Formulating and Evaluating Interaction Effects

Floryt van Wesel, Irene Klugkist, Herbert Hoijtink

Authors note

Floryt van Wesel, Irene Klugkist and Herbert Hoijtink, Department of Methodology and Statistics, Utrecht University

This research was supported by grant NWO-VICI-453-05-002 of the Netherlands Organization for Scientific Research.

Correspondence to: Floryt van Wesel, Department of Methodology and Statistics, Utrecht University. P.O.Box 80140, 3508 TC Utrecht, The Netherlands
e-mail: f.vanwesel@uu.nl tel:+31 30 253 1571 fax: +31 30 253 5797

Abstract

Previous research indicated that interaction effects in analysis of variance are difficult to formulate, to evaluate, and to interpret. In this paper we propose to look at interaction effects from a different angle: not in terms of analysis of variance induced main and interaction effects, but in terms of what a researcher expects about the ordering of and differences between group means. We introduce two methods which can be helpful in formulating expected orderings and differences. Furthermore, we compare three ways to evaluate such hypotheses: factorial analysis of variance with post-hoc tests, planned contrasts, and Bayesian model selection. Formulating, evaluating and interpreting interaction effects will be illustrated with examples of a 2x2 and a 2x6 factorial design.

Key words: ANOVA, Bayesian model selection, Contrasts, Inequality constraints, Interaction effects

Introduction

Researchers in the social sciences, whether conducting experiments or survey research, have expectations about the ordering of and the differences between group means. As the current practice tends to use analysis of variance (ANOVA) where there is an interest in group means, researchers are bound to think in terms of main effects and interaction effects. But do these effects correspond to what a researcher expects about the group means involved? We do not think so. When talking to social scientists and when reading their papers it becomes clear that they often formulate their expectations or hypotheses in terms of orderings of group means (Van Wesel, Boeije, & Hoijtink, submitted manuscript) not in terms of main or interaction effects. For example, in a 1x3 design, they write that they expect that the mean of group 1 is higher than the means of group 2 and 3, they do not write that they expect a main effect for groups. The same holds for expectations about main effects in more complicated, factorial, designs. When the expectation concerns an interaction effect, writing it down in exact terms (an ordering of the group means) gets more complicated and what is expected precisely is generally unclear. In this paper, we argue that expectations about interaction effects can be made more transparent by handling them in a similar way as expectations about main effects, in terms of orderings of and differences between group means. In the current article expectations and hypotheses are used interchangeable. In order to achieve transparency in interaction effect expectations we offer researchers in the social sciences two aids for formulating them: means plots and graphic representations of experimental designs.

The issues concerning interaction effects do not stop with formulating expectations about them; evaluating them can also lead to confusion. Currently, the most commonly used approach to evaluate interaction effect hypotheses in the context of factorial ANOVA is to follow up a significant (ANOVA) *F*-test with a post-hoc procedure to see where the differences are (e.g., Bonferroni, 1936; Keuls, 1952; Newman, 1939; Tukey, 1949). In such a post-hoc procedure one can test the simple main effects, which involves one test per level of each factor, or one can use pairwise comparisons, which involves one test per pair of means. As a result of such an analysis we end up with a considerable amount of *F*-test statistics, *t*-test statistics and

corresponding *p*-values, which can be difficult to draw conclusions from, especially with concern to the correctness of and amount of support for a specific expectation (Van de Schoot et al., 2011). These issues become more complex if the factorial design becomes more complicated (for instance in a three instead of two factorial design). Therefore, in order to achieve transparency in the evaluation of interaction effect expectations, we discuss two alternative analysis techniques that correspond better to what researchers are interested in (how correct their specific expectations about group means are) and whose results are easier to interpret: planned contrasts (e.g., Rosenthal, Rosnow, & Rubin, 2000) and Bayesian model selection for (in)equality constrained hypotheses (e.g., Hoijtink, Klugkist, & Boelen, 2008).

Difficulties concerning interaction effects in factorial ANOVA can already be found in a 2x2 design, as was initially shown by Rosnow and Rosenthal (1989a), which lead to a series of papers and replies on the topic. This will be elaborated on in the next section. After introducing the two examples used throughout this paper (a 2x2 factorial design and a 2x6 factorial design), we will propose two methods that can be helpful when formulating interaction effect hypotheses. Subsequently, two different methods to evaluate these hypotheses will be described and compared to the use of factorial ANOVA with post-hoc tests. Finally, conclusions will be drawn.

The interaction effect debate

An article by Rosnow and Rosenthal (1989a) opened a debate on the definition of and interpretation of interaction effects in analysis of variance. A second article on this topic followed a few years later (Rosnow & Rosenthal, 1995). In these articles Rosnow and Rosenthal state that interaction effects are often not correctly understood. According to them, the mistake most researchers make is to interpret a significant interaction effect in ANOVA by looking at the cell means. They state that every cell mean is composed of partly interaction effects and partly main effects. In their 1995 article they describe the hypothetical situation where researchers want to investigate the effects of patient's sex and therapist's sex on therapy effect (for the design see panel I of Table 1). The hypothetical researchers want to evaluate the following interaction effect hypothesis denoted by H_{RR}:

$$H_{RR}: \mu_1 > \mu_2 > \mu_3 = \mu_4,$$

where μ denotes a mean and its subscript indicates the group number. If the interaction effect in the omnibus (two way ANOVA) test they use is found to be significant, the researchers will conclude that there is indeed an interaction effect and they will refer the reader to look at the cell means (descriptive statistics) to see what the effect looks like. Although this approach is commonly used, according to Rosnow and Rosenthal, this interpretation of what the interaction effect looks like is incorrect. This incorrectness can be explained by unraveling the variance that is explained by each test in the two-way ANOVA. The total amount of variance consists of: 1) the variance that is explained by both main effects, leaving unexplained residual variance 2) this residual variance can be further explained by the interaction effect, still leaving some residual variance. From this it follows that the interaction effect explains part of the variance which is left after the main effects explained their part.

- Insert Table 1 about here –

Rosnow and Rosenthal propose the following approach to investigate hypothesized interaction effects: Do not look at the cell means but calculate the residual cell means and use these to reveal the pure interaction effect (interaction effects without the main effects). Panels II, III and IV of Table 1 are equal to Table 3 from Rosnow and Rosenthal (1995), and show how these residual cell means can be calculated. Panel II shows the hypothetical cell means, the row and column effects and a grand mean (G). The estimated cell means (panel III of Table 1) are calculated by the grand mean G plus the row effect plus the column effect. A residual cell mean (panel IV of Table 1) is calculated as the (hypothetical) cell mean minus the estimated cell mean.

- Insert Figure 1 about here –

Corresponding to panel II, III and IV of Table 1, Figure 1 shows means plots for the group means (left plot), the estimated cell means (middle plot), and the residual cell means (right plot). As can be seen in this figure, the interaction effect displayed by the hypothetical cell means (left plot) differs from the interaction effect displayed by the residual cell means (right plot). Rosnow & Rosenthal argue that when researchers use the left plot (cell means) to interpret the interaction effect they actually do not interpret the interaction effect in a pure manner, but that

they interpret it simultaneously with the main effects. When researchers interpret the interaction effect as shown in the right plot (residual cell means), they interpret the pure interaction without the main effects being present.

Abelson (1996), Meyer (1991), Petty, Fabrigar, Wegener, and Priester (1996) and Sohn (2001) replied to the arguments made by Rosnow and Rosenthal. Their main point concerned the definition of what an interaction effect is. These authors describe a difference in what an interaction effect is to an applied scientist and what it is to a statistician. For applied scientists, an interaction effect denotes the idea that an effect is different for different groups (or at different levels of a certain factor). They want to explain how the differences came about, they see the group means and wonder how they came to be as they are; i.e., they are interested in the cell means. For statisticians, an interaction effect is part of a factorial ANOVA and therefore it explains a part of the variance that could not be explained by the main effects; statisticians are interested in wherein simple main effects differ, i.e., the residual cell means. Rosnow and Rosenthal (1991, 1996) responded by repeating their previous arguments, that interaction effect cannot be interpreted by looking at the cell means.

This difference in perspective on what an interaction effect comprehends influences which analysis technique is appropriate to use. When a scientist has a hypothesis about the pure interaction effect - defined by the residual cell means as described by Rosnow and Rosenthal - it can be tested using an omnibus ANOVA. When a scientist has a hypothesis about the ordering of the group means - the cell means - planned contrasts or other appropriate tests should be used (Meyer, 1991; Petty et al., 1996).

Although Rosnow and Rosenthal are correct from a statistical point of view, this seems to have no practical value for applied social science researchers who are interested in a specific ordering of the group means. As the applied scientists are the ones in need of techniques to analyze their expectations adequately, we propose to take their perspective as our basic principle: Interaction effects as residual cell means (and thus factorial ANOVA) are not of interest; what is of interest is the ordering of and the differences between the group means, thereby eliminating the `factorial' in the design. Consequently, a researcher's expectation is of great importance; only

when such expectations are established can we start to think about how to analyze them. It should be noted that in al of the articles that contributed to the debate, hypotheses concerning interaction effect are written in terms of directional or inequality constrained hypotheses.

The debate opened by Rosnow and Rosenthal dealt only with the 2x2 factorial design for the effects of patient's sex and therapist's sex on therapy effect. This example will be used for explanatory purposes throughout this paper. In addition to the relatively simple 2x2 design, we introduce an example of a 2x6 factorial design in the next section.

Entrepreneurial intentions: a 2x2 example

This example is based on a paper by Gupta, Turban, and Bhawe (2008). The authors investigated the effect of implicit and explicit activation of gender stereotypes on men's and women's intentions to pursue a (masculine) career as entrepreneur. The experimental design they used was a 2 (participant gender) x 6 (stereotype activation conditions) between-subjects design. The six stereotype activation conditions were: control, explicit masculine stereotype, implicit masculine stereotype, explicit feminine stereotype, implicit feminine stereotype and nullified stereotype. The design and descriptive statistics can be found in Table 2.

- Insert Table 2 about here –

In the introduction section of their paper the authors stated three hypotheses concerning an interaction effect.

Interaction effect hypothesis 1 (H_1): *`Respondent gender and stereotype activation interact such that men will report stronger entrepreneurial intensions when presented with an implicit versus an explicit masculine stereotype whereas women will report stronger entrepreneurial intensions when presented with an explicit versus an implicit masculine stereotype.'* (Gupta et al., 2008, p. 1055)

Interaction effect hypothesis 2 (H_2): *`Respondent gender and stereotype activation interact such that women will report stronger entrepreneurial intensions when presented with an implicit versus an explicit feminine stereotype whereas men will report stronger entrepreneurial intensions when presented with an explicit versus an implicit feminine stereotype.'* (Gupta et al., 2008, p. 1055)

Interaction effect hypothesis 3 (H_3): `*Respondent gender and stereotype activation interact such that men will report significantly stronger intentions than women in the no stereotypical information condition (control), but men and women will report similar entrepreneurial intentions in the stereotype nullified condition.'* (Gupta et al., 2008, p. 1055)

Note that all hypotheses are stated in terms of `stronger than', indicating an ordering in the group means.

A total of 469 students participated in the study. These participants read a one-page article about entrepreneurship which contained one of the six experimental manipulations for stereotype activation. In the control condition, participants read an article that did not mention gender or gender differences in entrepreneurship. In the two masculine conditions entrepreneurs were associated with masculine characteristics (aggressive, risk taking, and autonomous) and in the two feminine conditions entrepreneurs were associated with feminine characteristics (caring, love to network, and humble). In the implicit condition, masculine or feminine characteristics were described. In the explicit condition, participants were told that entrepreneurs show characteristics of American masculinity/femininity and that so far as entrepreneurship is concerned, it pays to have masculine/feminine characteristics in addition to the aforementioned masculine or feminine characteristics. In the nullified condition, the article stated that entrepreneurs show characteristics of both men and women.

After reading the article the participants were asked to complete a four-item questionnaire with a 5-point scale to measure entrepreneurial intentions, resulting in a mean score between 1 and 5. The items concerned how interested the participants were in starting a business, acquiring a small business, starting and building a high-growth business, and acquiring and building a company into a high-growth business in the next 5 to 10 years (Zhao, Seibert, & Hills, 2005).

Formulating interaction effect hypotheses

The authors of the example described in the previous section stated expectations, expressed as hypotheses, in their article. As we argue that what is of interest to an applied researcher is the

starting point of developing analysis techniques, we think it necessary to carefully elicit what those researchers are interested in. Consequently, we want to know exactly what they expect. However, arriving at such specific hypotheses, specifically those concerning an interaction effect, can be quite difficult. For this reason in this section we will propose two aids that might be helpful in formulating hypotheses: means plots and graphic designs. Using the 2x2 and 2x6 examples we will show the advantages and disadvantages of both aids.

The 2x2 example

Let us conduct a thought experiment using the example of the effect of therapist's sex and patient's sex on treatment effect, as was presented in Table 1. In this experiment we elicit hypotheses from four hypothetical researchers (A, B, C and D) using both aids.

Aid 1: means plot. The first aid that can be used to elicit and formulate (interaction effect) hypotheses is a means plot. In our thought experiment we ask the four hypothetical researchers to draw a means plot of what they think the effect will look like in the population. When they draw such a plot, they need to think thoroughly about how each group mean relates to all the other group means. This is the point at which the hypotheses are actually elicited. The means plots of all four researchers can be found in Figure 2, where the circled numbers in the plots denote the four groups. The scale on the y-axis can be ignored at this stage and will be revisited in the section on analyzing interaction effect using planned contrasts. The means plots will be used to formulate a statistical hypothesis consisting of two parts: Part 1 is an ordering of the group means and Part 2 specifies the interaction effect. The following operators will be used to indicate an ordering of the means: smaller than `<', larger than `>', equal to `=' or no relation `,'. In addition to these operators, binary operators such as minus `-' and plus `+' can be used to specify differences between (combinations of) group means.

- Insert Figure 2 about here –

The means plot of researcher A can be seen in Figure 2A. For the first part of the statistical hypothesis we order the group means:

$$H_{A\ part\ 1}: \mu_4 < \mu_3 < \mu_2 < \mu_1.$$

For the second part of this hypothesis we describe a difference between differences (see the vertical arrows in Figure 2A):

$$H_{A\ part\ 2}: \mu_2 - \mu_4 = \mu_1 - \mu_3.$$

When both parts are combined, we get the complete statistical hypothesis describing the expectation of researcher A:

$$H_A: \mu_4 < \mu_3 < \mu_2 < \mu_1 \ \& \ \mu_2 - \mu_4 = \mu_1 - \mu_3.$$

The means plot of researcher B is represented by Figure 2B. For part 1 of the hypothesis (ordering the means) we get the same statistical hypothesis as we did for researcher A. The difference between the two expectations is expressed in the second part, where the difference between μ_2 and μ_4 is expected to be smaller than the difference between μ_1 and μ_3:

$$H_B: \mu_4 < \mu_3 < \mu_2 < \mu_1 \ \& \ \mu_2 - \mu_4 < \mu_1 - \mu_3.$$

The means plot of researcher C is shown in Figure 2C. Compared to researcher B's means plot, this plot shows a different expectation about the ordering of the group means and therefore a partly different expectation about the difference between differences, resulting in:

$$H_C: \mu_2 < \mu_3 < \mu_4 < \mu_1 \ \& \ \mu_4 - \mu_2 < \mu_1 - \mu_3.$$

The final means plot is drawn by researcher D and is displayed in Figure 2D. When ordering the group means and adding the difference between differences, it leads to:

$$H_D: \{\mu_2 = \mu_3 = \mu_4\} < \mu_1 \ \& \ \mu_2 - \mu_4 < \mu_1 - \mu_3.$$

Note that the second part of the statistical hypothesis D is not necessary to describe the expected effect; the complete hypothesis is implied by the combination of equality and inequality constraints in the first part of the hypothesis.

Aid 2: graphic design. Graphic representations of the experimental design can also be used to elicit expectations about the outcomes of the research. In this approach the four previously mentioned hypothetical researchers were asked to put constraints (< or > or =) between the experimental groups using a graphic representation of an experimental design.

- Insert Figure 3 about here –

The graphic designs for the four researchers can be found in Figure 3. Figure 3A displays the expectation of researcher A. He placed > signs between female and male therapists for both male and female patients representing his expectation that *female therapists will gain better treatment effects than male therapists* (better treatment effects equals higher scores). He also placed v signs (rotated >) between male and female patients representing his expectation that *female patients will gain better treatment effects than male patients*. In this example, with four groups, six relationships (one between each pair) and thus six constraints can be specified. Thus far only four of these relationships are specified. As this researcher has ideas about all relationships between the group means, the diagonal lines need to be taken into account as well. Therefore, the researcher added an inequality constraint on the diagonal line between female patients with a male therapist and male patients with a female therapist. It is not necessary to put a constraint on the other diagonal line, as the constraints present completely determine the ordering of all of the means. Now the first part of H_A, the ordering, is captured. To describe the specific expected difference between differences, indicators need to be added to the design. This is done by the superscript `I' between female and male patients with a female therapist and female and male patients with a male therapist. When constraints share a superscript it denotes which constraint is compared to which other constraint. As the constraints between these groups are similar, `v', it is stated that the difference between these groups is equal. This (non-existing) interaction effect expectation corresponds to Figure 2A and H_A.

Graphic design Figure 3B is the expectation of researcher B. He placed constraints between the group means that are in the same direction as those of researcher A, leading to an equal ordering of the means. However, his expectation concerning the specific interaction effect is different than that of researcher A. To indicate that *therapist's sex has a larger effect on female patients than on male patients* 1) superscripts are placed between female and male patients with a female therapist and female and male patients with a male therapist to indicate that those constraints will be compared and 2) a double larger than sign, >>, is placed between female and male patients with a female therapist to indicate that this difference will be larger than the difference between

female and male patients with a male therapist. This expectation corresponds to Figure 2B and H_B.

The expectations of researcher C are represented by Figure 3C. First, he placed constraints to determine the order of the group means and second, he placed superscripts on two constraints to indicate which ones are compared and a >> for the expected difference between differences. This expectation corresponds to Figure 2C and H_C.

Finally, researcher D's expectation is shown in Figure 3D. Note that he does not have to specify constraints on the diagonals as all relationships are already determined due to the equality signs. This expectation corresponds to Figure 2D and H_D.

The 2x6 example

Because this example is based on an existing paper, we can only guess how the researchers created their hypotheses would they have used a means plot or a graphic design. As will be shown, the hypotheses they state in their article are quite unspecific, leaving room for several competing statistical hypotheses.

Aid 1: means plot. When attempting to draw a means plot for the hypotheses stated in the introduction of Gupta et al. (2008), we observe that we do not know the relationships between all group means. We will illustrate this for H_1. We only know that *men will report stronger entrepreneurial intensions when presented with an implicit versus an explicit masculine stereotype* and that *women will report stronger entrepreneurial intension when presented with an explicit versus an implicit masculine stereotype*, telling us that the effect will be different for men and women (interaction effect). However, we do not know whether men will score higher than women in general (main effects), nor if there are any expectations concerning the other groups. Figure 4 illustrates the issue; it shows two possible means plots corresponding to H_1.

- Insert Figure 4 about here –

The left panel of Figure 4 shows a scenario in which we assume that there is a main effect for sex where entrepreneurial intentions of men are higher than entrepreneurial intentions of women, leading to:

$$H_{1a}: \mu_9 < \mu_2 < \mu_8 < \mu_3 \;\&\; \mu_8 - \mu_2 < \mu_3 - \mu_9 \;\&\; \mu_1, \mu_4, \mu_5, \mu_6, \mu_7, \mu_{10}, \mu_{11}, \mu_{12},$$

where the first part represents the ordering of the group means and the second part represents the difference between differences. The third part denotes that the other means are unconstrained, i.e., it is not hypothesized how they relate to each other or to the means in parts 1 and 2. The right panel of Figure 4 shows another possibility. In this scenario we expect the effect of (masculine) stereotype on entrepreneurial intentions to be inversely proportional for men and women leading to:

$$H_{1b}: \{\mu_2 = \mu_9\} < \{\mu_3 = \mu_8\} \;\&\; \mu_8 - \mu_2 = \mu_3 - \mu_9 \;\&\; \mu_1, \mu_4, \mu_5, \mu_6, \mu_7, \mu_{10}, \mu_{11}, \mu_{12}.$$

Again, the first part represents the ordering of the group means and the second part represents the (inversely proportional) difference between differences, although this part is redundant due to the implications imposed by the other constraints.

Aid 2: graphic design. Figure 5 displays the experimental design with constraints for H_1, H_2 and H_3 stated in Gupta et al. (2008). For H_1 we know that *men will report stronger entrepreneurial intensions when presented with an implicit versus an explicit masculine stereotype* leading to a "<" sign between μ_2 and μ_3, and that *women will report stronger entrepreneurial intensions when presented with an explicit versus an implicit masculine stereotype* leading to a ">" sign

- Insert Figure 5 about here –

between μ_8 and μ_9, shown in the H_1 row of Figure 5. This design leads to the following hypothesis:

$$H_1: \mu_2 < \mu_3 \;\&\; \mu_8 > \mu_9 \;\&\; \mu_1, \mu_4, \mu_5, \mu_6, \mu_7, \mu_{10}, \mu_{11}, \mu_{12}.$$

The H_2 row in Figure 5 shows that *men will report stronger entrepreneurial intensions when presented with an explicit versus an implicit feminine stereotype* leading to a ">" sign between μ_4 and μ_5, and that *women will report stronger entrepreneurial intensions when presented with an implicit versus an explicit feminine stereotype* leading to a "<" sign between μ_{10} and μ_{11}. This leads to:

$$H_2: \mu_4 > \mu_5 \;\&\; \mu_{10} < \mu_{11} \;\&\; \mu_1, \mu_2, \mu_3, \mu_6, \mu_7, \mu_8, \mu_9, \mu_{12}.$$

The final row of Figure 5 shows the third hypothesis which states that *men will report significantly stronger intentions than women in the no stereotypical information condition (control)*} leading to a "v" sign between μ_1 and μ_7, and that *men and women will report similar entrepreneurial intentions in the stereotype nullified condition* leading to the "=" sign between μ_6 and μ_{12}, resulting in:

$$H_3: \mu_1 > \mu_7 \;\&\; \mu_6 = \mu_{12} \;\&\; \mu_2, \mu_3, \mu_4, \mu_5, \mu_8, \mu_9, \mu_{10}, \mu_{11}.$$

The examples show that the use of means plots can be recommended when the relationships between all group means in the design are covered by the expectation, as was the case in the 2x2 example. Placing constraints in a graphic design is the most advantageous aid when not all relationships are clear, as was the case for the 2x6 example. Drawing a means plot is straightforward to use and commonly known. As all relationships need to be specified for this aid, it forces researchers to think more thoroughly about what they expect. On the other hand, researchers need to be aware of the fact that all relationships are specified and consequently that they might over-specify their expectations (when relationships are `unintentionally' established during drawing, see the example given in Figure 4). Another drawback of the means plots aid is that they become hard to draw for higher order factorial designs. Using a graphic design has the advantage of not having to specify all relationships. This might be especially convenient when a design is large, i.e., when there are a lot of group means or for higher order factorial designs. However, one needs to be aware of under-specifying expectations, as diagonal lines need to be drawn in the table, which are easy to forget. It should be noted that irrespective of which aid is preferred, both can be helpful for the elicitation of expectations, making formulating hypotheses a substantial part of conducting social science research.

Analyzing interaction effects

Once the directional hypotheses are formulated, a prudent choice has to be made about how to evaluate them. In the next subsections we will discuss three techniques that can be used to evaluate (in)equality constrained hypotheses, in which situations these techniques are appropriate

and what the advantages and disadvantages of each of the techniques are using the 2x2 example. Thereafter, the 2x6 example will be used as an illustration of each technique.

Analysis using factorial ANOVA

The most commonly used approach to evaluate hypotheses concerning group means is performing a factorial ANOVA. The number of performed *F*-tests depends on the experimental design as an *F*-test is performed for each main effect and for each interaction effect. These tests evaluate,

H_0*: there is no main effect,*

H_a*: there is a main effect,*

for all main effects and

H_0*: there is no interaction effect,*

H_a*: there is an interaction effect,*

for all interaction effects. When such an *F*-test results in a *p*-value smaller than .05 the omnibus ANOVA is followed up by a post-hoc procedure in order to investigate which group means differ from one another (e.g., Bonferroni, 1936; Keuls, 1952; Newman, 1939; Tukey, 1949).. We will discuss two commonly used post-hoc procedures: the pairwise comparisons approach and the simple main effects approach. These procedures have an exploratory character and consequently, *p*-values are two-sided.

In the pairwise comparisons approach, several *t*-tests are performed simultaneously for all possible sets of two group means. For example, in case of the 2x2 design there are 4 groups leading to 4 over 2 = 6 pairs of group means and corresponding *t*-tests (H_0*: the two group means are equal,* H_a*: the two group means differ*). Because multiple tests are performed simultaneously the *α*-level needs to be corrected (e.g., Ramsey, 2002; Shaffer, 1995). This can be done in several ways, such as the Bonferroni correction (Bonferroni, 1936), a correction by Tukey (1949) or a method for correction by Sheffé (Scheffé, 1953). Results of pairwise comparisons can be interpreted in terms of similarities and differences between two group means, as these tests are two-sided. For example: there is a significant difference between *female patients with a female therapist and female*

patients with a male therapist and *there is a significant difference between male patients with a female therapist and male patients with a male therapist*, but *female patients with a male therapist and male patients with a female therapist* not differ significantly. With respect to a hypothesized interaction effect the descriptive statistics need to be consulted in order to see whether or not the expected effect is present. Two things need to be noted here: 1) the expected effect is not tested and 2) this is exactly what Rosnow and Rosenthal (1996, 1995, 1991, 1989b, 1989a) argue is incorrect.

The simple main effects approach (or simple effects analysis) involves evaluating the effect of one factor at all levels of a second factor (e.g., Field, 2005, p. 412-413). This means that an *F*-test is performed per level of a factor, for each factor. For example, in the 2x2 example, this means performing 2 *F*-tests for the simple main effects of patient sex (i.e., *is there a difference between male and female patients with a male therapist? And is there a difference between male and female patients with a female therapist?*) and 2 *F*-tests for the simple main effects of therapist sex (i.e., *is there a difference between male and female therapists when it concerns male patients? And is there a difference between male and female therapists when it concerns female patients?*). For this procedure, as with the pairwise comparison approach, multiple tests are performed so the type 1 error needs to be controlled by adjusting the *a*-level. The results of these (two-sided) tests can be interpreted as the existence of effects of a factor (therapist's sex) within a level of another factor (patients' sex), i.e., *there is an effect of therapist's sex for male patients and there is an effect of therapist's sex for female patients*. With respect to a hypothesized interaction effect we would expect that the simple main effect at one level of a factor is different than the simple main effect at another level of the same factor. For example *there is an effect of therapist's sex for male patients but there is no effect of therapist's sex for female patients*. Equal to the pairwise comparison approach, the descriptive statistics need to be consulted in order to investigate whether or not the expected effect is present.

Analysis using Bayesian model selection

Another method is a Bayesian technique to evaluate (in)equality constrained hypotheses (Mulder, Hoijtink, & Klugkist, 2010; Van Wesel, Hoijtink, & Klugkist, in press; Hoijtink et al., 2008). More information, articles and software can be found on http://tinyurl.com/informativehypotheses.

This method concerns a model selection procedure (in contrast to a null hypothesis testing procedure). Therefore, the relevant question is `Which of the following hypotheses is the best hypothesis in terms of a balance between (model) fit and (model) complexity?' The hypotheses of interest in the 2x2 example can, for instance, be: the classical null hypothesis

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4,$$

the classical alternative hypothesis (H_a), now called the unconstrained hypothesis (H_{unc})

$$H_{unc}: \mu_1, \mu_2, \mu, \mu_4,$$

and the (in)equality constrained hypothesis A:

$$H_A: \mu_4 < \mu_3 < \mu_2 < \mu_1 \ \& \ \mu_2 - \mu_4 = \mu_1 - \mu_3.$$

The results of this analysis are expressed by one Posterior Model Probability (PMP) for each hypothesis. A PMP is a value between 0 and 1 that can be interpreted as the relative amount of support found in the data for each hypothesis under consideration. When, for example, PMP_0=.10, PMP_{unc}=.10 and PMP_A=.80 (note that these values add up to one) we learn that hypothesis A is the most likely hypothesis within this set of hypotheses and we conclude that *male patients with a male therapist gain the least therapy effects, followed by male patients with a female therapist, which in their turn are followed by female patients with a male therapist, leaving female patients with a female therapist as the group with the highest therapy effect.* Important is that in contrast to a classical *p*-value, a PMP does not render an accept/reject decision with respect to the null hypothesis. If, for example, the PMPs are .50, .00 and .50, respectively, the conclusion would be that the data at hand can't distinguish between H_0 and H_A i.e., both hypotheses are equally likely. As this technique concerns model selection it would also have been possible to add the other three (competing) hypotheses H_B, H_C and H_D to the previously mentioned set of hypotheses.

Analysis using planned contrasts

The last method we present here is that of planned contrasts. This technique is extensively described in Rosenthal et al. (2000) and involves specifying contrasts which are evaluated using the *F*- or *t*-test statistic. The ordering under H_A ($\mu_4 < \mu_3 < \mu_2 < \mu_1$) can, for instance, be captured in the standard linear contrast -3, -1, 1, 3, respectively. If the sample means are indeed increasing

(e.g., the sample means are 10, 20, 30, 40, respectively) then the contrast results in a positive value (e.g., C=-3*10 -1*20 + 1*30 + 3*40). A contrast value C can be calculated by $C = \sum_{j=1}^{J} c_j \cdot \mu_j$, where $j=1,...,J$ denotes the number of groups and c_j denotes a contrast value for group j. Consequently, a contrast test evaluates

$$H_0: C=0,$$

$$H_a: C>0,$$

As H_a is a directional hypothesis it concerns a one-sided test. For this approach it is not necessary to start with an omnibus ANOVA.

Each contrast value C is formulated by assigning specific contrast values to the groups (c_j) such that $C>0$ corresponds to support for the directional hypothesis. How this is done will be described next. For the 2x2 example the means plots in Figure 2 will be used as reference point. For eliciting the interaction effect hypotheses as described in the previous section, the scale on the y-axis of the plots was not important. However, for establishing contrast values this scale becomes very important, as the value of a group on the y-axis will be used as weight. Consequently, these weights do not only describe the ordering of the group means, they also describe a *relative effect*. Figure 2B show that the weights for H_B are 4, 2, 1, 0, however, they could also haven been, for instance, 8, 2, 1, 0. Using these latter weights, the relative distance between the first mean and the other means is much larger. Note that when means plots are used for this purpose, this relative effect needs to be elicited as well. Reading the values (weights) per group from the y-axis of Figure 2 for each hypothesis leads to the upper panel of Table 3. This panel also displays a mean weight per hypothesis.

- Insert Table 3 about here –

These weights are not contrast values yet, as the mean of the contrasts values should be zero (see H_0 and H_a above and Rosenthal et al., 2000). Therefore, the weights in the upper panel of Table 3 need to be converted. In order to do so, the mean weight of a hypothesis is subtracted from the group weights of that hypothesis, leading to the middle panel of Table 3. As can be seen, the mean weight per hypothesis is now zero. The contrast values (c_j) that are used in the

analysis are usually expressed as integers and consequently, the weights are multiplied to get such integers (Rosenthal et al., 2000). These integer values are called contrast values and are displayed in the lower panel of Table 3. Note that a weight of zero in the upper panel of Table 3 has a different interpretation than a contrast value of zero in the lower panel of the same table; in case of a weight, a zero indicates the position of a group on the y-axis of the means plot, and therefore this value is relative to the other weight values (in our example a zero denotes the smallest mean), in case of a contrast value, a zero means that the group with that contrast value does not influence the test statistic.

Hypotheses H_A, H_B and H_C can be evaluated using only one contrast, as is shown in the lower panel of Table 3. However, H_D involves equality constraints between μ_2, μ_3 and μ_4 and, therefore we cannot distinguish these three groups (they have the same contrast weight, 1). Extra contrasts need to be added to evaluate whether or not these three group means are equal. H_D contains three equal groups, and therefore two extra contrasts are needed: one contrast, *extra 1* (in the lower panel of the table), is for evaluating whether or not the second and third group means are equal, and one contrast, *extra 2*, is used to evaluate whether or not the second (and implicitly the third) and fourth group means are equal. As the null hypothesis states that the contrast is equal to zero, a non-significant test result is expected for both these extra contrasts concerning H_D. Note here that it can never be concluded that H_0 is correct, due to the underlying falsification procedure in null hypothesis testing (Cohen, 1990; Lehmann & Romano, 2005).

Each evaluated contrast results in an $F_{contrast}$-test statistic and corresponding one-sided p-value. Making inferences concerning H_A, H_B and H_C is straightforward as each hypothesis can be tested using one contrast. If the contrast is significant, it can be concluded that the group means are ordered according to the contrast weights (supporting the hypothesis), if not, the group means are equal. However, for H_D drawing conclusions is more complicated as this hypothesis is evaluated using three contrasts. In order to conclude that the hypothesis is supported, all three contrasts need to result in the expected direction, (significant or non-significant) after correcting for multiple testing: contrast H_D, involving the inequality constraint, needs to be significant and contrasts *extra 1* and *extra 2*, involving the equality constraints, need to be non-significant. When

one of the three contrasts has a result opposite to the expected direction, no support is found for the hypothesis.

Evaluating the 2x6 example

For the evaluation of the hypotheses belonging to the 2x6 example, a data set with the descriptive statistics in Table 2 was used. The hypotheses stated in the introduction of Gupta et al. (2008),

$$H_1: \mu_2 < \mu_3 \;\&\; \mu_8 > \mu_9 \;\&\; \mu_1, \mu_4, \mu_5, \mu_6, \mu_7, \mu_{10}, \mu_{11}, \mu_{12}.$$

$$H_2: \mu_4 > \mu_5 \;\&\; \mu_{10} < \mu_{11} \;\&\; \mu_1, \mu_2, \mu_3, \mu_6, \mu_7, \mu_8, \mu_9, \mu_{12}.$$

$$H_3: \mu_1 > \mu_7 \;\&\; \mu_6 = \mu_{12} \;\&\; \mu_2, \mu_3, \mu_4, \mu_5, \mu_8, \mu_9, \mu_{10}, \mu_{11}.$$

are analyzed using the factorial ANOVA followed up by the pairwise comparisons approach, the planned contrasts approach and the Bayesian approach. A summary of the results can be found in Table 4, where supported hypotheses are displayed in boldface.

- Insert Table 4 about here –

Equal to the analysis done by Gupta et al. (2008), a two-way ANOVA was performed, with condition and sex as fixed factors. This analysis was done using SPSS 16.0 statistical software, using the module: General linear Model - Univariate. First, consistent with their findings, a significant main effect for sex, $F(1,457)=39.62$, $p<.001$, and a significant interaction effect, $F(5,457)=3.18$, $p<.05$, were found. Second, in order to find out what the interaction effect looks like, post-hoc pairwise comparison tests, with Tukey's correction for multiple testing, were performed. Note that the p_t-values displayed in Table 4 are (Tukey) corrected for multiple testing, therefore, the familiar $a=.05$ is used as decision rule. Third, to interpret the results of this post-hoc analysis it was checked whether or not the expected results were found: for H_1: $\mu_2 \neq \mu_3$ and $\mu_8 \neq \mu_9$, for H_2: $\mu_4 \neq \mu_5$ and $\mu_{10} \neq \mu_{11}$ and for H_3: $\mu_1 \neq \mu_7$ and $\mu_6 = \mu_{12}$. Only H_3 was supported by the results of these tests. As expected there was no significant difference between μ_6 and μ_{12} (although actually we can only conclude that we cannot reject H_0) and there was the expected (two-sided) significant difference between μ_1 and μ_7. As these tests are two-sided the descriptive statistics need to be consulted to see whether the difference is in the right direction; The

descriptive statistics in Table 2 show that they are. The rest of the expected significant differences between group means were not found. Using this analysis support is found only for H_3.

To analyze the data using planned contrasts SPSS 16.0 statistical software, again the module: General linear Model - Univariate was used. First, the contrasts need to be established. As the statistical hypotheses consist of two separate inequality constraint parts, for example, for H_1: $\mu_2<\mu_3$ and $\mu_8>\mu_9$, each hypothesis was captured in two contrasts (C_1 and C_2). Consequently, for C_1 of H_1 we get a contrast value of -1 for μ_2 and 1 for μ_3 (as μ_2 is smaller than μ_3) and for C_2 of the same hypothesis we get a contrast value of 1 for μ_8 and -1 for μ_9 (as μ_9 is smaller than μ_8). Table 5 shows the contrasts values that we used to analyze the three hypotheses.

- Insert Table 5 about here –

Second for the hypotheses to be supported, we expect to find (one-sided) significant test results for both contrast of $H1$ and H_2, and for H_3 we expect the first contrast to be significant and the second contrast to be (two-sided) non-significant. The middle panel of Table 4 shows the (one-sided) results of this analysis. Note that this module in SPSS 16.0 gives *t*-test statistics. Furthermore, the panel displays the contrast value C in order to be able to check whether or not the direction of the contrast is as expected, i.e., if C is larger than 0. For H_1 both contrasts were significant and for both $C>0$ holds. For H_2 both contrasts (C_1 and C_2) were not significant. For H_3 the first contrast was significant and $C>0$ and the second (C_2) was not significant, as was expected. Therefore, it can be concluded that support was found in favor of H_1 and H_3.

In order to perform the Bayesian model selection method, software called BIEMS (downloaded from http://tinyurl.com/informativehypotheses) was used. The directional hypotheses H_1, H_2, H_3 were each compared to the corresponding null-hypothesis (H_0) and the unconstrained hypothesis (H_{unc}). The Posterior Model Probabilities for the sets H_1, H_0, H_{unc}, and H_2, H_0, H_{unc}, and H_3, H_0, H_{unc}, resulting from this analysis can be found in the lower panel of Table 4. Here it can be seen that all three inequality constrained hypotheses (H_1, H_2, H_3) are more likely than the null-hypothesis (H_0). Furthermore, for H_2, the amount of support for the directional hypothesis (PMP_2=.45) is similar to the amount of support for the unconstrained hypothesis (PMP_{unc}=.55), indicating that a hypothesis without any constraints is slightly more

likely and therefore that the support for H_2 is weak. It can be concluded that support is found in favor of H_1 and H_3.

As the boldfaced symbols in Table 4 display, the results from the techniques that actually evaluate the ordering of the group means (planned contrasts and Bayesian model selection), indicated by the hypotheses, lead to the same conclusions. In contrasts, the technique evaluating non-directional hypotheses (factorial ANOVA with pairwise comparisons post-hoc tests) shows a different result. Note that the results of both classical methods differ due to the fact that the number of multiple test, and thus the correction, differs. Concerning the hypotheses by Gupta et al. (2008), all three methods provide evidence in support of H_3. Furthermore, none of the three methods provide evidence for H_2. Different conclusions are drawn for H_1. Both the planned contrast method and the Bayesian method give evidence in support of this hypothesis, but the factorial ANOVA does not. Given these results, we would conclude that `*Men report stronger entrepreneurial intensions when presented with an implicit versus an explicit masculine stereotype whereas women report stronger entrepreneurial intensions when presented with an explicit versus an implicit masculine stereotype.*' And that `*Men report stronger intentions than women in the no stereotypical information condition, but men and women report similar entrepreneurial intentions in the stereotype nullified condition.*'.

Comparing these methods

The three methods described above have their own advantages and disadvantages. Table 6 shows a summary of the comparison on the following aspects: what each method tests, practicability, inferences and drawbacks.

- Insert Table 6 about here –

The factorial ANOVA with post-hoc tests can be seen as an explorative approach, whereas the planned contrasts method and Bayesian method can be seen as confirmative analysis approaches. In an explorative approach, a specific ordering of means is not tested explicitly, therefore the carefully formulated hypotheses are not used to their utmost potential. Conversely, the planned contrast method and the Bayesian method both evaluate whether a specific

hypothesis, and thus ordering of the group means, is supported or not (including the relative effect for planned contrasts).

Concerning practicability, the factorial ANOVA with post-hoc tests and the planned contrast method have an advantage over the Bayesian method as they are more familiar to most social scientists. Furthermore, both approaches can be performed in standard statistical software, as for example SPSS. In contrast, the Bayesian method is not well known and therefore not available in standard statistical software. Special software and tutorials for this method can be found on http://tinyurl.com/informativehypotheses.

The factorial ANOVA method results in multiple *F*- and *t*-test statistics and their corresponding (two-sided) *p*-values. In order to draw conclusions from these figures with concern to the amount of support for the specific hypothesis, we need to turn to the descriptive statistics. The larger the design, the more tests need to be performed. All these test results may make it difficult, if not impossible, to draw straightforward conclusions (Van de Schoot et al., 2011). The planned contrast method results in a number of *F*-test statistics and corresponding (one-sided) *p*-values equal to the number of formulated contrasts. When multiple contrasts are needed to evaluate one (directional) hypothesis, drawing straightforward conclusions can be difficult if some contrasts give the expected results and some do not. The Bayesian method results in a number of Posterior Model Probabilities equal to the number of competing hypotheses, where the highest PMP indicates the most likely hypothesis given the data and the set of hypotheses.

The drawbacks of the factorial ANOVA approach are the most substantial since this approach seems inappropriate for testing directional hypotheses as the hypotheses that are of interest are not tested at all. In addition the capitalization on chance, due to several test being performed at once, should be corrected. The drawbacks of the planned contrast approach involve the issue that establishing the contrast values can be rather complicated because relative effects need to be specified on top of the hypotheses. Furthermore, for both classical methods problems may arise when several test needed to evaluate one hypothesis have conflicting results

and H_0 can never be confirmed. The most important drawback of the Bayesian approach involves its practicability in terms of software and familiarity with the technique and its results.

Concluding from the above, the factorial ANOVA method can best be used when a researcher is interested in exploring and describing the data. The planned contrast method can best be used when a researcher had a directional hypothesis which can be captured in one contrast and the Bayesian method can be used in all situations where a researcher is interested in (competing) directional hypotheses.

Conclusions

In this paper we investigated how hypotheses concerning interaction effects can be formulated in terms of (in)equality constrained hypotheses and how these hypotheses can be evaluated, for a 2x2 and a 2x6 factorial ANOVA example. As we take the position that the expectations of researchers should be the starting point of developing analysis techniques, instead of the other way around, we proposed to formulate expectations about interaction effects by leaving out the *factorial* in the design. The sort of hypotheses we propose exist of two parts: the first part in which the ordering of the group (cell) means is described, and a second part in which the difference between differences is described. In order to establish these hypotheses, two aids for elicitation are proposed; means plots, which are useful when the research design is simple and all possible relationships between the parameters can be described, and graphic representations of experimental designs, which are useful when the design is more complicated and not all relationships between the parameters need to be specified. Furthermore, we proposed not to evaluate these inequality constrained hypotheses using factorial ANOVA but to evaluate them using planned contrasts or Bayesian model selection for (in)equality constrained hypotheses.

Interaction effects are of interest in a large part of social science research. This being the case, more effort should be made to formulate expectations concerning these effects in more detail. In addition, a more suitable method should be used to analyze these effects. In short, interaction effect should be handled with greater care.

References

Abelson, R. (1996). Vulnerability of contrast tests to simpler interpretations: An addendum to Rosnow and Rosenthal. *Psychological science, 7*, 242-246.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilit. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8*, 3-62.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Field, A. (2005). *Discovering statistics using SPSS*. London: Sage.

Gupta, V., Turban, D., & Bhawe, N. (2008). The effect of gender stereotype activation on entrepreneurial intentions. *Journal of Applied Psychology, 93(5)*, 1053-1061.

Hoijtink, H., Klugkist, I., & Boelen, P. A. (Eds.). (2008). *Bayesian evaluation of informative hypotheses*. NY: Springer.

Keuls, M. (1952). The use of the studentized range in connection with an analysis of variance. *Euphytica, 1*, 112122.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses. Third edition*. NY: Springer

.Meyer, D. (1991). Misinterpretation of interaction effects: A reply to Rosnow and Rosenthal. *Psychological Bulletin, 110(3)*, 571-573.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Inequality and equality constrained multivariate linear models: objective model selection using constrained posterior priors. *Statistical planning and inference, 140* , 887-906.

Newman, D. (1939). The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika, 31(1)*, 20-30.

Petty, R., Fabrigar, L., Wegener, D., & Priester, J. (1996). Understanding data when interactions are present or hypothesized. *Psychological science, 7*, 247-252.

Ramsey, P. H. (2002). Comparison of closed testing procedures for pairwise testing of means. *Psychological Methods, 7(4)*, 504523.

Rosenthal, R., Rosnow, R., & Rubin, D. (2000). *Contrasts and effect sizes in behavioral research*. Cambridge University Press.

Rosnow, R., & Rosenthal, R. (1989a). Definition and interpretation of interaction effects. *Psychological Bulletin, 105(1)*, 143-146.

Rosnow, R., & Rosenthal, R. (1989b). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44(10)*, 1276-1284.

Rosnow, R., & Rosenthal, R. (1991). If you're looking at the cell means, you're not looking at only the interaction (unless all main effects are zero). *Psychological Bulletin, 110(3)*, 574-576.

Rosnow, R., & Rosenthal, R. (1995). Some things you learn aren't so: Cohen's paradox, Asch's paradigm, and the interpretation of interaction. *Psychological science, 6*, 3-9.

Rosnow, R., & Rosenthal, R. (1996). Contrasts and interactions redux: Five easy pieces. *Psychological science, 7*, 253-257.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika, 40*, 87-104.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46*, 561-584.

Sohn, D. (2001). The interaction concept in scientific discourse and in the analysis of variance. *The Journal of psychology, 125(6)*, 621-629.

Tukey, J. (1949). Comparing individual means in the analysis of variance. *Biometrics, 5*, 99.

Van de Schoot, R., Hoijtink, H., Mulder, J., van Aken, M. A. G., de Castro, B. O., Meeus, W., et al. (2011). Evaluating expectations about negative emotional states using Bayesian model selection. *Developmental Psychology, 47*, 203-212.

Van Wesel, F., Boeije, H., & Hoijtink, H. (submitted manuscript). Use of hypotheses for analysis of variance models: Challenging the current practice.

Van Wesel, F., Hoijtink, H., & Klugkist, I. (2011). Choosing priors for constrained analysis of variance : methods based on training data. *Scandinavian Journal of Statistics*.

Zhao, H., Seibert, S. E., & Hills, G. E. (2005). The mediating role of self-efficacy in the development of entrepreneurial intentions. *Journal of Applied Psychology, 90(6)*, 1265-1272.

Table 1: Table of means with row and column effects leading to the estimates in the example from Rosnow and Rosenthal (1995)

I. Design		Therapist sex		
		female	male	
Patient sex	female	μ_1	μ_2	
	male	μ_3	μ_4	
II. Cell means				Row effect
		+2	0	+1
		-1	-1	-1
	Column effect	+0.5	-0.5	G =0
III. Estimated cell means				
		+1.5	+0.5	
		-0.5	-1.5	
IV. Residual cell means				
		+0.5	-0.5	
		-0.5	+0.5	

Note. Estimated cell mean = G + row effect + column effect Residual mean = cell mean - estimated cell mean

Table 2: Group numbers (*group*), means (*M*), standard deviations (*SD*) and number of participants (*N*) in the entrepreneurial intentions example

Condition			Masculine stereotype		Feminine stereotype		
		control	explicit	implicit	explicit	implicit	nullified
men	*group*	1	2	3	4	5	6
	M	3.44	2.94	3.48	3.37	3.49	3.09
	SD	1.01	1.07	1.08	1.01	0.94	1.03
	N	38	46	36	48	41	37
women	*group*	7	8	9	10	11	12
	M	2.66	2.93	2.43	2.56	2.68	2.94
	SD	1.09	1.05	1.00	0.99	0.94	1.12
	N	37	33	39	35	38	41

Table 3: Weights and contrasts for H_A, H_B, H_C and H_D

Weights indicated by means plot					
Hypothesis	Group number				mean
H_A	4	3	2	1	2.50
H_B	4	2	1	0	1.75
H_C	4	0	1	2	1.75
H_D	4	1	1	1	1.75
Weights summing up to zero					
	1	2	3	4	
H_A	1.5	0.5	-0.5	-1.5	0
H_B	2.25	0.25	-0.75	-1.75	0
H_C	2.25	-1.75	-0.75	0.25	0
H_D	2.25	-0.75	-0.75	-0.75	0
Contrasts in integer values (c_j)					
	1	2	3	4	
H_A	3	1	-1	-3	0
H_B	9	1	-3	-7	0
H_C	9	-7	-3	1	0
H_D	9	-3	-3	-3	0
extra 1	0	1	-1	0	
extra 2	0	1	0	-1	

Table 4: Results of the factorial ANOVA with pairwise comparison post-hoc tests (two-sided, Tukey corrected (p_t-values), planned contrasts (one-sided, p_o-values) and Bayesian model selection for H_1, H_2 and H_3

Factorial ANOVA and post-hoc tests			
Main effect sex	$F(1, 457) = 39.62, p < .001$		
Main effect condition	$F(5, 457) = 0.25, p = .94$		
Interaction effect	$F(5, 457) = 3.18, p = .01$		
	H_1	H_2	H_3
H_0	$\mu_2=\mu_3, p_t=.44$	$\mu_4=\mu_5, p_t>.999$	$\mu_1=\mu_7, p_t=.05$
H_0	$\mu_8=\mu_9, p_t=.65$	$\mu_{10}=\mu_{11}, p_t>.999$	$\mu_6=\mu_{12}, p_t>.999$
Planned contrasts			
	H_1	H_2	H_3
H_0:C_1=0	C_1=0.54	C_1=-0.12	C_1=0.78
	$t(457)=2.36, p_o=.01$	$t(457)=0.55, p_o=.29$	$t(457)=3.29, p_o<.001$
H_0:C_2=0	C_2=0.50	C_2=0.12	C_2=0.15
	$t(457)=2.06, p_o=.02$	$t(457)=0.50, p_o=.31$	$t(457)=0.64, p_o=.26$
Bayesian model selection			
	H_1	H_2	H_3
$PMP_{1;2;3}$	0.79	0.45	0.81
PMP_0	0.00	0.00	0.06
PMP_{unc}	0.21	0.55	0.13

Table 5: Contrasts for H_1, H_2 and H_3

		μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}	μ_{11}	μ_{12}
H_1	C_1	0	-1	1	0	0	0	0	0	0	0	0	0
	C_2	0	0	0	0	0	0	0	1	-1	0	0	0
H_2	C_1	0	0	0	1	-1	0	0	0	0	0	0	0
	C_2	0	0	0	0	0	0	0	0	0	-1	1	0
H_3	C_1	1	0	0	0	0	0	-1	0	0	0	0	0
	C_2	0	0	0	0	0	1	0	0	0	0	0	-1

Table 6: Summary of comparison of the three methods

Method	What is tested	Practicability	Inferences	Drawbacks
Factorial ANOVA and post hoc tests	Omnibus ANOVA testing main and interaction effects and post hoc testing for differences between all pairs of means or simple main effects	easy to test in standard statistical software (SPSS)	ANOVA *F* and corresponding *p* values for main and interaction effects and post hoc *t* or *F* and corresponding *p*-values for differences between means. Descriptive statistics needed to check specific ordering	ANOVA does not evaluate directional hypotheses, capitalization on chance, possible conflicting multiple test results and equalities cannot be confirmed
Planned contrasts	testing a specific ordering of means, including the relative effects	easy to test in standard statistical software (SPSS)	*F*(or *t*) and corresponding (one-sided) *p* values for each contrast	the relative effects need to be elicited, possible conflicting multiple test results and equalities cannot be confirmed
Bayesian Model selection	evaluating one specific ordering of means (or multiple orderings)	special soft-ware needed (BIEMS)	Posterior model probabilities for each hypothesis in the set	software and method are unfamiliar

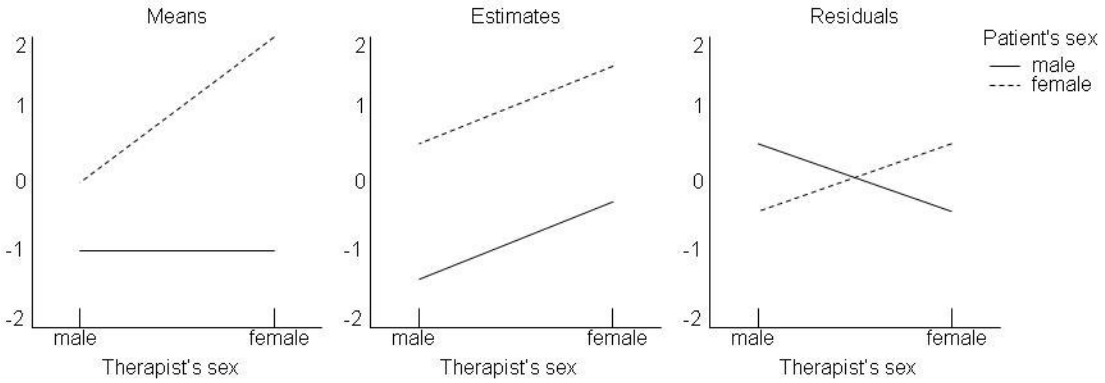

Figure 1. Interaction plots for the means, estimates, and residuals in the example from Rosnow and Rosenthal (1995)

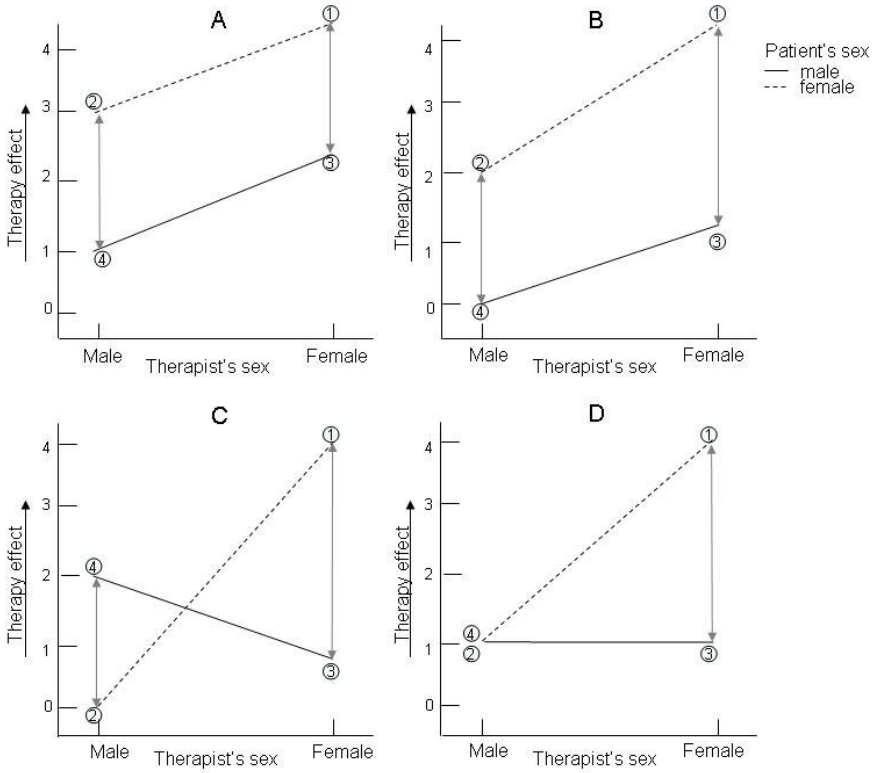

Figure 2. Means plots for the expectations of researchers A,B,C and D on the effect of therapist's sex and patient's sex on treatment effect

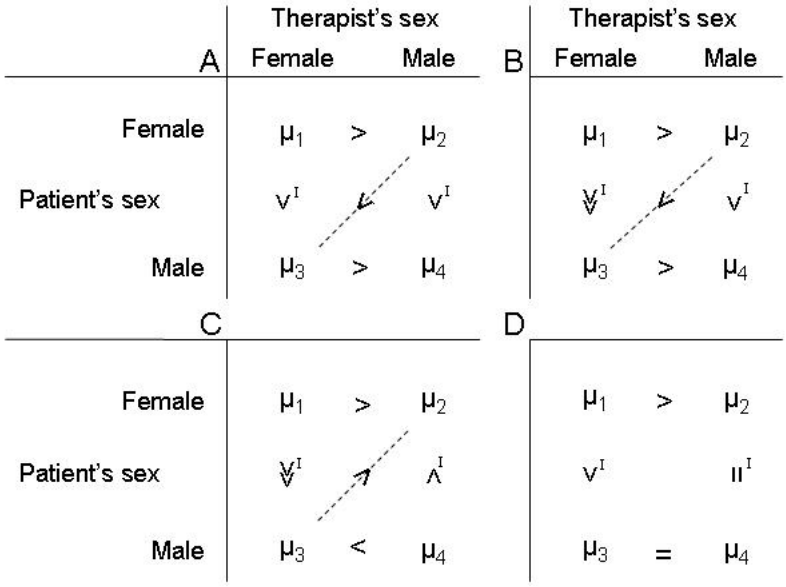

Figure 3. Graphic designs specifying the expectations of researcher A,B,C and D with respect to the effect of therapist's sex and patient's sex on treatment effect

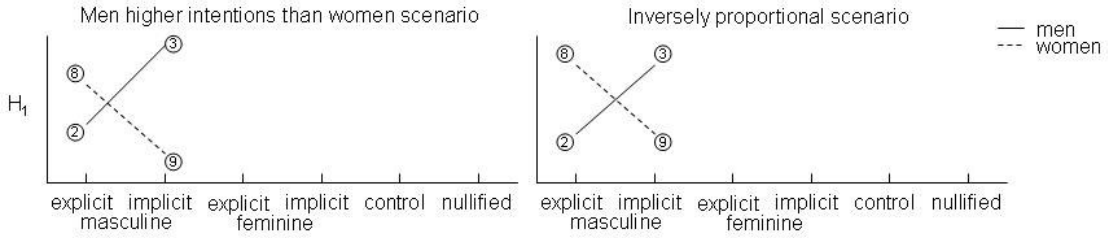

Figure 4. Means plots for two possible scenarios of H_1

	Condition	Masculine stereotype			Feminine stereotype				
		explicit		implicit	explicit		implicit	control	nullified
H_1	Men	μ_2	<	μ_3	μ_4		μ_5	μ_1	μ_6
	Women	μ_8	>	μ_9	μ_{10}		μ_{11}	μ_7	μ_{12}
H_2	Men	μ_2		μ_3	μ_4	>	μ_5	μ_1	μ_6
	Women	μ_8		μ_9	μ_{10}	<	μ_{11}	μ_7	μ_{12}
H_3	Men	μ_2		μ_3	μ_4		μ_5	μ_1	μ_6
								V	II
	Women	μ_8		μ_9	μ_{10}		μ_{11}	μ_7	μ_{12}

Figure 5. Graphic designs for H_1, H_2 and H_3