (first, third, and fifth grade) to recall the names of their classmates ("class recall"; $A_1$) or to learn a lists of categorically related words ("free recall"; $A_2$). Amount of recall was assessed by the percentage of correctly reproduced items and amount of clustering by the Adjusted Ratio of Clustering (ARC) of Roenker, Thompson, and Brown (1971). The data on the two dependent variables were analyzed separately, i.e. by means of separate univariate analyses, and only the ARC measure will be considered here. For this measure, the statistical analyses rested
on mean scores (p. 800-801). Following the standard $F$ test for task (Factor A), the authors applied the standard $F$ test for interaction, and then they performed a separate $F$ test for class recall and Newman-Keuls tests for free recall. Overall, the data did not completely agree with the authors' expectations.

Now let us next consider one psychological hypothesis that can be associated with the Experiment 1 by Bjorklund and Zeman (1982) and that condenses the information the authors give as the background for their experiments (p. 800-801). Based on these information I assume that they could have had the following psychological hypothesis PH-1 in mind: "Because of the demand structure of the task, especially children's expert knowledge about the material to be recalled, amount of class recall is independent of age, whereas amount of free recall depends on aspects of children's knowledge base that develop with age. This leads to better clustering performance with higher age levels."

After having chosen operationalizations for the psychological variables and a design, both of which constitute necessary antecedents for derivation, the next step is to derive a (set of) prediction(s). This set of predictions refers to the operationalized psychological variables (ARC scores in the example) and for this reason, it will be called a psychological prediction (PP-1). Usually, a psychological prediction consists of several partial predictions each of which usually refers to fewer treatment levels or treatment combinations than the complete design contains. One of the central questions in face of factorial designs then concerns the level of the design the predictions should be derived for. This question is answered by the type of psychological hypothesis.

The psychological hypothesis PH-1 alleged to the Bjorklund and Zeman (1982) Experiment 1 addresses two factors and their combined effects on the dependent variable; the PH-1 may be called a "two-factorial hypothesis." Whenever a hypothesis addresses "combined effects" the level of the design (here: level of cells or - alternatively - level of main effects) has to be identified that contains these combined effects. The treatment combinations are located at the cells of the design, so here cells are the conceptual units of interest. There is no need for considering the levels of the two main effects for the evaluation of the psychological hypothesis. In spite of this fact, it usually is possible to derive predictions for at least one of the main effects, but these predictions are "only" implied by the predictions for the cells whereas these implied predictions do not allow to construct the predictions for the cells (Betz & Levin, 1982; Hager, 1992a; Hager & Hasselhorn, 1995). For this reason, the predictions that can be derived for the
level of design called main effect(s) deserve no further consideration from the point of view taken here.

For the 2x3 design chosen by Bjorklund and Zeman and the dependent variable ARC scores, the set of predictions called PP-1 derived from the hypothesis PH-1 are those presented in Table 1.

**Table 1. Possible predictions concerning mean differences across age groups applied to the experiment by Bjorklund and Zeman (1982, p. 801) (prediction PP-1).**

<table>
<thead>
<tr>
<th>Factor A: Task</th>
<th>Factor B: Age group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st grade (B₁)</td>
</tr>
<tr>
<td>Class recall (A₁)</td>
<td>ARC₁₁ = ARC₁₂ = ARC₁₃</td>
</tr>
<tr>
<td>Free recall (A₂)</td>
<td>ARC₂₁ &lt; ARC₂₂ &lt; ARC₂₃</td>
</tr>
</tbody>
</table>

*Note. Homogeneity (equality) of the ARC scores across the age groups is predicted under class recall, "ARC₁₁ = ARC₁₂ = ARC₁₃". "ARC₂₁ < ARC₂₂ < ARC₂₃" meaning that "ARC₂₁ is smaller than ARC₂₂ and ARC₂₂ is smaller than ARC₂₃". Nothing is said, however, about the distances between these ARC scores across factor A that can be equal or unequal without contradicting the prediction. For the statistical prediction, the entries “ARC” are changed to expected mean values, μ.*

For the statistical analysis, the statistical counterpart of the psychological prediction has to be derived from the latter. In general, this derivation leads to a particular statistical hypothesis, and hypotheses of this kind no longer refer to psychological variables, but to statistical concepts, such as expected means or mean ARC scores. If, in addition, the statistical hypothesis derived from the psychological prediction constitutes an adequate and exhaustive derivation from the psychological prediction it will be called a "statistical prediction" (SP).

The main difference between the two types of predictions, which usually is also the sole difference, lies in the referents of the predictions: psychological variables on the one and statistical concepts on the other hand. The signs ">", "≠", "=" and so on that have been identified as connectors between treatment levels or combinations of treatment levels, respectively, must be the same in the statistical as well as in the psychological prediction. Both types of predictions encompass all those (patterns of) results that agree with the psychological hypothesis, but none that disagree (see also Westermann & Hager, 1986).
This particular type of statistical hypothesis then constitutes one step in deriving testable statistical hypotheses from a psychological one. If only two treatment conditions are considered, most of the statistical predictions you may conceive of are equivalent to either a statistical $H_0$ or a statistical $H_1$. This is not true, however, if more than two treatment conditions are involved, where it is the exception rather than the rule that the statistical prediction is identical to either an $H_0$ or an $H_1$.

In order to maintain the content of the (statistical) prediction or to be exhaustive, it has to be decomposed into testable (partial) hypotheses. There are three types of decompositions: First, the psychological hypothesis and prediction leads to a statistical prediction consisting of one or more alternative hypotheses ($H_{1,r}$); second, the psychological hypothesis is adequately and exhaustively examined by a statistical prediction to be decomposed into one or more null hypotheses ($H_{0,s}$); and a mixture of testable $H_{0,s}$ and $H_{1,r}$ is derived from the third type of psychological hypothesis and statistical prediction (see Westermann & Hager, 1986, for further details).

4. How to test statistical hypotheses concerning particular data patterns in the ANOVA model

But which statistical concepts the hypotheses to be derived should refer to? Although several other choices are possible, I restrict my considerations to the concepts connected with parametric ANOVA. When applying the ANOVA $F$ test for interaction, one can use the cell means model (see Timm & Carlson, 1975, p. 13-17):

The cell means model uses the model equation

$$\gamma_{ijk} = \mu_{jk} + \varepsilon_{ijk}$$  \hspace{1cm} (4)

and leads to the following interaction hypotheses:

$$H_0(AxB): \mu_{jk} - \mu_{j'k'} - \mu_{jk'} + \mu_{j'k} = 0 \text{ for all quadruples } <j,j',k,k'>;$$
$$j = 1,2 \text{ and } k = 1,2,3; \text{ and}$$

$$H_0(AxB): \mu_{jk} - \mu_{j'k'} - \mu_{jk'} + \mu_{j'k} \neq 0 \text{ for at least one quadruple } <j,j',k,k'>;$$
$$j = 1,2 \text{ and } k = 1,2,3. \hspace{1cm} (6)$$

$$H_0(AxB): \mu_{jk} - \mu_{j'k'} - \mu_{jk'} + \mu_{j'k} = 0 \text{ for all quadruples } <j,j',k,k'>;$$
$$j = 1,2 \text{ and } k = 1,2,3; \text{ and}$$

$$H_0(AxB): \mu_{jk} - \mu_{j'k'} - \mu_{jk'} + \mu_{j'k} \neq 0 \text{ for at least one quadruple } <j,j',k,k'>;$$
$$j = 1,2 \text{ and } k = 1,2,3. \hspace{1cm} (6)$$
Since Bjorklund and Zeman (1982) deal with mean ARC scores, the cell means model seems adequate. Thus, cell means are used as statistical equivalents or operationalizations of the (non-observable psychological) variable "clustering performance", as assessed by the (observable) ARC scores. This decision can and should not prevent other authors from using other statistical concepts, such as medians, probabilities, correlations, and so on.

By choosing the cell means model, however, the well known difficulty is avoided that the interaction terms $\gamma_{jk}$ of the alternative effect model usually have no meaningful psychological interpretation and - in addition - are hard to interpret from the very beginning (see Bernhardson, 1973, as well as the debate between Rosenthal & Rosnow, 1989, 1991, and Meyer, 1991). But does this decision not lead to the difficulties that have been discussed by the authors just mentioned?

Part of these difficulties results from the fact that the problem is viewed solely from the ANOVA perspective, which in the factorial case usually involves interactions and main effects. The above authors agree as far as the relevance of cell means is concerned, but they disagree with respect to the relationships between main effects and interactions. Undoubtedly, an answer to this question is important if one begins data analysis with ANOVA and then proceeds with identification of the type of interaction, namely ordinality and disordinality, which is essential for meaningful interpretations (see, for example, Boik, 1993; Hager & Westermann, 1983, and Kirk, 1982). But the problem of making sense of an interaction following ANOVA $F$ test does not arise here at all, since one "level of analysis" (that is, the cells of the design as opposed to the main effects) according to the psychological hypothesis of interest is chosen and adhered to, disregarding of the other levels. (If the psychological hypothesis addresses only one factor its examination should of course take place on the level of the respective main effect, such as age group.) Then the predictable relationships between cells is specified in a way as to optimally conform with the psychological hypothesis under scrutiny.

Next, those statistical hypotheses and tests are identified that allow for an adequate and exhaustive testing for this particular pattern.

If one substitutes the scores of the ARC variable of Table 1 by the (expected) mean of these scores, $\mu_{jk}$, the following statistical prediction SP-1 results, where "$\wedge$" symbolizes a conjunctive linkage (see below):

$$SP-1: (\mu_{11} = \mu_{12}) \wedge (\mu_{21} < \mu_{22}) \wedge (\mu_{12} = \mu_{13}) \wedge (\mu_{22} < \mu_{23}).$$

(7)
The decomposition of the SP is given in testable partial hypotheses, which form two testing instances, each consisting of a nondirectional null hypothesis and by a directional alternative hypothesis. The pattern predicted is one of statistical two-way interaction.

Let us consider the first quadruple of cell means which can be stated as 
\[
(\mu_{22} - \mu_{21} - \mu_{12} + \mu_{11})
\]
This expression can be re-arranged algebraically into:
\[
(\mu_{22} - \mu_{21} - \mu_{12} + \mu_{11}) = (\mu_{22} + \mu_{11}) - (\mu_{21} + \mu_{12}) = (8a)
\]
\[
(\mu_{22} - \mu_{21} - \mu_{12} + \mu_{11}) = (\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11}) = (8b)
\]
\[
(\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22}) = (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}). (8c)
\]
The four selected re-arrangements in (8b) and in (8c) represent *differences of mean differences* (or contrasts of contrasts), and this is one way to conceptualize and understand statistical first order interactions in ANOVA, but of course not the only one (see also Marascuilo & Levin, 1970, and Meyer, 1991). This interpretation of interactions and this formulation of what usually is called an "interaction contrast" (Meyer, 1991), however, will grossly simplify the further presentations and analyses. All formulations concerning the (expected) means are equivalent and again lead to identical tests and decisions - all other things being equal (see above). As will be shown below, there are also differences between the various re-arrangements.

Returning to the representation of the main aspects of the prediction in Table 1, an obvious decomposition of the prediction SP-1 is addressed which is adequate, but not exhaustive:
\[
SP - 1: \{[H_{13} : \psi_1 = (\mu_{22} - \mu_{21}) - (\mu_{12} - \mu_{11}) > 0] \land
[H_{14} : \psi_4 = (\mu_{23} - \mu_{22}) - (\mu_{13} - \mu_{12}) > 0]\}.
\]
This (incomplete) decomposition of the prediction SP-1 consists of a conjunction ("\land") of partial hypotheses ($H_{1,r}$). Partial hypotheses - much like partial predictions - refer to fewer parameters than the design incorporates; contrast hypotheses are special partial hypotheses (see also Betz & Levin, 1982). To connect these partial hypotheses by a conjunction in addition means that the whole conjunction called SP-1 will only be accepted if all its partial hypotheses can also be accepted. Using the conjunctive connection then is tantamount to applying a strict decision criterion with respect to
whole set called statistical prediction (SP) (see Westermann & Hager, 1986, for further details).

With respect to the mean ARC scores the part of the SP-1 says that the pairwise mean differences among age groups are greater for free recall than for class recall. Although \( H_1: \psi = (\mu_{23} - \mu_{21}) - (\mu_{13} - \mu_{11}) > 0 \) referring to non-adjacent age groups can be derived, too, this partial hypothesis can be omitted without any negative consequences, since it is redundant when using the conjunctive criterion of connection. For if at least two adjacent means do not differ significantly, the strict criterion cannot be fulfilled, and the respective statistical prediction leading to the partial prediction cannot be accepted. If, on the other hand, each two adjacent means differ significantly, the non-adjacent means must differ significantly, too, all other things being equal. Thus, comparing non-adjacent means is of no relevance for the decision about the SP-1.

Negative consequences, however, may arise from the fact that the conjunction of two partial hypotheses does not represent an exhaustive derivation although it is adequate: The representation in Table 1 clearly shows different partial predictions within free recall and within class recall. The lack in exhaustivity then is due to neglecting these particular predictions that are not represented by the \( H_{1,1} \) in (9). Non-exhaustive derivations like this can quite easily lead to test results that are not unequivocally interpretable with respect to the specific predictions.

As an example of this consider \( H_{1,1} \) just mentioned saying \( \psi_1 > 0 \). Apparently, \( \psi_1 > 0 \) arises not only when \( \psi_2 = \mu_{12} - \mu_{11} = 0 \) and when \( \psi_3 = \mu_{22} - \mu_{21} > 0 \) (as should be according to predictions), but also when, for example, \( \psi_3 = \mu_{22} - \mu_{21} < 0 \) and when in addition \( \psi_2 = \mu_{12} - \mu_{11} > 0 \) as well as \( |\psi_2| > |\psi_3| \) which is counter to predictions. Consider the fictitious, though possible mean values \( \mu_{11} = .51, \mu_{12} = .35, \mu_{21} = .16, \) and \( \mu_{22} = .06, \) which lead to \( \psi_1 = \psi_3 - \psi_2 = (.06 - .16) - (.35 - .51) = .06 > 0 \). In order to avoid statistically ambiguous results like the foregoing one, which may (erroneously) lead to a corroborating evaluation of the psychological hypothesis, it is advisable to derive and test some additional partial hypotheses concerning adjacent means across the age groups within class recall (A_1) and within free recall (A_2), respectively. If these partial hypotheses referring to simple mean differences are added to the above decomposition we get the adequate and exhaustive derivation from the psychological hypothesis PH-1 alleged to Bjorklund and Zeman (1982) and from the statistical prediction SP-1 (see also Table 1):
SP-1 \Rightarrow \left[ \left[ H_{14} : \psi_4 = (\mu_{23} - \mu_{22}) - (\mu_{13} - \mu_{12}) > 0 \right] \land \left[ H_{13} : \psi_3 = (\mu_{22} - \mu_{21}) > 0 \right] \land \left[ H_{15} : \psi_5 = (\mu_{13} - \mu_{12}) = 0 \right] \land \left[ H_{16} : \psi_6 = (\mu_{23} - \mu_{22}) > 0 \right] \right] \quad (10)

"\Rightarrow" symbolizes the adequate and exhaustive derivation of partial hypotheses. The two testing instances within the square brackets are linked conjunctively. Although the contrasts are not mutually orthogonal and although the validity of some hypotheses is strictly implied by the validity of others (see Betz & Levin, 1982, for the details), the decisions about the partial hypotheses are not completely determined in the sense that knowing all minus one decisions allows perfect prediction of the remaining one that has been chosen in advance. And even when using only one error term (\(MS_e\) of ANOVA) for all tests (which is the usual practice with multiple comparisons following significant overall test; Thompson, 1994) the six decisions will not be completely dependent from each other, which can easily be shown by some appropriate numeric examples (see Hager, 1987, 1992a, and Hager & Hasselhorn, 1995, for a demonstration of the difference between "strict" [stochastic] dependencies among the hypotheses and "weak" dependencies between the data-based decisions on the hypotheses.) Because of the weak dependencies among the decisions, none of the four one- and the two two-sided (\(t\)) tests should be omitted, for which the possible cumulation of error probabilities is not as high as one might presume, since an increase in dependency leads to a decrease in magnitude of cumulation of these probabilities (see Miller, 1981). Only testing of all derived hypotheses ensures that all decisions are based on tests and not on inspection of the empirical means (see also Huberty & Morris, 1988, p. 572). Decisions that are mainly or exclusively based on inspection of data violate the basic demand for applying testing strategies which are consistent in the sense that all decisions on statistical hypotheses or significance should be based on tests. (Thus, applying tests of multiple comparisons following a significant \(F\) for a main effect on the one hand and deciding on the type of interaction by inspection of a graph of the cell means following a significant \(F(AxB)\) on the other hand, represent two procedures that are not consistent with each other, though they are often applied in conjunction.)

One might take issue with the lack of independence of the several contrasts and hypotheses. But remember that the statistical partial hypotheses and the tests have been chosen solely on the basis of their meaningfulness with respect to the given
psychological hypothesis to examine: "In practice the comparisons that are constructed are those having some meaning in terms of the experimental variables; whether these comparisons are orthogonal or not makes little difference" (Winer, 1971, p. 175; see also Edwards, 1985, p. 151; Huberty & Morris, 1988; Kirk, 1982, p. 95, 106, 114, and Myers, 1972, p. 362). To put it into other words: In order to represent an exhaustive derivation of partial predictions even at the level of statistical tests, all tests on partial hypotheses should be performed. On the other hand, none of the standard tests of ANOVA (for the main effects and for the interaction) is necessary nor any other test not referring to a particular prediction.

In passing we note that the number of tests (and the degree of dependency) can be reduced by combining the two null hypotheses \( H_{0,2} \) and \( H_{0,7} \) to form the

\[
H_0(B \text{ within } A_1) : \mu_{11} = \mu_{12} = \mu_{13}.
\] (11)

Considering relationships between the several hypotheses, this \( H_0(B \text{ within } A_1) \) is equivalent to the two nulls \( H_{0,2} \) and \( H_{0,7} \), but testing the \( H_0(B \text{ within } A_1) \) may well lead to a different decision concerning means across \( B \) (grade level) within \( A_1 \) (class recall) than testing the two separate (partial) null hypotheses (see above). Although both ways of testing can be chosen since they represent adequate and exhaustive derivations, they also represent different testing strategies, and despite of the greater cumulation of error probabilities² I favor testing the separate nulls since this procedure is consistent with the separate hypotheses and separate tests for free recall \( (A_2) \).

In the example considered three testing instances each consisting of quadruples of means can be constructed, and two of them refer to adjacent grade levels \( (B_1 \text{ and } B_2; B_2 \text{ and } B_3) \) whereas the third uses non-adjacent grade levels \( (B_1 \text{ and } B_3) \). This third testing instance need not be considered when using the conjunctive operator in connecting the various instances, but it should be considered when using a disjunctive operator among the testing instances. The disjunctive operator reflects a lenient or benevolent criterion (see Westermann & Hager, 1986) in deciding on accepting or rejecting the statistical prediction and - in the end - also in deciding on corroboration or non-corroboration of the psychological hypothesis. Under this benevolent criterion the latter kind of hypothesis will be taken as corroborated even if the partial predictions

² I shall not deal with any problem arising from cumulation of error probabilities, since this has been done elsewhere (Hager, 1992a; Westermann & Hager, 1986).
turn out to hold only for the two extreme grade levels (first and fifth grade; B1 and B3). This type of decision, then, is the same that would have resulted if the experimenters had planned their experiment with only the two "extreme" grade levels B1 (first graders) and B3 (fifth graders) from the very beginning, for example, if they had chosen to execute their experiment with only the minimum number of treatment conditions or one testing instance.

Even if you prefer the benevolent criterion to the strict or severe one, represented by the conjunctive operator, it is not possible to use the disjunctive operator ("\(\lor\)") for the partial hypotheses that refer to a single testing instance, that is to four cells in the case of statistical interaction. For if you do so, the criterion of exhaustivity of derivation is violated; the reasons why have been discussed in some detail by Hager (1992a) and will not be repeated here.

It should be noted in addition that no prediction concerning relative magnitude of means with respect to the task factor (A) has been considered until now. For this reason, the psychological hypothesis leading to the (statistical) prediction SP-1 does not refer to partial predictions concerning mean differences among the tasks (A1 and A2). Accordingly, its acceptance will not be affected by any differences among the means for A1 and A2 (that is within each age group: \(\mu_{11} - \mu_{21}\), \(\mu_{12} - \mu_{22}\), and \(\mu_{13} - \mu_{23}\) ) be they positive or negative or (partly) null.

The pattern of prediction deduced from the Bjorklund and Zeman experiment can be less often encountered than a more typical one that also refers to differences of mean differences, where, however, each of the simple mean differences differs from zero. This type of pattern can be predicted, for example, for Experiment 1 of Schwanenflugel, Guth, and Bjorklund (1986) and in Hasselhorn (1992).

In the former study, subjects had to rate the importance of concept attributes, and the theoretical introduction preceding this experiment allows to predict that the differences in mean ratings of attributes classified as high or medium important for related concepts by adults increase with age or grade level, respectively (Schwanenflugel et al., 1986, p. 422-424). No prediction, however, seems possible with respect to relationships across different grade levels. For simplicity's sake, this type of prediction is applied to the Bjorklund and Zeman (1982) study, assuming however differences decreasing with grade levels in order to take into a better account the circumstances of
this study. These considerations lead to the subsequent statistical prediction SP-2 shown in Table 2:

\[ \text{SP-2 : } \{(\mu_{1k} - \mu_{2k}) > 0 \text{ for all } k, k = 1, 2, 3 \} \land [(\mu_{11} - \mu_{21}) > (\mu_{12} - \mu_{22}) > (\mu_{13} - \mu_{23})]\}. \quad (12) \]

Table 2. Possible predictions concerning mean differences within age groups applied to the experiment by Bjorklund and Zeman (1982, p. 801) (prediction SP-2).

<table>
<thead>
<tr>
<th>Factor A: Task</th>
<th>Factor B: Age group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st grade (B₁)</td>
</tr>
<tr>
<td>Class recall</td>
<td>( \mu_{11} )</td>
</tr>
<tr>
<td></td>
<td>( \lor )</td>
</tr>
<tr>
<td>Free recall</td>
<td>( \mu_{21} )</td>
</tr>
<tr>
<td></td>
<td>( \lor )</td>
</tr>
</tbody>
</table>

Note. A monotonic trend for mean differences for recall scores is predicted, that is mean differences for recall scores are expected to decrease with age: "(\(\mu_{11} - \mu_{21}\)) > (\(\mu_{12} - \mu_{22}\)) > (\(\mu_{13} - \mu_{23}\))". No partial predictions, however, are (held) possible for means or mean differences across grade levels.

The altered pattern of prediction has to be taken into account when deriving the statistical partial hypotheses. Applying the same principle of deriving adequate and exhaustive partial hypotheses from a statistical prediction (here: SP-2), we get:

\[ \text{SP-2 : } \{(H_{1,1} : \psi_1 = (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}) > 0) \land (H_{1,2} : \psi_2 = (\mu_{11} - \mu_{21}) > 0) \land \]
\[ (H_{1,3} : \psi_3 = (\mu_{12} - \mu_{22}) > 0) \} \land \{(H_{1,4} : \psi_4 = (\mu_{12} - \mu_{22}) - (\mu_{13} - \mu_{23}) > 0) \land \]
\[ (H_{1,5} : \psi_5 = (\mu_{13} - \mu_{23}) > 0) \}. \quad (13) \]

(The \(H_{1,3}\) is part of both testing instances or quadruples, but need only be tested once, of course. It has been listed with the partial hypotheses of the second testing instance in order to demonstrate the construction of each such instance.) It may be enlightening to compare this decomposition of a statistical prediction (SP-2) with the respective decomposition of the SP-1 above.

Because of applying the conjunctive connector it is not necessary to derive partial hypotheses for the possible third testing instance referring to non-adjacent means. But in cases like the foregoing one the experimenters might prefer to choose the disjunctive criterion for connecting the three instances which leads to accepting the statistical
prediction even if the partial hypotheses of only one testing instance can be accepted. This type of criterion then is of interest when the researchers operate with "adjacent" age groups, namely first grade, second grade, third grade, and so on. For in this case mean ages might be too close together to demonstrate the expected effects for adjacent grade levels, whereas these effects can easily be demonstrated for at least one pair of non-adjacent grade levels. Remember, by the way, that Bjorklund and Zeman (1982) chose the first, third, and fifth grade levels and omitted the second and the fourth grade levels - maybe for reasons similar to the one just sketched.

This leads to the following derivation ("∨" signifies the disjunctive operator):

\[
\text{SP-2} \Rightarrow \left[ \left[ H_{1,1} : \psi_1 = (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}) > 0 \right] \land \left[ H_{1,2} : \psi_2 = (\mu_{11} - \mu_{21}) > 0 \right] \right] \land \left[ H_{1,3} : \psi_3 = (\mu_{13} - \mu_{23}) > 0 \right] \land \left[ H_{1,4} : \psi_4 = (\mu_{21} - \mu_{12}) - (\mu_{23} - \mu_{13}) > 0 \right] \land \left[ H_{1,5} : \psi_5 = (\mu_{11} - \mu_{21}) - (\mu_{13} - \mu_{23}) > 0 \right] \land \left[ H_{1,7} : \psi_7 = (\mu_{11} - \mu_{21}) - (\mu_{13} - \mu_{23}) > 0 \right] \land \left[ H_{1,6} : \psi_6 = (\mu_{13} - \mu_{23}) > 0 \right] \land \left[ H_{1,8} : \psi_8 = (\mu_{21} - \mu_{12}) - (\mu_{23} - \mu_{13}) > 0 \right] \land \left[ H_{1,9} : \psi_9 = (\mu_{21} - \mu_{12}) - (\mu_{23} - \mu_{13}) > 0 \right].
\]

Bjorklund and Zeman (1982, p. 801) did test for differences for the task factor, too (F test for main effect of task). Let us assume that they had a psychological hypothesis PH-3 in mind that enables to derive the partial predictions of both the SP-1 and the SP-2. The statistical prediction SP-3 then takes this form (see also Table 3):

\[
\text{SP-3} : \{(\mu_{11} = \mu_{12} = \mu_{13}) \land (\mu_{21} < \mu_{22} < \mu_{23}) \land (\mu_{1k} - \mu_{2k}) > 0 \text{ for each } k; k = 1,2,3 \} \land \{(\mu_{11} - \mu_{21}) > (\mu_{12} - \mu_{22}) > (\mu_{13} - \mu_{23})\}.
\]

**Table 3.** Possible predictions concerning mean differences within and across age groups applied to the experiment by Bjorklund and Zeman (1982, p. 801) (prediction SP-3).

<table>
<thead>
<tr>
<th>Factor A: Task</th>
<th>Factor B: Age group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st grade (B₁)</td>
</tr>
<tr>
<td>Class recall (A₁)</td>
<td>µ₁₁ = µ₁₂ = µ₁₃</td>
</tr>
<tr>
<td></td>
<td>∨</td>
</tr>
<tr>
<td>Free recall (A₂)</td>
<td>µ₂₁ &lt; µ₂₂ &lt; µ₂₃</td>
</tr>
</tbody>
</table>

*Note.* The predictions SP-1 and SP-2 have been combined to form the SP-3.
In order to achieve adequate and exhaustive tests, the following partial hypotheses should be derived from the SP-3, again using the strict (conjunctive) criterion:

\[ \text{SP-3} \Rightarrow \{[H_{1,1} : \psi_1 = (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}) > 0] \land \\ [H_{1,2} : \psi_2 = (\mu_{11} - \mu_{21}) > 0] \land [H_{1,3} : \psi_3 = (\mu_{12} - \mu_{22}) > 0] \land \\ [H_{0,4} : \psi_4 = (\mu_{11} - \mu_{12}) = 0] \land [H_{1,5} : \psi_5 = \mu_{22} - \mu_{21} > 0] \land \\ \{[H_{1,6} : \psi_6 = (\mu_{12} - \mu_{22}) - (\mu_{13} - \mu_{23}) > 0] \land H_{1,3} : \psi_3 > 0) \land \\ (H_{1,7} : \psi_7 = (\mu_{13} - \mu_{23}) > 0] \land (H_{0,8} : \psi_8 = \mu_{13} - \mu_{12} = 0) \land \\ [H_{1,9} : \psi_9 = (\mu_{13} - \mu_{23}) > 0]) \} \]

16

This derivation of testable statistical hypotheses is adequate and exhaustive with respect to the new psychological hypothesis (PH-3) that is of course different from the PH-1 and the PH-2 considered above: It is more precise, and therefore it enables more partial predictions, i.e. the sum of partial prediction of both the PH-1 and the PH-2. This psychological hypothesis PH-3, then, is easier to disconfirm ("falsify") than the PH-1 and the PH-2 (Popper, 1992) and harder to corroborate.\(^3\)

The three examples shed a little light on the statement above according to which the various algebraic re-arrangements of the term "(\(\mu_{jk} - \mu_{j'k'} - \mu_{jk'} + \mu_{j'k}\))" lead to identical values of the contrasts and do not alter the statement of the respective partial hypotheses, but that there are differences, too. These differences concern the basic mean differences which for one type of prediction have to be formed within the rows (task; \(A_j\)).

This type of construction is represented best by the arrangement "(\(\mu_{jk} - \mu_{jk'} - \mu_{jk'} + \mu_{j'k}\))", leading to "(\(\psi_1 = (\mu_{22} - \mu_{21}) - (\mu_{12} - \mu_{11}) > 0\)" (see SP-1 above). For another type (SP-2) these differences have to be formed within the columns (grade levels; \(B_k\)), represented best by the expression "(\(\mu_{jk} - \mu_{j'k} - \mu_{jk'} + \mu_{j'k}\))", leading, for example, to "(\(\psi_1 = (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}) > 0\)." In still another type of pattern (SP-3) the mean differences have to be constructed within the rows as well as within the columns of the design. Here both of the foregoing representations are adequate, although only one

\(^3\) "Deciding on a psychological hypothesis" here means attaching the label "corroborated " or "not corroborated " to this hypothesis. This type of labelling agrees with "sophisticated falsificationism" as advocated by Lakatos (1970), although it does not imply any statement concerning the "truth" of a psychological hypothesis. Moreover, I am not interested in the "degree of confidence" into such an hypothesis which is another question.
of them will actually be considered because they always have the same value. Which way of constructing is more appropriate in a given context depends solely on the psychological hypothesis and the predictions preceding the formulation of the partial hypotheses.

5. Demonstration of the testing strategy

In this section, the proposed testing strategy is applied to the examination of the psychological hypothesis PH-1 associated with Experiment 1 of Bjorklund and Zeman (1982, p. 800-802) and to the hypothesis PH-2 leading to the statistical prediction SP-2. The original means given by these authors have been modified, and the data the following tests are based on are given in Table 4. $MS_e$ was recomputed to be .022 on the basis of $F(1,140) = 196.34$ and the relevant mean difference being .353, assuming, however, equal cell frequencies ($n = 23; N = 138$) and a between-subject variation for the task factor, whereas Bjorklund and Zeman (1982, p. 800-801) used within-subject variation of task with unequal $n$'s.

Table 4. Some partly fictitious data to demonstrate application of the tests proposed for testing the predictions SP-1 and SP-2.

<table>
<thead>
<tr>
<th>Factor A: Task</th>
<th>Factor B: Age group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st grade ($B_1$)</td>
</tr>
<tr>
<td>Class recall</td>
<td>$M_{11} = .51$</td>
</tr>
<tr>
<td>Free recall</td>
<td>$M_{21} = .10$</td>
</tr>
</tbody>
</table>

Note. Means for class recall scores have been taken from Bjorklund and Zeman (1982, p. 801), and means for free recall scores have been altered on the basis of the means reported by Bjorklund and Zeman.

The statistic chosen for testing the partial hypotheses is $t$, and the formula applied is the one Kirk (1982) gives on p. 96. The cumulation of error probabilities ($\alpha$ and $\beta$) is disregarded; $\alpha = .05$ is chosen for each test, thinking the power high enough with $n = 23$ per cell. This value is one-sided when testing against the directional alternatives $H_{1,1}$, $H_{1,3}$, $H_{1,4}$, and $H_{1,5}$ and it is two-sided when testing the non-directional nulls $H_{0,2}$ and $H_{0,6}$. 
(see Table 5 for a recapitulation of the relevant statistical hypotheses). With respect to the first contrast (associated with \( H_{1,1} \)), we get

\[
t_1 = \frac{(0.24 - 0.10) - (0.53 - 0.51)}{\sqrt{\frac{(MS_e) * (4/23)}}} = 1.94
\]

This value is statistically significant (s.); it leads to accepting the derived \( H_{1,1} \). The second \( t \) value is:

\[
t_2 = \frac{(0.53 - 0.51)}{\sqrt{0.022 * (2/23)}} = .46,
\]

which is not significant (n.s.), and it leads to retaining the \( H_{0,2} \). This result, however, agrees with the predictions in the same way as accepting the \( H_{1,1} \) above. For the remainder of \( t \) values see Table 5.

**Table 5. Results of the statistical tests for the data in Table 4 aimed at the psychological hypothesis PH-1.**

<table>
<thead>
<tr>
<th>Derived statistical hypothesis</th>
<th>( t )</th>
<th>ASH?</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_{1,1} ): ( \psi_4 = (\mu_{2i} - \mu_{1i}) - (\mu_{12} - \mu_{11}) &gt; 0 )</td>
<td>1.94</td>
<td>yes</td>
</tr>
<tr>
<td>( H_{0,2} ): ( \psi_2 = (\mu_{12} - \mu_{11}) = 0 )</td>
<td>.46</td>
<td>yes</td>
</tr>
<tr>
<td>( H_{1,3} ): ( \psi_3 = (\mu_{22} - \mu_{21}) &gt; 0 )</td>
<td>2.74</td>
<td>yes</td>
</tr>
<tr>
<td>( H_{1,4} ): ( \psi_4 = (\mu_{23} - \mu_{22}) - (\mu_{33} - \mu_{32}) &gt; 0 )</td>
<td>3.56</td>
<td>yes</td>
</tr>
<tr>
<td>( H_{0,5} ): ( \psi_5 = (\mu_{13} - \mu_{12}) = 0 )</td>
<td>-.69</td>
<td>yes</td>
</tr>
<tr>
<td>( H_{1,5} ): ( \psi_8 = (\mu_{23} - \mu_{22}) &gt; 0 )</td>
<td>4.34</td>
<td>yes</td>
</tr>
</tbody>
</table>

**Note.** ASH: Acceptance of derived statistical hypothesis. - \( MS_e \) assumed to be .022, \( n_{jk} = n = 23 \) per cell \((N = 138)\), and no repeated measures on the task factor (A). - Critical \( t(120) \) at \( \alpha = .05: t = 1.658 \) (one-sided), \( t = \pm 1.980 \) (two-sided). ASH: Acceptance of the derived statistical hypothesis. - Each derived statistical hypothesis can be accepted, and so is the statistical prediction as the conjunction of all statistical hypotheses.

The pattern of decisions concerning the partial hypotheses is in complete accordance with the predictions. Since all these hypotheses have been connected by a conjunction, the SP-1 can be accepted, too. Based on the premise that no severe violations of
experimental validity have occurred in the experiment, the psychological hypothesis PH-1 can be taken as corroborated.

Applying the testing strategy to the partial hypotheses derived from the SP-2 leads to rejecting this prediction (see Table 6), since one of the derived partial hypotheses cannot be accepted. The respective psychological hypothesis PH-2, then, cannot be called corroborated. This demonstrates the validity of the argument concerning the variety of patterns and the necessity of differential testing. Needless to say that applying the same standard tests of ANOVA to the different predictable patterns will preclude test-based decisions.

Table 6. Results of the statistical tests for the data in Table 5 aimed at the psychological hypothesis PH-2.

<table>
<thead>
<tr>
<th>Derived statistical hypothesis</th>
<th>t</th>
<th>ASH?</th>
</tr>
</thead>
<tbody>
<tr>
<td>H_{1,1}: \psi_1 = (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}) &gt; 0</td>
<td>1.94</td>
<td>yes</td>
</tr>
<tr>
<td>H_{1,2}: \psi_2 = (\mu_{11} - \mu_{21}) &gt; 0</td>
<td>9.39</td>
<td>yes</td>
</tr>
<tr>
<td>H_{1,3}: \psi_3 = (\mu_{12} - \mu_{22}) &gt; 0</td>
<td>6.64</td>
<td>yes</td>
</tr>
<tr>
<td>H_{1,4}: \psi_4 = (\mu_{12} - \mu_{22}) - (\mu_{13} - \mu_{23}) &gt; 0</td>
<td>3.56</td>
<td>yes</td>
</tr>
<tr>
<td>H_{1,5}: \psi_5 = (\mu_{13} - \mu_{23}) &gt; 0</td>
<td>1.60</td>
<td>no</td>
</tr>
</tbody>
</table>

Note. Same specifications and abbreviations as in Table 5. - Since one of the derived statistical hypotheses (H_{1,5}) cannot be accepted, the statistical prediction SP-2 should be rejected. The respective psychological hypothesis PH-2 therefore is not confirmed.

If M_{23} in Table 4 would have been smaller than .43, H_{1,5} could also have been accepted. In this case, the statistical prediction SP-2 could have been accepted and the psychological hypothesis PH-2 could have been called corroborated. This result, however, would affect the corroboration of the PH-1, because the two psychological hypotheses and the various predictions derived from them refer to different patterns.

Since H_{1,5} is part of the SP-3 and since it is rejected, the SP-3 cannot be accepted and the respective psychological hypotheses cannot be called corroborated. On the other hand, you might say that only one disconfirming result out of nine is reason enough to call the hypothesis PH-3 corroborated.

Let’s return to the two-factorial patterns. Applying the above tests to the original data leads to rejecting the SP-1, since none of the alternatives can be accepted. For the
original means, then, the psychological hypothesis is not corroborated. The analysis by Bjorklund and Zeman (1982) have resulted in the same overall decision, under the premise they had intended to test the same psychological hypothesis. So nothing is gained with the new strategy?

Not quite so, since the different testing strategies are not comparable. The reason why is that they have been constructed to achieve different aims. The strategy favored here focuses on statistical hypotheses (a priori) and their adequate and exhaustive connection to a preceding psychological hypothesis. Questions of efficiency or relative power are only attacked in the second place. As is well-known, planned contrasts are more powerful the ANOVA $F$ tests and multiple comparisons which only allow two-sided tests on means whereas planned comparisons allow one- and two-sided tests. Most standard strategies, on the other hand, focus on overall tests to begin data analysis with, followed by bidirectional post hoc tests in most cases. Questions of minimizing cumulation of $\alpha$ risks and of efficiency of tests are of particular interest with the bulk of these strategies (see Hager, 1992a, for a discussion). In the strategy presented here questions of efficiency are not neglected, however. First, the same type of parametric tests as those used in standard analyses is chosen; second, cumulation of error probabilities (which encompasses cumulation of both $\alpha$ and $\beta$) is kept low by restricting to those statistical tests that are necessary and sufficient with respect of the psychological hypothesis, and the cumulation for these tests is adjusted for by Bonferroni methods (see Westermann & Hager, 1986, for the details, especially with respect to adjusting $\beta$). And third, desirable power values are achieved by appropriate power analysis. Wahlsten's (1991) article is quite helpful for planning of tests especially on one degree-of-freedom contrast hypotheses on differences of mean differences (see also Cohen, 1988, and Hager, 1987, on these matters).

The data served as a demonstration of the testing strategy and its sensitivity to different hypotheses. Although only one pattern of means has been considered, the strategy is very sensitive to any other pattern one might conceive of (see Hager, 1992a).

6. Conclusions

The testing strategy proposed in this article is based on hypotheses on contrasts planned a priori. In order to formulate meaningful contrasts in advance it is necessary to refer to a psychological hypothesis or at least to expectations concerning
experimental results. The former may have arisen from a more comprehensive theory, the latter from a preceding experiment or simply from plausibility considerations. Without such hypotheses or expectations made explicit and used for guiding experimental design no appropriate planning of an experiment can result (Hager, 1987, 1992a). From such hypotheses as well as from expectations, predictions can be derived that always refer to a specific layout for experimental design. In order to perform statistical tests that refer to statistical hypotheses which are meaningful with respect to the psychological hypothesis or to the expectations, the statistical hypotheses should be chosen so as to represent the predictions adequately and exhaustively. This means that the set of results or pattern of results should be unequivocally divided into two subsets, one of which contains those (patterns of) results that agree with the psychological hypothesis whereas the remaining subset contains all those (patterns of) results that disagree.

The choice of statistical hypotheses (and tests), then in the first place is guided by the psychological hypothesis and by the predictions derived from it and despite of whether they fulfill the statistical criterion of orthogonality or not. From this it follows that I have dealt exclusively with the examination of psychological hypotheses by means of statistical hypotheses connected with the former.

The statistical testing strategy proposed in this article is more appropriate for achieving the aim of valid examinations of psychological hypotheses than other strategies currently in use, because of its versatility which is the outstanding feature of planned contrasts in general. It is a one-stage strategy as opposed to the ubiquitous multi-stage procedures when a overall tests is followed by multiple comparisons. Moreover, tests on planned contrasts based on the cell means model are easily and unambiguously to interpret. This is not true with the standard tests of ANOVA (or multiple regression) because of the very unspecific alternatives the $F$ test is testing against. This lack of specificity is the main reason that leads to (routine) application of post hoc tests that are less powerful than focussed contrasts and that are bidirectional, while most psychological hypotheses are directional.

No claim is made to offer a new test or even a completely new testing strategy - planned contrasts have been known for decades, but presumably because of some only footnote-like recommendations by textbook authors (and a text by Rosenthal & Rosnow, 1985, focussing on focussed one degree-of-freedom contrasts) this strategy has only rarely been used. What may be new with my proposal is associating planned
contrasts with a specific (although again not novel) interpretation of first order interaction in ANOVA, namely as differences of mean differences (regardless of presence of one or more within-subject factors), and is identifying sets of contrast hypotheses that grasp the information relevant with respect to the psychological hypothesis. These sets of contrasts always refer to what has been called the testing instance signifying the minimum number of treatment conditions (or cells) necessary to represent the factors addressed by the psychological hypothesis. In the case of the Bjorklund and Zeman experiment, some hypotheses with two factors (task and age group) have been reconstructed and postulated so that the respective testing instance consists of four cells. In the case of a single factor hypothesis the instance consists of two treatment conditions, and so on. Whenever more treatment levels than 2x2 for a two-factor hypothesis are present, several of these testing instances can be constructed. The strategy is, of course applicable to this case and to higher order interactions.

It has been argued on the basis of the criteria of adequateness and exhaustivity of derivation that the (statistical) partial hypotheses referring to one testing instance always have to be connected by a conjunctive rule. The conjunction of hypotheses represents a severe criterion of decision, since each derived partial prediction or hypothesis must be accepted, before a positive decision concerning the psychological hypothesis is possible.

In addition, I have capitalized on the tests necessary and exhaustive for an appropriate decision concerning the predictions and the psychological hypothesis. This focus is not meant to preclude tests of any other hypotheses that the researcher might find interesting (a priori or a posteriori). But these additional hypotheses and tests are not part of the statistical prediction which exclusively refers to a preceding psychological hypothesis; and they do not take part in the cumulation of error probabilities with respect to the psychological hypothesis, since the conceptual unit for cumulation is defined to be a single statistical prediction linked to one psychological hypothesis.

Usually, the type of problem addressed here is attacked from the perspective of (greater) efficiency of a (novel) testing procedure. This criterion is not adopted here. Instead, I propose to let the choice of statistical hypotheses be guided by other criteria - at least in the first place. But if a researcher is interested in more efficient tests, say, more powerful ones, he or she is free to plan her or his tests accordingly by means of
power analysis. Power analysis is easier done with one-stage than with multi-stage strategies.

References


