

An Introduction to Bayesian Evaluation of Informative Hypotheses

Herbert Hoijtink¹, Irene Klugkist¹, and Paul A. Boelen²

¹ Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, the Netherlands h.hoijtink@uu.nl and i.klugkist@uu.nl

² Department of Clinical and Health Psychology, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, the Netherlands p.a.boelen@uu.nl

1.1 Bayesian Evaluation of Informative Hypotheses

Null hypothesis significance testing (NHST) is one of the main research tools in social and behavioral research. It requires the specification of a null hypothesis, an alternative hypothesis, and data in order to test the null hypothesis. The main result of a NHST is a p -value [3]. An example of a null hypothesis and a corresponding alternative hypothesis for a one-way analysis of variance is:

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

and

$$H_a : \mu_1, \mu_2, \text{ and } \mu_3, \text{ are not all equal,}$$

where μ_1 , μ_2 , and μ_3 represent the average score on the dependent variable of interest in three independent groups. The implication of the null hypothesis has been often criticized. Cohen [1] calls it the “nil hypothesis” because he finds it hard to imagine situations, especially in psychological research, where “nothing is going on,” and three means are exactly equal to each other. The meaning of the alternative hypothesis can also be criticized. If H_0 is rejected, and thus H_a is implicitly accepted, we find ourselves in a situation that can be labelled “something is going on but we don’t know what.” Knowing that three means are not all equal (H_a) does not tell us which means are different or what the order of the means is. Stated otherwise, the null hypothesis describes the population of interest in an unrealistic manner, and the alternative hypothesis describes the population of interest in an uninformative manner.

In this book we will introduce and exemplify the use of informative hypotheses. An informative hypothesis can be constructed using inequality ($<$ denotes smaller than and $>$ denotes larger than) and about equality (\approx) constraints. Two examples of informative hypotheses are

$$H_{1a} : \mu_1 > \mu_2 > \mu_3$$

and

$$H_{1b} : \{\mu_1 \approx \mu_2\} > \mu_3.$$

The first hypothesis states that μ_1 is larger than μ_2 and that μ_2 is larger than μ_3 . The second hypothesis states that μ_1 is about equal to μ_2 and that both are larger than μ_3 . The inequality constraints $<$ and $>$ can be used to add theoretical expectations to the traditional alternative hypothesis H_a , thus making it more informative. The about equality constraint \approx has two advantages. First of all, if, like Cohen [1], researchers consider the traditional null hypothesis H_0 to be a “nil hypothesis,” they can replace it by

$$H_{1c} : \mu_1 \approx \mu_2 \approx \mu_3.$$

Of course it has to be specified what is meant by \approx . In words, $\mu_1 \approx \mu_2$ means that μ_1 is not substantially different from μ_2 . In a simple formula this means that

$$|\mu_1 - \mu_2| < \delta,$$

where δ is the smallest difference between two means that is considered to be relevant by the researchers formulating the hypotheses. This immediately leads to the second advantage of using \approx constraints. If the traditional null hypothesis is rejected by a significance test, it still has to be determined whether the effect found is relevant or not. If $H_{1c} : \mu_1 \approx \mu_2 \approx \mu_3$ is rejected, this is not necessary, because the relevance of an effect is already included in H_{1c} .

Classical and informative hypotheses differ not only in the manner in which they are formulated but also in the manner in which they are evaluated. The classical null and alternative hypotheses can be evaluated using a p -value. For a one-way analysis of variance the p -value is, loosely speaking, the probability that the differences in means observed in the data or larger differences come from a population where the null hypothesis is true. According to a popular rule, the null hypothesis is rejected if the p -value is smaller than .05. Informative hypotheses can be evaluated using Bayesian model selection. The main result of Bayesian model selection is the posterior probability [2]. The posterior probability represents the support in the data for each hypothesis under investigation. For H_{1a} and H_{1b} these probabilities could, for example, be .90 and .10, respectively. This implies that after observing the data, H_{1a} is nine times as probable as H_{1b} .

1.2 Overview of the Book

The book consists of four parts. The first part, “Bayesian Evaluation of Informative Hypotheses,” consists of Chapters 2 through 5. Subsequently, informative hypotheses, Bayesian estimation, Bayesian model selection, and the usefulness of the traditional null, alternative and informative hypotheses will

be discussed in the context of analysis of variance models. The first part is an introduction and tutorial in which all the steps involved in the formulation and evaluation of informative hypotheses are subsequently discussed. This part of the book is suited for both social scientists and statisticians.

The second part, “A Further Study of Prior Distributions and the Bayes Factor,” which consists of Chapters 6 through 9, contains a further elaboration of the use of Bayesian model selection for the evaluation of informative hypotheses. The second part of the book is rather technical and aimed at statisticians. Subsequently, different specifications of prior distributions, Bayesian alternatives for the use of posterior model probabilities, and the contrast between classical and Bayesian analysis will be discussed.

The third part of the book, “Beyond Analysis of Variance,” discusses the application of informative hypotheses beyond the context of analysis of variance. It consists of Chapters 10 through 13, in which the application of informative hypotheses will subsequently be discussed for analysis of covariance, latent class models, models for contingency tables, and multilevel models. The third part of the book is suited for both social scientists and statisticians.

The fourth part of the book, “Evaluations,” consists of Chapters 14 through 16. The concept of informative hypotheses will be discussed from the perspectives of psychologists, statisticians, and philosophers of science. This part of the book is also suited for both social scientists and statisticians.

1.3 Software

For many of the models and approaches discussed in this book software, and manuals are available. Software for inequality constrained analysis of variance and covariance, inequality constrained latent class models, and models for contingency tables as discussed in Chapters 3, 4, 10, 11, and 12, respectively, can be found at <http://www.fss.uu.nl/ms/informativehypotheses>. Software for inequality constrained analysis of variance as discussed in Chapter 6 can be found at <http://rosselldavid.googlepages.com>. Readers interested in software for the other approaches/models discussed should contact the authors of the respective chapters.

References

- [1] Cohen, J.: The earth is round ($p < .05$). *American Psychologist*, **49**, 997–1003 (1994)
- [2] Kass, R.E., Raftery, A.E.: Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795 (1995)
- [3] Schervish, M.J: P values: what they are and what they are not. *The American Statistician*, **50**, 203–206 (1996)

Illustrative Psychological Data and Hypotheses for Bayesian Inequality Constrained Analysis of Variance

Paul A. Boelen¹ and Herbert Hoijtink²

¹ Department of Clinical and Health Psychology, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, the Netherlands p.a.Boelen@uu.nl

² Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, the Netherlands h.hoijtink@uu.nl

2.1 Introduction

In this chapter, three datasets from existing psychological research programs will be introduced that allow for an investigation of differences between groups on a single outcome variable. The first dataset was gathered to study amnesia in people with Dissociative Identity Disorder. The second dataset was originally generated to study emotional reactivity and emotional regulation in children subjected to different kinds of social evaluation by peers. The third dataset was obtained from research on coping with loss that, among other purposes, was used to study gender differences in coping. These three datasets will be used in subsequent chapters to illustrate Bayesian inequality constrained analysis of variance. To set the stage for these illustrations, in the current chapter, we will provide background information for the three datasets and will introduce theories and corresponding hypotheses that can be tested with these datasets. Some of these hypotheses were (implicitly) formulated by the researchers who gathered the data. However, since the approach introduced in this book allows more flexibility (*viz.* construction of hypotheses using inequality constraints), additional hypotheses will be formulated. We will describe how traditional hypothesis testing could be used to evaluate these hypotheses. Limitations of more conventional approaches to hypothesis testing will also be addressed. In the chapters that follow, these hypotheses will be evaluated using Bayesian inequality constrained analysis of variance.

2.2 Amnesia in Dissociative Identity Disorder: The Investigation of a Controversy

In psychiatry and clinical psychology, the Diagnostic and Statistical Manual of Mental Disorders (DSM [1]) is probably the most frequently used system to

classify mental disorders. It includes specific criteria for dozens of disorders, subdivided into several categories among which are the categories of mood disorders, anxiety disorders, and personality disorders. There is a lot of controversy surrounding the system. Among other things, critics have noted that many of the disorders in DSM lack reliability and construct validity [40]. One of the most controversial disorders in the DSM is the Dissociative Identity Disorder (DID) – among the lay public also known as Multiple Personality Disorder. According to the last edition of the DSM, DID is defined as present when the person has at least two distinct identities or personality states that recurrently take control of the person’s behaviour and has an inability to remember important personal information that cannot be explained by ordinary forgetfulness [1].

The controversy surrounding DID basically comes down to the question if it is indeed possible that people can have two or more separable identities (so called “alters”), with currently dominant identities being amnesic for events experienced by the other identity. Some say that this is indeed possible, whereas others have questioned if this is indeed so (for a review see [20]). A lot of research has been conducted to study this topic. Yet, as with the disorder itself, much of this research has been criticized. For instance, some studies have simply asked one alter of a DID-patient if he/she remembered what was experienced by the other alter. Potentially problematic is that, say, subjective experience of amnesia does not necessarily reflect the objective presence of this phenomenon as present in people with, for instance, dementia or other organic mental disorders. To curb this problem, researchers have used implicit measures of amnesia that allow for an examination of amnesia without study participants being aware that amnesia is tested. Elegant examples of this approach are represented in several studies by Huntjens [16], who used experimental designs to answer the question if inter-identity amnesia reported by DID-patients represents true, objectively verifiable amnesia or is perhaps attributable to other processes. In the present book, one of the studies by Huntjens et al. [17] will be used for illustrative purposes in several chapters.

As a starting point of their study, Huntjens et al. [17] observed that it is still uncertain whether or not the amnesia of DID-patients is “real” amnesia, if it is iatrogenic (i.e., induced by therapists), or if it is caused by suggestive influence of media and cultures on suggestible individuals. Noticing limitations of extant studies on this topic, they felt it was timely to further examine the issue of symptom simulation in DID-patients. To this end, a group of DID-patients ($N_{pat} = 19$) and three controlgroups were subjected to a recognition task. In the first phase of this task, which was the learning episode, patients were subjected to part of the Wechsler Memory Scale-revised (i.e., the Logical Memory-story A and the Visual Reproduction subtests [39]); patients were told a brief story and they were shown several drawn figures after which they performed a recall test of both the story and the figures. Then, after a delay, patients were asked to switch to another alter that was subjected to the second phase of the task. In this phase, these other alters were subjected

to a recall test and a 15-item multiple-choice recognition test. Ten multiple-choice questions asked participants about particular story details, offering three possible answers for each question (e.g., “Was the story about a man, a woman, or an animal?”). Five questions asked participants to pick out the previously seen figure among five alternatives including four foils. The number of correct answers were summed to obtain a total “Recognition Score.” This score ranged from 0 to 15 and represented an index of remembering.

As noted, apart from the DID-patients ($N_{pat} = 19$), three control groups were included. The first control group was a normal control group ($N_{con} = 25$). The second group consisted of normal people who were asked to deliberately simulate inter-identity amnesia ($N_{sim} = 25$). One week before the experiment took place, participants in this group were extensively informed about DID and its possible iatrogenic nature. They were asked to make up an imaginary, amnesic identity and to practice switching from one identity to the other. The third control group consisted of normal people who only underwent the second phase of the experimental task and had to guess the right answers to the recognition questions ($N_{amn} = 25$). As such, they represented a “truly amnesic control group” in that they were truly unable to remember anything about the story because they didn’t get a chance to hear the story in the first place.

The design allowed the authors to compare the overall memory performance (recognition scores) among true DID-patients, Controls, Simulators, and True amnesiacs. There were at least two hypotheses implicated in the study. A first hypothesis was based upon the viewpoint that DID-patients actually suffer from “real amnesia.” From this viewpoint, it could be hypothesized that, in terms of differences in memory performance between groups, Controls would remember the story details best and that both DID-patients and True amnesiacs would perform worse than normals but would not differ from each other. Finally, Simulators could be expected to perform worst because they knew the correct answer and thus could deliberately choose a wrong answer. Their score could be expected to be worse than the score of the True amnesiacs who had to guess the right answer and obviously occasionally guessed right. Using equality and inequality constraints, this hypothesis could be represented as

$$H_{1a} : \mu_{con} > \{\mu_{amn} = \mu_{pat}\} > \mu_{sim}, \quad (2.1)$$

where μ denotes the mean recognition score in the group indicated and $>$ means larger than. A second hypothesis was based upon the viewpoint that DID-patients actually feign their amnesia. From this viewpoint, it could be hypothesized that the memory performance of DID-patients would be similar to that of Simulators and that both groups would have a poorer memory performance than both normals and True amnesiacs. This hypothesis could be depicted as

$$H_{1b} : \mu_{con} > \mu_{amn} > \{\mu_{pat} = \mu_{sim}\}. \quad (2.2)$$

Table 2.1. Recognition scores in the DID data

	<i>M</i>	<i>SD</i>	<i>N</i>
1. DID-patients	3.11	1.59	19
2. Controls	13.28	1.46	25
3. Simulators	1.88	1.59	25
4. True amnesiacs	4.56	1.83	25

As a means to enhance clarity on the validity of both hypotheses, Huntjens et al. [17] compared the memory performance (recognition scores) of the four groups, using analysis of variance (see [32] for a nice introduction) followed by pairwise comparisons of means (see [37] for a nice introduction). Table 2.1 describes the recognition scores in each group.

Analysis of variance can be used to test the hypothesis

$$H_0 : \mu_{con} = \mu_{amn} = \mu_{pat} = \mu_{sim} \quad (2.3)$$

versus

$$H_2 : \mu_{con}, \mu_{amn}, \mu_{pat}, \mu_{sim}, \quad (2.4)$$

that is, testing “nothing is going on” versus “something is going on but I don’t know what.” Note that this is not what the researchers wanted to know. They wanted to know which of the hypotheses H_{1a} and H_{1b} was the best. As can be seen in Table 2.2, the significance of H_0 is .00, implying that “something is going” on. To further clarify this, Scheffe’s posthoc tests were computed. All pairwise tests had significance smaller than .05, except the comparison of DID-patients with Simulators (a significance of .11). We additionally conducted other pairwise comparison procedures (Tukey HSD, Sidak, Gabriel, Hochberg), which all rendered the same result. These results are in accordance with H_{1b} but not with H_{1a} . Further support for H_{1b} is obtained from a visual inspection of the means in the four groups in Table 2.1: Controls indeed scored higher than True amnesiacs and both groups scored higher than DID-patients and Simulators.

All in all, outcomes indicated that, in terms of memory performance, the performance of DID-patients was worse than the performance of True amnesiacs and close to those who simulated DID. These findings led Huntjens et

Table 2.2. Significances for the analysis of variance and Scheffe’s post-hoc tests

Hypothesis	Significance
$H_0 : \mu_{con} = \mu_{amn} = \mu_{pat} = \mu_{sim}$.00
$H_0 : \mu_{con} = \mu_{pat}$.00
$H_0 : \mu_{pat} = \mu_{sim}$.11
$H_0 : \mu_{amn} = \mu_{pat}$.04
$H_0 : \mu_{con} = \mu_{sim}$.00
$H_0 : \mu_{con} = \mu_{amn}$.00
$H_0 : \mu_{amn} = \mu_{sim}$.00

al. [17] to conclude that DID-patients are similar to Simulators in providing incorrect answers to questions about the recognition of information to which they were previously subjected.

The study represents an elegant example of testing amnesia in DID-patients, without them being made aware that amnesia is tested. As such, it adds to our knowledge of DID in showing that it is an oversimplification to say that DID-patients indeed suffer from true amnesia. However, it is questionable whether hypothesis testing using analysis of variance and Scheffe's post hoc tests is the most elegant way to evaluate H_{1a} and H_{1b} (note that the authors did not have access to the Bayesian procedures that will be introduced in this book). First of all, the authors were not particularly interested in H_0 and H_2 . Hence, testing these hypotheses is a rather indirect manner to evaluate H_{1a} and H_{1b} . Second, testing H_0 versus H_2 did not provide all answers: post hoc tests needed to be executed, followed by a visual inspection of the sample means, in order to be able to reach a conclusion. A number of issues threaten the viability of this procedure:

- The results might not be in agreement with either H_{1a} or H_{1b} , in which case no conclusion can be obtained.
- For the example at hand, the procedure consists of testing seven hypotheses. Of course Scheffe can be used to control the probability of type I errors, also known as errors of the first kind (i.e., the probability of incorrectly rejecting null hypotheses), but also leads to a reduction of power, which will increase the probability of type II errors, also known as an errors of the second kind (i.e., incorrectly accepting one or more null hypotheses). Since power issues are important in psychological research because sample sizes are often limited, this is an undesirable feature of procedures that are used to control the amount of type I errors.
- A visual inspection of means is not a well-established formalized procedure that can be used to evaluate informative hypotheses (i.e., hypotheses formulated using inequality constraints). If four means are expected to be ordered from small to large, and the sample means are 0.2, 0.1, 4, and 8, a formal procedure might show substantial support for the expectation even though the order of the first two means is reversed.
- Finally, the results of pairwise comparisons of means may be inconsistent. Results like "accept $H_0: \mu_A = \mu_B$," "accept $H_0: \mu_B = \mu_C$," and "reject $H_0: \mu_A = \mu_C$ " cannot be straightforwardly interpreted because they are logically inconsistent: The three conclusions cannot be simultaneously true (see also [12]).

In the next chapters, Bayesian inequality constrained analysis of variance will be introduced. There are a number of differences between this Bayesian and traditional analysis of variance:

- The Bayesian approach requires a researcher to translate a number of competing theories into inequality constrained hypotheses before looking

at the data (see (2.1) and (2.2) and the examples that will be given in the next sections).

- Subsequently the data will be used to quantify the support for each hypothesis under investigation. This quantification is called the posterior probability and will be introduced and discussed in Chapter 4 and later chapters, the interested reader is also referred to Chapter 7 of [14]. For the example at hand it might turn out that the posterior probabilities of H_{1a} and H_{1b} are .2 and .8, respectively. Such a result would imply that after observing the data, H_{1b} is four times as likely as H_{1a} .
- With the Bayesian approach, a conclusion with respect to H_{1a} and H_{1b} will always be obtained (a possible conclusion is that both are equally likely). There is no multiple testing problem and a visual inspection of means is not necessary in order to reach a conclusion.

However, Bayesian analysis can benefit from an evaluation of H_0 and H_2 in addition to H_{1a} and H_{1b} : If neither H_{1a} nor H_{1b} is a better model than H_2 , it can be concluded that both sets of constraints are not in accordance with the data. Furthermore, if H_0 is a better model than H_{1a} and H_{1b} , the conclusion for the present DID example could be that the experimental manipulation has completely failed.

Note, finally, that there is an alternative to the formulation of H_{1a} and H_{1b} :

$$H_{1c} : \mu_{con} > \{\mu_{amn}, \mu_{pat}\} > \mu_{sim} \quad (2.5)$$

and

$$H_{1d} : \mu_{con} > \mu_{amn} > \{\mu_{pat}, \mu_{sim}\}. \quad (2.6)$$

Here it is no longer required that $\mu_{amn} = \mu_{pat}$ but only that they are located between μ_{con} and μ_{sim} . Something similar can be observed for H_{1b} . The alternative formulation is less restrictive than the original, without altering the core message of each hypothesis. Which formulation should be preferred, if any and whether or not it is wise to include H_0 and H_2 in the set of hypotheses under investigation will be further discussed in Chapter 4.

2.3 The Effect of Peer Evaluation on Mood in Children High and Low in Depression

A second study that will be used to illustrate the subsequent chapters addressed the influence of depression severity on emotional reactivity after different types of peer evaluation feedback in preadolescent children [26]. The interplay of moods (diffuse, slow moving feeling states) and emotions (quick moving reactions) is an intriguing study topic. Intuitively, it makes sense to think that moods potentiate like-valenced or matching emotions in a way that irritable mood strengthens angry reactions, anxious mood facilitates the experience of panic, and depressed mood facilitates reactions of sadness [29].

Yet, studies on the interplay of moods and emotions do not unambiguously support this so-called mood-facilitation hypothesis; there is evidence for this mood facilitation effect in anxiety [2], but little evidence that depressed mood inflates depressed emotional reactivity [29].

Reijntjes et al. [26, 27] studied the interaction between depressed mood and emotions in preadolescent children as one of the topics in a larger research program designed to enhance knowledge on emotional reactivity, emotional regulation, and depression in children. In this program, social evaluation by peers, manipulated by the authors, was chosen as a means to generate change in emotions. The reason to choose peer evaluation as a way of manipulating affect was chosen, because peer praise and rejection are common emotion-eliciting events in childhood [11] and exert an important influence on the development and maintenance of both externalizing problems (e.g., conduct disorder) [13] and internalizing problems (e.g., depression, cf. [21]). For the purpose of the present book, we will focus on a small aspect of this research program, in which the link between depressed mood and changes in current emotions in response to the manipulated peer evaluation was explored. This issue was addressed in detail in one of the studies of Reijntjes et al. [27].

In this study, 139 children, between the ages of 10 and 13 years, were led to believe that they were participating in an Internet version of a peer evaluation contest that was based on and named after an American television show called *Survivor*. Each participant was seated in front of a laptop computer and told that he/she participated in the online computer contest, together with four same-sex contestants. In actuality, contestants were fictitious. The objective of the game was presented as getting the highest “likeability” score from a jury consisting of 16 members, 8 boys and 8 girls. To this end, children were asked to provide information about themselves. They were asked questions about their favourite musical group, hobbies, and future occupation, and a number of character traits (e.g., sense of humour, agreeableness, intelligence, trustworthiness). Apart from that, their picture was taken with a web-cam.

All participants were led to believe that all the information and the picture would be transmitted to the judges over the Internet, who would then give them a “likeability” score ranging from 0 to 100. A short while after transmitting the information, the outcome of the likeability contest was presented. Specifically, the names of the players with the highest and the lowest score appeared on the screen. This is where the manipulation of the authors came in, that is, all participants were randomly assigned to a success feedback, a failure feedback, and a neutral feedback condition. Participants in the *success feedback* condition were told that they had obtained the highest score, those in the *failure feedback* condition were led to believe that they got lowest score, and those in the *neutral feedback* condition were told that they received neither the highest nor the lowest score. This last condition represented a control condition.

All participants completed a number of questionnaires over the course of the experiment. For the assessment of depressed mood, they completed the

Children’s Depression Inventory (CDI) before the contest. The CDI is a 27-item measure for the assessment of social, behavioural, and affective symptoms of depression in children [19]. Each of the 27 items asks children to pick out the statement that applies to them from 3 alternatives (e.g., I like myself, I do not like myself, I hate myself). For each question, the first answer (absence of the symptom) is rated as 0, the second answer (mild symptom) is rated as 1, and the third answer (definite symptom) is rated as 2. Item scores are summed to form an overall depression score that can range from 0 to 54.

To assess changes in affect induced by the peer evaluation, children completed the Positive And Negative Affect Schedule (PANAS; [38]) before and after the contest. In the current illustration, we focus on the Positive Affect scale, the PANAS-P. This scale presents children with 10 mood-related adjectives (e.g., enthusiastic, active, alert) and asks them to rate the extent to which these moods are experienced on 5-point scales ranging from “very slightly or not at all” to “extremely”. Scale scores can range from 10 to 50.

The data of Reijntjes et al. [27] allowed us to test several hypotheses about the influence of depressed mood on emotional reactivity induced by peer evaluation. The *improvement in affect/emotion*, defined as post-Survivor positive affect minus pre-Survivor positive affect, represents emotional reactivity and was the dependent variable. Positive values of this emotional reactivity (like in the high depressed success feedback group; see Table 2.3) indicate an improvement in positive mood, and negative values (like in the low depressed failure feedback group) indicate a decrease in positive mood. To compare between children with different levels of depression, the sample was divided in three groups of equal size: the “Low Depressed” group, the “Moderately Depressed” group, and, the “High Depressed” group. The hypotheses for this study addressed differences in emotional reactivity for children in these three groups, who were subjected to success, failure, or neutral peer feedback. This created nine groups of children. Table 2.3 shows the labelling of these groups, as well as the mean Emotional Reactivity Scores across groups.

We were particularly interested in reactivity after success feedback and after failure feedback. Mood improvement after success feedback is defined as emotional reactivity in the success condition minus emotional reactivity in the neutral condition ($\mu_1 - \mu_2$, $\mu_4 - \mu_5$, and $\mu_7 - \mu_8$). Note that the subscripts correspond to the group labelling displayed between [.] in Table 2.3. Positive differences between means indicate that mood improved more in the success condition than in the neutral condition. Negative differences indicate the opposite. Similarly, mood improvement after failure feedback is defined as emotional reactivity in the failure condition minus emotional reactivity in the neutral condition ($\mu_3 - \mu_2$, $\mu_6 - \mu_5$, and $\mu_9 - \mu_8$). Here positive differences between means indicated that mood improved more in the failure than in the neutral condition, and negative differences indicate that the opposite occurred. The neutral condition was chosen as a reference group because it may well be that the emotional reactivity in the neutral condition is different for different levels of depression. Our interest was not primarily in whether

Table 2.3. Emotional reactivity: Mean (M), standard deviation (SD), and sample size (N) per subgroup of depression by feedback condition

Depression	Feedback condition								
	Positive			Neutral			Negative		
	M	SD	N	M	SD	N	M	SD	N
Low	[1] 0.27	4.67	18	[2] 0.29	4.76	17	[3] -9.33	8.77	12
Moderate	[4] 0.41	4.96	12	[5] -1.50	5.99	14	[6] -5.78	5.75	19
High	[7] 5.76	4.39	17	[8] -0.56	4.25	16	[9] -3.85	6.49	14

Note: Numbers in square brackets denote the group labelling as used in the hypotheses.

emotional reactivity differs among depression levels in the success and failure conditions, but whether these differences persisted if we controlled for differences in emotional reactivity in the neutral condition.

Considering the issue of emotional reactivity, we could define at least three competing theories on the influence of depressed mood on emotional reactivity after peer praise (i.e., success feedback) and peer rejection (i.e., failure feedback). As noted, the mood-facilitation hypothesis states that mood potentiates matching emotions. Based on this hypothesis, we could expect that success feedback would elicit a less pronounced mood improvement in depressed compared to nondepressed children and that failure feedback would generate a more pronounced worsening of mood in depressed compared to nondepressed children; cf. [28]. Translated into an inequality constrained hypothesis this becomes

$$\begin{aligned}
 H_{1a} : \{ \mu_7 - \mu_8 \} &< \{ \mu_4 - \mu_5 \} < \{ \mu_1 - \mu_2 \}, \\
 \{ \mu_9 - \mu_8 \} &< \{ \mu_6 - \mu_5 \} < \{ \mu_3 - \mu_2 \}.
 \end{aligned}
 \tag{2.7}$$

One alternative hypothesis can be drawn from recent research on emotional context insensitivity in depression. Emotional context insensitivity comes down to the notion that the presence of depression causes attenuated emotional reactivity to both positive and negative stimuli; that is, depressed mood states are said to prompt withdrawal and to cause reduction in motivated activity and pessimism, such that both positive and negative triggers from the environment lead to little or no emotional reactivity [30]. In a recent study among adults, Rottenberg et al. [31] found evidence that depression coincides with such insensitivity. Based on this viewpoint, it could be expected that success feedback would lead to a weaker increase in mood in depressed compared to nondepressed children and that failure feedback would lead to stronger mood improvement (or, stated otherwise, failure feedback has less impact on depressed children and will thus lead to a smaller reduction in positive mood) in depressed compared to nondepressed children:

$$\begin{aligned}
H_{1b} : \{ \mu_7 - \mu_8 \} &< \{ \mu_4 - \mu_5 \} < \{ \mu_1 - \mu_2 \}, \\
\{ \mu_9 - \mu_8 \} &> \{ \mu_6 - \mu_5 \} > \{ \mu_3 - \mu_2 \}.
\end{aligned}
\tag{2.8}$$

We could think of an additional third hypothesis, one that is interesting but not explicitly mentioned and examined in existing literature. It is possible that, since depressed mood coincides with negative cognitions about self-worth [3], children higher in depression expected to get lower scores from their peers in the Survivor contest. From this viewpoint it could be expected that success feedback would actually lead to a stronger increase in mood in depressed compared to nondepressed children, because this success was highly discrepant with their self-view and expectations. In a related vein, it would be possible that failure feedback would lead to a weaker mood reduction (or, taking into account the way that we have defined our dependent variable, a stronger mood improvement) in depressed compared to nondepressed children, because, at some level, the depressed children were already expecting to fail; cf. [15]. Translated into an inequality constrained hypothesis, what could be called discrepancy hypothesis could be formulated as

$$\begin{aligned}
H_{1c} : \{ \mu_7 - \mu_8 \} &> \{ \mu_4 - \mu_5 \} > \{ \mu_1 - \mu_2 \}, \\
\{ \mu_9 - \mu_8 \} &> \{ \mu_6 - \mu_5 \} > \{ \mu_3 - \mu_2 \}.
\end{aligned}
\tag{2.9}$$

Evaluation of H_{1a} , H_{1b} , and H_{1c} using a traditional two-way analysis of variance does not render a straightforward evaluation of the hypotheses of interest. Both main effects are significant ($p = .00$ for the feedback condition; and, $p = .01$ for the depression condition) but the interaction effect is not (a significance of .08).

Using Figure 2.1 to interpret the significant main effects, the following conclusions seem valid: The main effect of feedback condition can be described as a decrease of emotional reactivity from the success via the neutral to the failure feedback condition. The main effect of depression level is less clear, but there is a tendency for more emotional reactivity in children with high levels in depression compared to those with moderate and low levels of depression. For a number of reasons, it is questionable whether these results can and should be used to evaluate H_{1a} , H_{1b} , and H_{1c} :

- The significant main effects do not address emotional reactivity in the success and failure condition corrected for emotional reactivity in the neutral condition.
- As mentioned earlier, if the null hypothesis (“nothing is going on”) is rejected in favour of the alternative hypothesis (“something is going on but I don’t know what”), a further inspection of tabled data summaries (e.g., a table of means) or a visual representation of the data (e.g., using figures like Figure 2.1) may be needed in order to determine what is going on. Evaluation of tabled data summaries and figures is never straightforward.

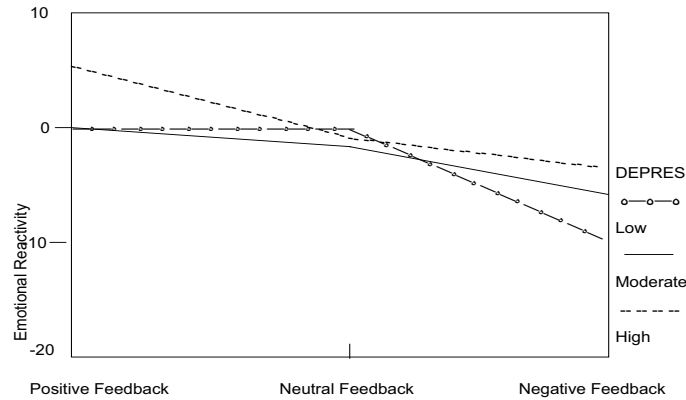


Fig. 2.1. Visual display of average emotional reactivity in the nine experimental groups

Without additional testing it is not clear whether the main effect of depression level was significant because of mean differences in one or more of the success, neutral, and failure feedback conditions. If, for example, standard errors of the means in the success and failure feedback condition would be large compared to the standard errors in the neutral feedback condition, the main effect of depression level would mainly be caused by the neutral feedback condition, which would change the “prima facie” interpretation of Figure 2.1 such that it is not in accordance with either of the three hypotheses.

- With additional testing to support the interpretation of Figure 2.1 (e.g., a pairwise comparison of means within each of the three feedback conditions), there is again a multiple testing problem; that is, due to the fact that more than one hypothesis is tested, the probability of type I errors will increase. This can be remedied using procedures that control the probability of type I errors like Scheffe and Bonferroni. However, this will lead to an increase of type II errors, that is, a reduction in power (the probability to correctly reject the null hypothesis). Furthermore, results may not be consistent with either one of H_{1a} , H_{1b} , or H_{1c} .
- A better approach is to execute four smaller two-way analyses of variance. In each analysis the focus is on the interaction effect in the subtable consisting of the first or last two groups of the factors feedback condition and depression level. Consider, for example, the subtable consisting of low/moderate levels of depression, and success/neutral feedback, that is, groups 1, 2, 4, and 5. If $\{\mu_4 - \mu_5\} < \{\mu_1 - \mu_2\}$ like in H_{1a} , or $\{\mu_4 - \mu_5\} > \{\mu_1 - \mu_2\}$ like in H_{1c} , there should be an interaction among feedback and depression in the subtable at hand. However, the significance

of the interaction was .46. This implies that $\{\mu_4 - \mu_5\} = \{\mu_1 - \mu_2\}$, which is not consistent with either H_{1a} , H_{1b} , or H_{1c} . In a similar manner the interaction in the other three subtables were tested. Neither of these was significant (moderate/high and success neutral, $p = .09$; low/moderate and neutral/failure, $p = .10$; moderate/high and neutral/failure, $p = .12$) and thus not helpful to choose the best of H_{1a} , H_{1b} , or H_{1c} .

Briefly stated, it appears that traditional hypothesis testing using two-way analyses of variance is not very helpful when evaluating H_{1a} , H_{1b} , and H_{1c} . In the following chapters it will be shown that Bayesian inequality constrained analysis of variance renders informative results and is able to select the best of the three hypotheses. Like in the previous example, one can extend H_{1a} , H_{1b} , and H_{1c} with the traditional null hypothesis (i.e., “nothing is going on”) and an alternative hypothesis (i.e., “something is going on but I don’t know what”). As in the DID example these hypotheses could be used to verify if the experimental manipulation was at all successful and whether at least one of H_{1a} , H_{1b} , and H_{1c} is supported by the data in the sense that it is a better model than the unconstrained alternative hypothesis.

2.4 Coping with Loss: The Influence of Gender, Kinship, and Time from Loss

Gender differences in psychological problems after stressful life events have received considerable attention in the literature. A recent review of the literature on gender differences in posttraumatic stress disorder (PTSD) has shown that in the general population this disorder is more prevalent among women than among men, even when controlling for the fact that men generally have a greater chance of being confronted with events that are potentially traumatizing [36]. Stated otherwise, among people who are confronted with traumatic events, women have a greater chance of developing PTSD than do men. It has been argued that this difference may well be due to the fact that base rates of psychopathology are higher in women than in men. Indeed, studies have shown that women are generally more prone to develop anxious and depressive symptoms (e.g., [22, 33]). Nevertheless, there is evidence that women have a greater risk of developing problems, even when controlling pretrauma levels of distress [9].

The issue is obviously a complex one. For instance, some have noted that the relative severity of posttrauma psychopathology levels in women compared to men may well be at least partially attributable to the fact that men generally tend to underreport depressive and anxious symptoms [34, 36]. In addition, it is conceivable that gender differences in emotional problems after traumatizing events are moderated by demographic variables such as age and marital status and trauma-related variables such as the nature of the event and the severity of the event. With respect to the severity, it is noteworthy

that Kendler et al. [18] found women to have a considerable greater chance of developing depression after exposure to a relatively minor threat, whereas the excess risk in women nearly disappeared in groups exposed to a more severe threat.

So, although female gender seems to be a risk factor for the development of psychological problems after confrontation with adversity, this conclusion is far from definitive to say the least. As a third example of Bayesian inequality constrained analysis of variance, we focussed on gender differences in the consequences of a stress full life-event different from trauma, namely the loss of a loved one. More specifically, we sought to address gender differences in the development of complicated grief. Complicated grief refers to a group of grief-specific symptoms that have been found to be distinct from depression and anxiety and to be predictive of severe mental health impairments [24]. These include intense and persistent yearning, difficulties in accepting the loss, avoidance, and shattered worldview.

Little is known about gender differences in this area. To our knowledge, there are four studies that addressed this issue, using the same conceptualization of complicated grief and the same questionnaire to measure it. Very roughly, these generated different results: In one study there were indications that men were worse off after a loss [4], in another study no differences were found [5], and two studies showed that women suffered more [10, 41]. As an additional illustration of Bayesian approaches, we examined the impact of gender further. In doing so, we focussed on the consequences of losing a partner and the consequences of losing a child, given that these losses are generally regarded as the most devastating (and mostly studied) losses that people may suffer [35]. Our aim was not to get a definitive answer to the question “who suffers more?” but, as with the other studies described in this chapter, to illustrate how the Bayesian inequality constrained analysis of variance can be used to test hypotheses about scores of groups on a single variable.

There were two hypotheses that guided our examination. The first of these is that women have a greater risk of developing complicated grief than do men. This hypothesis was fuelled by the fact that two of four studies that addressed gender differences found women to suffer more and by the fact that research on PTSD has shown that women are more at risk for problems after other types of adversity [36]. The second hypothesis was fuelled by the notion that the loss of partner or child may well work out differently for men and women. It has been claimed that women generally grow more attached to their children than do men [8]. This could make them more prone to develop emotional complications after loss. At the same time, there are reasons to believe that men are more vulnerable to get problems after the loss of a partner for instance, because their well-being is more strongly dependent on their relationship with the partner than that of women (cf. [25]). So, our second key hypothesis was that the influence of gender is moderated by kinship such that women suffer more after losing a child and men suffer more after losing a partner. In keeping with this notion, the study that addressed gender differences in parental bereaved

Table 2.4. Complicated grief: Mean (M), standard deviation (SD) and sample size (N) per subgroup of the Gender by Kinship by Time from loss design

Gender	Kinship	Time from loss							
		Recent			Remote				
		M	SD	N	M	SD	N	N	
Men	Partner	[1]	84.91	21.59	106	[2]	78.60	20.31	131
	Child	[3]	79.77	21.88	26	[4]	77.79	22.37	52
Women	Partner	[5]	86.42	18.56	229	[6]	78.36	19.28	374
	Child	[7]	84.88	17.33	91	[8]	83.02	21.74	165

Note: Numbers in square brackets denote the group labelling as used in the hypotheses.

individuals showed that women suffered more [41], whereas such results were not found in two studies that focussed on conjugal loss [4, 6].

To test these hypotheses, data were obtained from research programs on grief conducted by the first author of this chapter. We selected 760 bereaved partners and parents from a group of 1321 mourners who participated in a study on the distinctiveness of complicated grief, depression, and anxiety [6] and 419 similarly bereaved individuals from a group of 943 mourners included in a study on the dimensionality of complicated grief [7]. The final sample for this illustration thus included $N = 1179$ mourners.

To assess complicated grief symptoms, all participants completed the Dutch version of the Inventory of Complicated Grief-revised (ICG-r). The ICG-r is a self-report measure that taps each of the symptoms of complicated grief as well as other potentially problematic responses to loss [23]. The 29-item Dutch version was developed by Boelen et al. [7]. Examples of items are “I feel that I have trouble accepting the death” and “I feel myself longing and yearning for the lost person.” Respondents rate the presence of these symptoms in the preceding month, on 5-point scales ranging from “almost never” to “always.” The overall complicated grief severity score is calculated by summing item scores. This score ranged from 29 to 145.

As we wished to include possible effects of time from loss, the group was divided into those who were bereaved less than 1 year (the “recent loss” group) and those whose losses occurred more than 1 year ago (the “remote loss” group). Table 2.4 shows complicated grief severity scores across the groups that were included in the analyses.

As noted, our key hypotheses were that (a) women generally have a higher risk of developing severe complicated grief symptoms than do men and, alternatively, (b) that losing a child leads to higher complicated grief levels in women, whereas the loss of a partner leads to higher complicated grief levels in men. Yet, for illustrative purposes, we included these hypotheses in a sequence of expectations, also including (main) effects of time from loss and kinship to the deceased. The first hypothesis in this sequence was based on the fact that studies have convincingly shown that, in general, complicated

grief levels are stronger in the early months of bereavement than later [24]. So, the first hypothesis was formalized as follows:

- The directional (an inequality constraint is used to specify the direction of the effect) simple effect of time in the three-factor design at hand; that is, recent loss leads to more grief than remote loss: H_{1a} : $\mu_1 > \mu_2$, $\mu_3 > \mu_4$, $\mu_5 > \mu_6$, $\mu_7 > \mu_8$. Note that the subscripted numbers refer to the group numbers in Table 2.4 listed between [.]

To investigate our first key hypothesis that women always have higher complicated grief levels than men, H_{1a} could be extended with the following:

- The directional simple main effect of gender in the three-factor design at hand; that is, the grief scores for the women are always higher than the grief scores for the men in the corresponding groups to render H_{1b} : $\mu_5 > \mu_1$, $\mu_6 > \mu_2$, $\mu_7 > \mu_3$, $\mu_8 > \mu_4$, $\mu_1 > \mu_2$, $\mu_3 > \mu_4$, $\mu_5 > \mu_6$, $\mu_7 > \mu_8$. Note that this is a model containing the simple main effects of time and gender.

Although there may be an interaction between gender and kinship, some scholars have claimed that, both for men and for women, losing a child poses a stronger risk for the development of complicated grief than does losing a partner (cf. [41]). This could be investigated via an extension of H_{1b} with the following:

- The directional simple effect of kinship in the three-factor design at hand; that is, losing a child leads to more complicated grief than losing a partner: H_{1c} : $\mu_3 > \mu_1$, $\mu_4 > \mu_2$, $\mu_7 > \mu_5$, $\mu_8 > \mu_6$, $\mu_5 > \mu_1$, $\mu_6 > \mu_2$, $\mu_7 > \mu_3$, $\mu_8 > \mu_4$, $\mu_1 > \mu_2$, $\mu_3 > \mu_4$, $\mu_5 > \mu_6$, $\mu_7 > \mu_8$. Note that this is a model containing the simple main effects of time, gender, and kinship.

Whether losing a child is more devastating for women could be investigated with the following hypothesis, which contains three elements:

- The simple effect of time (i.e., H_{1a}), the notion that losing a child is more severe for women than losing a partner and the notion that losing a child is more severe for women than for men. Together these three components render the following hypothesis: H_{1d} : $\mu_1 > \mu_2$, $\mu_3 > \mu_4$, $\mu_5 > \mu_6$, $\mu_7 > \mu_8$, $\mu_7 > \mu_5$, $\mu_8 > \mu_6$, $\mu_7 > \mu_3$, $\mu_8 > \mu_4$.

Our second key hypothesis that losing a child is more devastating for women, whereas men are worse off after partner loss could be investigated via an extension of the previous hypothesis with two elements:

- Complicated grief levels of men who lose a partner are higher than complicated grief levels of men who lose a child, and complicated grief levels of men who lose a partner are higher than complicated grief levels of women who lose a partner. This renders the following model: H_{1e} : $\mu_1 > \mu_2$, $\mu_3 > \mu_4$, $\mu_5 > \mu_6$, $\mu_7 > \mu_8$, $\mu_7 > \mu_5$, $\mu_8 > \mu_6$, $\mu_7 > \mu_3$, $\mu_8 > \mu_4$, $\mu_1 > \mu_3$, $\mu_2 > \mu_4$, $\mu_1 > \mu_5$, $\mu_2 > \mu_6$.

Table 2.5. Three-way analysis of variance of complicated grief scores

Effect	Significance
Main effect of Gender	.064
Main effect of Time	.004
Main effect of Kinship	.652
Interaction between Gender and Time	.794
Interaction between Gender and Kinship	.148
Interaction between Time and Kinship	.093
Three-way interaction Gender, Time, Kinship	.765

As in the previous two examples, it may be useful to add the traditional null hypothesis and an unconstrained hypothesis to H_{1a} until H_{1e} . The unconstrained hypothesis $H_2: \mu_1, \dots, \mu_8$ is particularly useful in this example. If none of the models under investigation has a better fit than H_2 , none of the sets of constraints is supported by the data.

It is difficult to obtain information with respect to the inequality constrained hypotheses under investigation, using traditional analysis of variance. A straightforward three-way analysis of variance does not test the hypotheses under investigation (combinations of sets of directional simple effects where inequality constraints are used to specify the direction of the effects) since it evaluates only main, two- and three-way interaction effects. For illustrative purposes we executed such a three-way analysis of variance; the results are displayed in Table 2.5. As can be seen, only the main effect of time was found to have a significance smaller than .05. Although this provides some support for H_{1a} , in subsequent chapters it will be shown that using these results to evaluate H_{1a} through H_{1e} will render misleading results, because the hypotheses tested are not directly related to the hypotheses under investigation.

Another approach could be to apply Scheffe's procedure for pairwise comparisons of means to the eight groups that constitute the three-way design at hand (note that other pairwise comparison procedures rendered similar results). The results with a significance smaller than .10 are displayed in Table 2.6. As can be seen, the results are not clearly in agreement with one of the hypotheses under investigation. Furthermore, as will be shown in subsequent chapters, it would be wrong to conclude that none of the hypotheses under investigation is supported by the data. In following chapters it will also be shown how Bayesian inequality constrained analysis of variance can straight-

Table 2.6. Pairwise comparisons with a significance smaller than .10 for the complicated grief scores

Group	Group	Significance
Women, Recent, Partner vs. Women, Remote, Partner		.002
Women, Recent, Partner vs. Men, Remote, Partner		.077

forwardly be used to determine to which degree each hypothesis is supported by the data.

2.5 Conclusions

In the context of analysis of variance models this chapter has illustrated at least two things. First, theories can often be translated into hypotheses that are formulated using inequality constraints among a set of means. Second, it may be difficult, relatively ad hoc, or even impossible to evaluate these hypotheses using traditional null hypotheses testing. The following features appeared in the three examples that have so far been discussed:

- Traditional null (“nothing is going on”) and alternative hypotheses (“something is going on but I don’t know what”) do not render straightforward evaluations of the informative (specified using inequality constraints) hypotheses under investigation.
- An informal evaluation of the sample means is often needed in addition to traditional hypothesis testing to be able to evaluate informative hypotheses. Since informal analysis, sometimes also referred to as “eyeball testing,” can be misleading, this is not a proper mode of inference.
- Results obtained with traditional hypothesis testing may not be in agreement with any of the informative hypotheses under investigation.
- Often multiple smaller null hypotheses and alternative hypotheses have to be tested in order to be able to evaluate informative hypotheses. There are two drawbacks of multiple hypothesis testing: It is hard to control both the type I and type II errors and results obtained using multiple hypothesis testing may be mutually inconsistent.

In the following chapters it will be shown that it is relatively easy to evaluate a set of informative hypotheses using Bayesian inequality constrained analysis of variance. The use of inequality constraints is not limited to analysis of variance type models. In the third part of the book it will be shown how inequality constrained hypotheses can be formulated in the context of analysis of covariance models, latent class models, multilevel models, and models for contingency tables.

The following types of constraints have been used in this chapter:

- Mean restricted to be larger than another mean (e.g., $\mu_1 > \mu_2$).
- Combination of means restricted to be larger than another combination of means (e.g., $\{\mu_1 - \mu_2\} > \{\mu_3 - \mu_4\}$). Note that the restriction in which the minus sign is replaced by a plus sign is also possible.

Restrictions that have not been used but that can also be helpful to translate a theory into an inequality constrained hypothesis are as follows:

- Mean restricted to be larger than a number (e.g., $\mu > 0$). This restriction will be used in Chapter 13 of this book, which deals with inequality constrained multilevel analysis.
- Mean restricted to be larger than another mean plus a number, for instance, $\mu_1 > \{\mu_2 + 2\}$. This restriction can be used if researchers want to include a minimally required effect size in their hypotheses; that is, the first mean should be at least 2 larger than the second mean. The number 2 is not the traditional effect size in terms of Cohen's d , in which case it would be written as $\frac{\mu_1 - \mu_2}{\sigma} > d$. However, rewriting Cohen's d , we obtain $\mu_1 - \mu_2 > d\sigma$; stated otherwise, the number 2 is Cohen's d multiplied with the within-group standard deviation σ .
- Absolute difference between means restricted to be smaller than a number (e.g., $|\mu_1 - \mu_2| < 2$). This restriction can be used to specify that two means are not relevantly different, where, again, researchers should specify the minimally required effect size needed for two means to be required relevantly different.

All these types of restrictions can be combined, thus rendering a flexible tool that can be used to translate one or more theories into a number of competing hypotheses. In this book many examples will be given, and references to software with which these hypotheses can be evaluated will be provided.

References

- [1] American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders (4th ed. Text Revision). Washington, DC, American Psychiatric Association (2000)
- [2] Barlow, D.H.: Anxiety and its Disorders: The Nature and Treatment of Anxiety and Panic (2nd ed). New York, Guilford Press (2002)
- [3] Beck, A.T., Rush, A.J., Shaw, B.F., Emery, G.: Cognitive Therapy of Depression. New York, Guilford Press (1979)
- [4] Bierhals, A.J., Prigerson, H.G., Fasiczka, A., Frank, E., Miller, M., Reynolds III, C.: Gender differences in complicated grief among the elderly. *Omega: Journal of Death and Dying*, **32**, 303–317 (1996)
- [5] Boelen, P.A., Bout, J. van den: Gender differences in traumatic grief symptom severity after the loss of a spouse. *Omega: Journal of Death and Dying*, **46**, 183–198 (2003)
- [6] Boelen, P.A., Bout, J. van den: Complicated grief, depression, and anxiety as distinct post-loss syndromes: A confirmatory factor analysis study. *American Journal of Psychiatry*, **162**, 2175–2177 (2005)
- [7] Boelen, P.A., Keijsers, J. de, Bout, J. van den: Psychometrische eigenschappen van de Rouw VragenLijst (RVL) [Psychometric properties of the Inventory of Traumatic Grief]. *Gedrag & Gezondheid*, **29**, 172–185 (2001)
- [8] Bowlby, J.: Attachment and Loss: Volume 3, Loss: Sadness and Depression. New York, Basic Books (1980)

- [9] Breslau, N., Davis, G.C., Andreski, P., Peterson, E.L., Schultz, L.R.: Sex differences in posttraumatic stress disorder. *Archives of General Psychiatry*, **54**, 1044–1048 (1997)
- [10] Chen, J.H., Bierhals, A.J., Prigerson, H.G., Kasl, S.V., Mazure, C.M., Jacobs, S.: Gender differences in the effects of bereavement-related psychological distress in health outcomes. *Psychological Medicine*, **29**, 367–380 (1999)
- [11] Coie, J.D.: Toward a theory of peer rejection. In: Asher, S.R., Coie, J.D. (eds) *Peer Rejection in Childhood* (pp. 365–401). New York, Cambridge University Press (1990)
- [12] Dayton, C.M.: Information criteria for pair wise comparisons. *Psychological Methods*, **8**, 61–71 (2003)
- [13] Dodge, K.A., Lansford, J.E., Salzer Burks, V., Bates, J.E., Pettit, G.S., Fontaine, R., Price J.M.: Peer rejection and social information-processing factors in the development of aggressive behavior problems in children. *Child Development*, **74**, 374–393 (2003)
- [14] Gill, J.: *Bayesian Methods. A Social and Behavioral Sciences Approach*. London, Chapman & Hall (2002)
- [15] Gladstone, T.R.G., Kaslow, N.J.: Depression and attributions in children and adolescents: A meta-analytic review. *Journal of Abnormal Child Psychology*, **23**, 597–606 (1995)
- [16] Huntjens, R.J.C.: *Apparent Amnesia. Interidentity Memory Functioning in Dissociative Identity Disorder*. Ph.D. thesis, Utrecht University (2003)
- [17] Huntjens, R.J.C., Peters, M.L., Woertman, L., Bovenschen, L.M., Martin, R.C., Postma, A.: Inter-identity amnesia in dissociative identity disorder: A simulated memory impairment? *Psychological Medicine*, **36**, 857–863 (2006)
- [18] Kendler K.S., Kuhn J.W., Prescott C.A.: The interrelationship of neuroticism, sex, and stressful life events in the prediction of episodes of major depression. *American Journal of Psychiatry*, **161**, 631–636 (2004)
- [19] Kovacs, M.: Rating scales to assess depression in school-aged children. *Acta Paedo Psychiatria*, **46**, 305–315 (1981)
- [20] Lilienfeld, S.O., Lynn, S.J.: Dissociative identity disorder: Multiple personality, multiple controversies. In: Lilienfeld, S.O., Lohr, J.M., Lynn, S.J. (eds) *Science and Pseudoscience in Clinical Psychology*. New York, Guilford Press (2003)
- [21] Nolan, S.A., Flynn, C., Garber, J.: Prospective relations between rejection and depression in young adolescents. *Journal of Personality and Social Psychology*, **85**, 745–755 (2003)
- [22] Nolen-Hoeksema, S.: *Sex Differences in Depression*. Stanford, CA, Stanford University Press (1990)
- [23] Prigerson, H.G., Jacobs, S.C.: Traumatic grief as a distinct disorder: A rationale, consensus criteria, and a preliminary empirical test. In: Stroebe, M.S., Hansson, R.O., Stroebe, W., Schut, H.A.W. (eds) *Handbook of Bereavement Research. Consequences, Coping, and Care* (pp. 613–647). Washington, DC, American Psychological Association Press (2001)
- [24] Prigerson H.G., Vanderwerker L.C., Maciejewski P.K.: Prolonged grief disorder as a mental disorder: Inclusion in DSM. In Stroebe, M.S., Hansson,

- R.O., Stroebe, W., Schut, H.A.W. (eds) *Handbook of Bereavement Research and Practice: 21st Century Perspectives*. Washington, DC, American Psychological Association Press (in press)
- [25] Rando, T.A.: *Treatment of Complicated Mourning*. Champaign, IL, Research Press (1993)
- [26] Reijntjes, A.: *Emotion-Regulation and Depression in Pre-adolescent Children*. Ph.D. thesis, Free University Amsterdam (2004).
- [27] Reijntjes, A., Dekovic, M., Vermande, M., Telch, M.: The role of depressive symptoms in early adolescents online emotional responding to a peer evaluation challenge. *Depression and Anxiety* (in press)
- [28] Rosenberg, E.L.: Levels of analysis and the organization of affect. *Review of General Psychology*, **2**, 247–270 (1998)
- [29] Rottenberg, J.: Mood and emotion in major depression. *Current Directions in Psychological Science*, **14:3**, 167–170 (2005)
- [30] Rottenberg, J., Gotlib, I.H.: Socioemotional functioning in depression. In: Power, M. (ed) *Mood Disorders: A Handbook of Science and Practice* (pp. 61–77). New York, Wiley (2004)
- [31] Rottenberg, J., Gross, J. J., Gotlib, I. H.: Emotional context insensitivity in major depressive disorder. *Journal of Abnormal Psychology*, **114**, 627–639 (2005)
- [32] Rutherford, A.: *Introducing ANOVA and ANCOVA, a GLM Approach*. London, Sage (2001)
- [33] Shear, M. K., Feske, U., Greeno, C.: Gender differences in anxiety disorders: Clinical implications. In: Frank, E. (ed) *Gender and its Effects on Psychopathology* (pp. 151–165). Washington, DC, American Psychiatric Publishing (2000)
- [34] Sigmon, S.T., Pells, J.J., Boulard, N.E., Whitcomb-Smith, S., Edenfield, T.M., Hermann, B.A., LaMattina, S.M., Schartel, J.G., Kubik, E.: Gender differences in self-reports of depression: The response bias hypothesis revisited. *Sex Roles*, **53**, 401–411 (2005)
- [35] Stroebe, W., Schut, H.: Risk factors in bereavement outcome: A methodological and empirical review. In: Stroebe, M.S., Hansson, R.O., Stroebe, W., Schut, H. (eds) *Handbook of Bereavement Research. Consequences, Coping, and Care* (pp. 349–372). Washington, DC, American Psychological Association Press (2001)
- [36] Tolin, D., Foa, E.B.: Sex differences in posttraumatic stress disorder: A quantitative review of 25 years of research. *Psychological Bulletin*, **132**, 959–992 (2006)
- [37] Toothaker, L.E.: *Multiple Comparison Procedures*. London, Sage (1993)
- [38] Watson, D., Clark, L.A., Tellegen, A.: Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, **54**, 1063–1070 (1988)
- [39] Wechsler, D.: *Wechsler Memory Scale-Revised*. San Antonio, TX, The Psychological Corporation (1987)
- [40] Widiger, T.A., Clark, L.A.: Toward DSM-V and the classification of psychopathology. *Psychological Bulletin*, **126**, 946–963 (2000)
- [41] Wijngaards-de Meij, L., Stroebe, M., Schut, H., Stroebe, W., Bout, J. van den, Heijden, P. van der, Dijkstra, I.C.: Couples at risk following the death of their child: Predictors of grief versus depression. *Journal of Consulting and Clinical Psychology*, **73**, 617–623 (2005)