Informative Hypotheses

How to move beyond classical null hypothesis testing

Rens van de Schoot

Utrecht University
Facullty of Social Sciences
Heidelberglaan 1
3584CS Utrecht

This research is financed by
The Netherlands Organization for Scientific Research
(NWO-VICI-453-05-002).

ISBN 978-90-393-5367-7

Printed in the Netherlands by
ZuidamUithof Drukkerijen.

Explanation of cover. The picture on the cover is perhaps the best evidence to reject the hypothesis that the Earth is flat as was issued in the introduction of this dissertation. The question of the shape of the earth was a recurring issue in scientific debate during the era of Aristotle. I will tell the story of Aristotle's scientific investigations in the introduction chapter to set the stage for my main point: that more can be learned from data by using informative hypotheses than the traditional null hypothesis.

The photo on the front cover is taken from *www.manyfreewall papers.com.*

Informative Hypotheses

How to move beyond classical null hypothesis testing

Informatieve Hypothesen

Rechtstreeks Verwachtingen Evalueren of de Nul Hypothese Toetsen?

(met een samenvatting in het Nederlands)

proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. J.C. Stoof, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 29 oktober 2010 des ochtends te 10.30 uur

door

Adrianus Gerardus Joannes van de Schoot
geboren op 9 juli 1979
te Eindhoven

Promotor:	Prof. dr. H. Hoijtink
Co-promotor:	Dr. J-W. Romeijn

Leden van de Beoordelingscommissie: Prof. Dr. W. Meeus
Prof. Dr. R. Meijer
Prof. Dr. H. Kelderman
Prof. Dr. S. Hartmann

Contents

1

Introduction

Although Wainer (1999) argues in "One Cheer for Null Hypothesis Significance Testing" that traditional null hypothesis testing can be useful in certain cases, many researchers have no particular interest in this hypothesis ('nothing is going on'). So why test the null hypothesis if one is not interested in it? Cohen (1994) aptly summarised the criticism of traditional null hypothesis testing in the title of his paper "The earth is round ($p < .05$)", which stresses the fact that the null hypothesis is almost never a realistic representation of the population of interest. Let us elaborate on this using an example inspired by this title.[1]

The question of the shape of the earth was a recurring issue in scientific debate during the era of Aristotle (384BC-322BC Rusell, 1997). By that time, scientists were agreed that the ancient speculations in Persian writings that the earth was a seven-layered ziggurat or a cosmic mountain, were off

[1] The historical figure Aristotle never denied that the earth was round; in fact, from the third century B.C. onwards, no educated person in the history of Western civilization believed that the earth was flat (Rusell, 1997). Indeed, Erasthenes (276-195 B.C.) gave a reasonable approximation of the earth's circumference and provided strong support for the hypothesis that the earth is round. However, Aristotle was one of the first scientists to provide evidence of the earth's roundness.

the mark. In fact, the Greek idea that the earth was round had dominated scientific thinking since as early as the fifth century B.C.. The only serious opponents were the atomists Leucippus and Democritus, who still believed that the earth was a flat disk floating in the ocean, as certain ancient Mesopotamian philosophers had maintained.

Now let us embark on some historical science fiction. In what follows, we will tell the story of Aristotle's scientific investigations using different ways of evaluating hypotheses. In order to falsify the Mesopotamian hypothesis and provide evidence for the earth's roundness, we say that Aristotle might have used an approach based on testing the null hypothesis. Naturally, it is not our intention to provide an accurate historical account here. We simply want to provide an example and set the stage for our main point: that more can be learned from data by using informative hypotheses than the traditional null hypothesis.

1.1 Traditional Null Hypothesis Testing

If Aristotle had used traditional null hypothesis testing, he might have tested the following null and alternative hypotheses:[2]

- H_0: the shape of the earth is a flat disk,
- H_1: the shape of the earth is not a flat disk.

Aristotle would have gathered data about the shape of the earth and found evidence against the null hypothesis, for example: (1) stars that were seen in Egypt were not seen in countries north of Egypt, while stars that were never beyond the range of observation in northern Europe were seen to rise and set in Egypt; (2) in eclipses of the moon a curved shadow passed across the face of the moon; (3) as shown by Eratosthenes, the shadow of an obelisk at

[2] Clearly, these hypotheses are not statistical hypotheses and no actual statistical inference could be carried out; these fictitious hypotheses are purely designed to serve as an example.

Alexandria extended out from its base at noon during the summer solstice, while at the same time the sun was directly overhead at Syene, approximately 500 miles south of Alexandria. Together, these observations could not be taken as evidence of a flat earth. In sum, H_0 would have been rejected, leading Aristotle to conclude that the earth cannot be represented by a flat disk. But what can actually be learned from this conclusion? Not much! It tells us that the earth is not a flat disk, but we remain ignorant of the earth's actual shape. This ignorance is a result of the alternative hypothesis, which includes all shapes that are non-flat, for example spherical, but also pear-shaped, triangular, and so forth.

In actual fact, Aristotle agreed with Pythagoras (582BC - ca. 507BC), who believed that all astronomical objects have a spherical shape, including the earth. So, once again embarking on an episode of imaginary history, Aristotle could also have tested the following null and alternative hypotheses:

- H_0: the shape of the earth is a sphere,
- H_1: the shape of the earth is not a sphere.

At that time, no one could see the earth as a whole and know it to be a sphere by direct observation. But one can derive other conclusions from the hypothesis that the earth is a sphere and use these to test the null hypothesis. For example, one could predict that if someone sailed west for a sufficient amount of time, this person would come back to where they started (Magellan did this). Or one could predict that if the earth was a sphere, ships at sea would first show their sails above the horizon, and then later as they sailed closer, their hulls (Galileo observed this). These precise predictions, if exactly confirmed, would establish a provisional objective reality for the idea that the earth is a sphere.

Now, imagine that Aristotle continued his search for data and that he gathered data that yielded evidence against the null hypothesis: while standing on a mountain top, he noticed the many peaks and valleys and had already concluded that the shape of the earth was not a perfect

sphere. Apparently, the earth's surface has many irregularities and if enough irregularities are observed it could provide just enough evidence to reject the null hypothesis. Moreover, most large rock formations do not even vaguely resemble spheres so, by analogy, the earth could not be expected to have that shape either. And so it may have happened that Aristotle once again rejected the null hypothesis, concluding that the earth is not a sphere (Cohen: "The earth is round ($p < .05$)"). What can be learned from this conclusion? Again, not much! The null hypothesis has been rejected, but we are still ignorant with regard to the shape of the earth.

1.2 Falsification and Beyond

Admittedly, not all methodologists agree on this point. In response to Aristotle's imagined disappointment, Popper (1959) would have argued that this insight is all that Aristotelian science, or any science for that matter, can hope for. When it comes to general hypotheses, or hypotheses that are beyond the reach of direct verification (which, in the spaceship-deprived era of Aristotle was certainly the case), we can only be sure of their falsification. Direct positive evidence for hypotheses about the shape of the earth cannot be obtained, so there would be no reason for Aristotle to be disappointed. Popper would have argued there is no way to prove that the earth is spherical, therefore we can only hypothesize that it is the shape of a sphere. Since Aristotle found evidence demonstrating that the earth is not spherical, this hypothesis is rejected. In fact, according to Popperian reasoning, Aristotle should rejoice in the fact that at least he now knows the earth is not a sphere! Nevertheless, is that all we can learn from our observations?

We have gone through two falsifications and rejected both null hypotheses. However, it might be of interest to know which of the two falsified hypotheses in our example is best supported by the observations made. Rather than using the examples given above, we might argue that Aristotle was actually interested in evaluating the two hypotheses directly against each other:

- H_A: the shape of the earth is a flat disk,
- H_B: the shape of the earth is a sphere.

After comparing these two hypotheses with Aristotle's observations, we can deduce that of the two hypotheses, the second one is more likely and the shape of the earth is better represented by a sphere than by a flat disk. Using this approach, it would have been clear to Aristotle that his observations provided more support for H_B than for H_A. In other words, he would have learned something positive after all!

1.2.1 What Does This Historical Example Teach Us?

Evaluating your expectations directly produces more useful results than sequentially testing traditional null hypotheses against catch-all rivals. In this dissertation I take this one step further and show that researchers are in fact interested in the evaluation of what we will call 'informative hypotheses'. These are hypotheses that contain information about the ordering of parameters.

For example, consider an example taken from Strohmeier, Fandrem, Spiel and Stefanek (2009), see also, Fandrem, Strohmeier and Roland (2009) about the extent to which the goal of being accepted by friends is an underlying function of aggressive behaviour in adolescents, and whether this function is more predictive than reactive aggression for aggressive behaviour in first and second generation immigrants compared with the native population. In their introduction, the authors formulated clear expectations: "Concretely, the study investigates the predictive power of the goal to be accepted by friends as underlying function of aggressive behaviour in comparison with reactive aggression. We are interested whether the goal to be accepted by friends operates differently in native, first and second generation immigrant boys and girls. We speculate that the goal to be accepted by friends as underlying function for aggressive behavior might be more important for first generation immigrants compared with natives or second generation immigrants. This

is because first generation immigrants who migrated themselves and who experienced resettlement are more vulnerable regarding their peer relations than natives and second generation immigrants. This vulnerability might be a reason that they also use antisocial means - that is aggressive behaviour - to reach their goal to be affiliated with and accepted by their peers."

Note that 'more important' can be translated into an informative hypothesis: among first-generation immigrants, acceptance by friends is a stronger predictor of aggressive behaviour and reactive aggression is a weaker predictor of aggressive behaviour than among the other two groups. If you can formulate such a hypothesis, it is argued in this dissertation that you should evaluate this informative hypothesis directly.

Nonetheless, evaluating informative hypotheses presupposes that prior knowledge is available. If that were not the case, it would make no sense to evaluate informative hypotheses and testing the traditional null hypothesis would be appropriate. In most applied articles, however, prior knowledge is available in the form of expectations. Therefore, the assumption of availability of prior knowledge is not really an issue. M. D. Lee and Pope (2006) contend that much information is already known before data are collected (see also M. D. Lee & Wagenmakers, 2005). Jaynes (2003) puts it more dramatically: "If we humans threw away what we knew yesterday in reasoning about our problems today, we would be below the level of wild animals; we could never know more than we can learn in one day, and education and civilization would be impossible" (p. 87). These statements are supported by applied articles where researchers evaluated informative hypotheses, see Chapters 7 and 8, but see also Kammers, Mulder, De Vignemont and Dijkerman (2009); Meeus, Van de Schoot, Klimstra and Branje (2010); Van Well, Kolk and Klugkist (2009).

1.3 Different Types of Hypotheses

In this dissertation we discuss four different types of hypotheses. Let us illustrate these using an example from Van de Schoot, Velden, Boom and Brugman (2010). The authors investigated the association between popularity and antisocial behaviour in a large sample of young adolescents from preparatory vocational schools (VMBO) in the Netherlands. In this setting, young adolescents are at increased risk of becoming or becoming more antisocial. Five, so-called, sociometric status groups were defined in terms of a combination of social preference and social impact: a popular, rejected, neglected, controversial, and an average group of adolescents.

Suppose we want to compare these five sociometric status groups in terms of the number of committed offences reported to the police last year (minor theft, violence, and so on) and let μ_1 be the mean of the number of committed offences for the popular group, μ_2 for the rejected group, μ_3 for the neglected group, μ_4 for the controversial group and μ_5 for the average group. Different types of hypothesis can now be formulated.

Informative Hypotheses

The main focus of the current dissertation is on so-called *informative* hypotheses, denoted by H_{I_1}, H_{I_2}, . . ., H_{I_N} for N hypotheses. These are hypotheses containing information about the ordering of the means and can consist of the following constraints:

1. larger than, denoted by '>'
2. smaller than, denoted by '<'
3. equal, denoted by '='.

Such expectations about the ordering of parameters can stem from previous studies, a literature review or even academic debate. If no information is available about the ordering, this is denoted by a comma.

An informative hypothesis can consist of any combination of constraints. Consider an imaginary hypothesis that the controversial group reports the largest number of offences, followed by the rejected group, the average group and the popular group, while the neglected group has the lowest score. In this situation the informative hypothesis then becomes: $H_{I_1} : \mu_3 < \mu_1 < \mu_5 < \mu_2 < \mu_4$. Another expectation could be that the popular, rejected and average groups report more offences than the neglected group, and fewer offences than the controversial group. Where the popular, rejected and average group can have any score as long as it is lower than the neglected group and higher than the controversial group, $H_{I_2} : \mu_3 < \{\mu_1, \mu_5, \mu_2\} < \mu_4$.

Traditional null hypothesis

Second, there is the traditional null hypothesis (denoted by H_0), which states that nothing is going on, $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$.

Unconstrained hypothesis

If no constraints are imposed on any of the means, and any ordering is equally likely, the hypothesis is called unconstrained (denoted by H_U): $H_U : \mu_1 \ , \ \mu_2 \ , \ \mu_3 \ , \ \mu_4 \ , \ \mu_5$.

1.4 Evaluating Informative Hypotheses

In the literature different procedures are described that allow for the evaluation of informative hypotheses. First of all, there are approaches that render a p-value for the comparison of H_{I_n} (which sometimes has the role of null-hypothesis) with an alternative hypothesis. A good overview is given by the books of Barlow, Bartholomew, Bremner and Brunk (1972); Robertson, Wright and Dykstra (1988); and Silvapulle and Sen (2004). In addition, the *Journal of Statistical Planning and Inference* published in 2002 a special issue on testing inequality constraint hypotheses. Furthermore, an adaptation of

the classical F-test for analysis of variance (ANOVA) has been proposed by Silvapulle, Silvapulle and Basawa (2002), see also Kuiper and Hoijtink (2010); Silvapulle and Sen (2004). This new test is called the F-bar test, which is a confirmatory method to test one single informative hypothesis in two steps, for example:

$$\begin{aligned} H_0 &: \mu_3 = \mu_1 = \mu_5 = \mu_2 = \mu_4 \\ \text{versus} & \\ H_{I_1} &: \mu_3 < \mu_1 < \mu_5 < \mu_2 < \mu_4 \ , \end{aligned}$$

and

$$\begin{aligned} H_{I_1} &: \mu_3 < \mu_1 < \mu_5 < \mu_2 < \mu_4 \\ \text{versus} & \\ H_U &: \mu_3, \mu_1, \mu_5, \mu_2, \mu_4 \ , \end{aligned}$$

where in the second hypothesis test H_{I_1} serves as the null hypothesis.

A second way of evaluating an informative hypothesis is to use a model selection approach. Model selection is not a test of the model in the sense of hypothesis testing, rather it is an evaluation between models using a trade-off between model fit and model complexity. Given a data set, several competing statistical models may be ranked according to their value on the model selection tool used and the one with the best trade-off is the winner of the model selection competition. Problems with standard model selection tools arise because default model comparison tools, for example AIC (Akaike, 1981), BIC (Schwarz, 1978), and DIC (Spiegelhalter, Best, Carlin & Van Der Linde, 2002) are not equipped to deal with inequality constraints, see Chapters 3 and 6. Alternative model selection procedures are the Paired-Comparison Information Criterion (PCIC) proposed by Dayton (2003), see also an application in Taylor et al. (2007). The literature contains also one modification of Akaikes information criterion that can be used in the context of inequality constrained analysis of variance models. It is called the order-restricted information criterion (ORIC, Anraku, 1999; Kuiper & Hoijtink,

2010) with an application in for example Hothorn, Vaeth and Hothorn (2009) and for an introduction to the ORIC see Chapter 3.

One method of model selection which receives particular attention in this dissertation involves using Bayes factors. In this method each hypothesis of interest is provided with a 'degree of support' which tells us exactly how much support there is for each of the informative hypotheses under investigation. This process involves collecting evidence that is meant to provide support for or against a given hypothesis and as evidence accumulates, the degree of support for a hypothesis increases or decreases. Several technical articles have been published on how to evaluate informative hypotheses using Bayes Factors (Hoijtink, 1998, 2001; Hoijtink & Klugkist, 2007; Hoijtink, Klugkist & Boelen, 2008; Kato & Hoijtink, 2006; Klugkist, Laudy & Hoijtink, 2005; Laudy, Boom & Hoijtink, 2005; Laudy, Zoccolillo et al., 2005; Laudy & Hoijtink, 2007; Mulder, Hoijtink & Klugkist, 2009; Mulder, Klugkist et al., 2009). For an introduction to this method see Chapter 4 (see also the book of Hoijtink, Klugkist & Boelen, 2008) and for a comparison with other methods of evaluating hypotheses see Chapter 2 (see also, Kuiper & Hoijtink, 2010; Kuiper, Klugkist & Hoijtink, 2010).

1.5 Outline of Dissertation

The main topic of in the current dissertation is *informative hypotheses* and I take three different perspectives on this kind of hypothesis.

In the first part of this dissertation I take a philosophical approach to the question of why one should evaluate informative hypotheses in the first place. In Chapter 2 I provide arguments as to why the evaluation of informative hypotheses goes wrong when traditional null hypothesis testing is used. Bayesian model selection is then used to evaluate the hypotheses of interest and the results are compared with the results of traditional hypothesis testing. In Chapter 3 background knowledge, or prior information, is philosophically defined and its relationship with model selection procedures is

investigated. It is argued that the measure for complexity in model selection criteria needs to be refined. Two alternative model selection criteria are introduced and investigated in relation to the notion of simplicity for the evaluation of informative hypotheses.

In part II I adopt a statistical perspective and focus on the scope for extending the literature on the evaluation of informative hypotheses. In Chapter 4, I first provide an introduction for non-statisticians to the method of evaluating informative hypotheses using Bayesian model selection as described in, for example, Hoijtink, Klugkist and Boelen (2008), see also Mulder, Klugkist et al. (2009). In Chapter 5, a method based on the parametric bootstrap to evaluate informative hypotheses in structural equation models is described. The software program Mplus (Muthén & Muthén, 2007) is used for this method. It is also shown that the alpha level needs to be calibrated when evaluating informative hypotheses using the parametric bootstrap procedure. In Chapter 6, the prior predictive deviance criterion, or *prior* DIC, is derived as an alternative to the well known DIC, of Spiegelhalter et al. (2002). It is shown that there are situations in which the *prior* DIC can be used to evaluate informative hypotheses whereas the DIC of Spiegelhalter and colleagues fails to do so. However, it is shown that even the *prior* DIC fails in situations where inequality hypotheses are not supported by the data, and an alternative loss function is proposed that can be approximated by a new information criterium, called the Prior Information Criterion, or PIC.

In part III, two applications in the field of psychology are provided in which the research question is evaluated using informative hypotheses. In Chapter 7 I investigate the levels of self-concept (high or low) of delinquent young adults. Different expectations of low and high self-concept and antisocial behaviour are evaluated using Bayesian model selection. In Chapter 8 the progression and stability of adolescent identity formation is evaluated. On the basis of previous research, several assumptions about the increase and decrease of identity statuses over time are investigated.

PART I

Philosophy

2

Evaluating Expectations about Negative Emotional States of Aggressive Boys using Bayesian Model Selection

Van de Schoot, R., Hoijtink, H., Mulder, J., Van Aken, M.,
Orobio de Castro, B., Meeus, W. & Romeijn, J.-W.

In press for *Developmental Psychology*

Abstract

Researchers often have expectations about the research outcomes in regard to inequality constraints between, for example, group means. Consider the example of researchers who investigated the effects of inducing a negative emotional state in aggressive boys. It was expected that highly aggressive boys would, on average, score *higher* on aggressive responses towards other peers than moderately aggressive boys, who in turn score *higher* than non-aggressive boys. In most cases, null hypothesis testing is used to evaluate such hypotheses. We will show, however, that hypotheses formulated using inequality constraints between the group means are generally not evaluated properly. The wrong hypotheses are tested, i.e. the null hypothesis that group means are equal. In this paper, we propose an innovative solution to these above-mentioned issues using Bayesian model selection, which we illustrate using a case study.

Many psychology researchers rely on regression analysis, analysis of variance or repeated measures analysis to answer their research questions. The default approach in these procedures is to test the classical null hypothesis that 'nothing is going on': regression coefficients are zero, there are no group differences, etc. We argue that many researchers have some very strong prior beliefs about various components outcomes of their analyses and are not particularly interested in testing a traditional null hypothesis (see, e.g., Cohen, 1990, 1990). For example, a researcher might expect that highly aggressive boys would, on average, score *higher* on aggressive responses towards other peers than moderately aggressive boys, who in turn would score *higher* than non-aggressive boys. Note that we will refer to such explicit expectations as *informative* hypotheses.

This aforementioned explicit expectation is clearly not the same as the traditional null hypothesis: all scores for the boys are equal. Often researchers are not particulary interested in this null hypothesis. However, the average researcher specifies the traditional null hypothesis in a rather robotic way. Note that this is a critique of them not of the method, since classical null hypothesis testing is very useful for testing the null hypotheses if you are interested in it. Even so, there are already researchers who actually use prior beliefs directly in their data analysis, see for example Kammers et al. (2009); Meeus, Van de Schoot, Keijsers, Schwartz and Branje (2010); Meeus, Van de Schoot, Klimstra and Branje (2010); Van de Schoot and Wong (2010); Van de Schoot, Hoijtink and Doosje (2009); Van Well et al. (2009); Wong and Van de Schoot (2010).

In the current paper we show how subjective beliefs influence analyses in hidden ways and how they might be incorporated explicitly in data analysis. That is, we describe, by means of a case study, what can happen if a researcher has informative hypotheses, and uses traditional frequentist analysis or thoughtful frequentist analysis. Subsequently, we elaborate on an

alternative strategy: the evaluation of informative hypotheses by means of Bayesian model selection (Hoijtink, 1998, 2001; Hoijtink & Klugkist, 2007; Hoijtink, Klugkist & Boelen, 2008; Kato & Hoijtink, 2006; Klugkist et al., 2005; Klugkist, Laudy & Hoijtink, 2010; Kuiper & Hoijtink, 2010; Laudy, Zoccolillo et al., 2005; Laudy & Hoijtink, 2007; Mulder, Hoijtink & Klugkist, 2009; Mulder, Klugkist et al., 2009). Furthermore, we use one of our own studies (Orobio De Castro, Slot, Bosch, Koops & Veerman, 2003) in the area of experimental psychology to illustrate that our aim is not to disregard any specific study, but to discuss a problem very common to psychological research, a problem encountered in our own research as well.

2.1 Example: Emotional State in Aggressive Boys

Orobio De Castro et al. (2003) investigated the effects of inducing a negative emotional state in aggressive boys. It was questioned whether inducing negative emotions would make boys with aggressive behaviour problems attribute more aggressive responses and hostile intentions to their peers in comparison to the group of non-aggressive boys. The authors examined three levels of aggression: high, moderate, and no aggression.

The highly aggressive group consisted of boys referred to special education for aggressive behaviour problems. Informed consent was obtained from all participants and their parents. The moderately aggressive group consisted of boys in regular education with teacher-rated externalizing behavior problems scores on the Teacher's Report Form (TRF) in the borderline or clinical range. No SES information was made available from the original paper.

Mild negative emotions were induced by manipulating participants' performance in a computer game. Each participant completed two conditions: a neutral-emotion condition prior to playing a computer game (neutral) and a negative-emotion condition following emotional manipulation after unjustly losing the game (negative). Hostile intent attributions and aggressive

responses to other peers were assessed by presenting the boys with eight vignettes concerning ambiguous provocation by peers, for example:

> Imagine: You and a boy in your class are taking turns at a computer game. Now it's your turn, and you are doing great. You are reaching the highest level, but you only have one life left. You never came this far before, so you are trying very hard. The boy you are playing with watches the game over your shoulder. He sees how far you have come. Then he shouts "Watch out! You've got to be fast now!" and he pushes a button. But it was the wrong button, and now you have lost the game!

Two open-ended questions were asked directly after listening to each vignette: (1) why the provocateur in the vignette acted the way he did; (2) how the participants would respond if they were to actually experience the events portrayed in the vignette. Answers to the first question were coded as benign, accidental, ambiguous, or hostile. The reactions of the boys to the second question were coded as aggressive, coercive, solution attempt, or avoidant. By counting the number of vignettes in each condition with a hostile or an aggressive response to the questions, respective scores for hostile intentions and aggressive responses were calculated.

2.1.1 Expectations

The first expectation (A) was that negative emotion manipulation would invoke more hostile intentions and aggressive responses at all levels of aggression. This expectation was based on Dodge (1985), who hypothesized that a negative emotional state makes children more prone to attribute hostile intentions to other children they interact with. The constraints corresponding to the informative hypothesis $H_{A,host}$ in relation to hostile attribution are displayed in Table 2.1. It can be seen, for example, that the mean score for non-aggressive boys in the neutral condition is expected to be lower than the mean score for non-aggressive boys in the negative condition,

$M_{neu,non} < M_{neg,non}$. Note that the same constraints hold for aggressive responses ($H_{A,aggr}$).

A second expectation (B) was that emotion manipulation would influence aggressive boys more than less aggressive boys. Consequently, the tendency to attribute more hostile intentions to peers in ambiguous situations was expected to increase more in highly aggressive boys than in moderately aggressive and non-aggressive boys. As was argued by Orobio De Castro et al. (2003), this hypothesis seems plausible, given the fact that many children with aggressive behaviour problems have histories of abuse, neglect, and rejection (Coie & Dodge, 1998). As a result, these highly aggressive boys exhibit a greater tendency to attribute hostile intentions to peers in ambiguous situations than non-aggressive boys do (see also, Orobio de Castro, Veerman, Koops, Bosch & Monshouwer, 2002). The constraints corresponding to the informative hypothesis for hostile attribution ($H_{B,host}$) are displayed in the middle of Table 2.1. These constraints imply, for example, that the difference between the negative and neutral conditions is smaller for the non-aggressive group than for the moderately aggressive group, $[M_{neu,non} - M_{neg,non}] < [M_{neu,mod} - M_{neg,mod}]$. The same constraints also hold for aggressive responses ($H_{B,aggr}$).

A third expectation (C) was a combination of expectation A and B. The authors expected that negative emotion manipulation would invoke more hostile intentions and aggressive responses at all levels of aggression and at the same time that emotion manipulation would influence aggressive boys more than less aggressive boys. The difference between the neutral and the negative condition would become larger if boys are more aggressive. The hypotheses $H_{C,host}$ and $H_{C,aggr}$ combine the constraints presented in the upper part of Table 2.1 with the constraints presented in the middle of Table 2.1.

The research question investigated throughout the current paper is which of these three informative hypotheses, H_A, H_B, or H_C, is best supported by

Table 2.1: Constraints for Hypotheses A, B and C for Hostile Attribution

		Aggression level		
	Condition	No Aggression	Moderate	High
$H_{A,host}$	Neutral	$M_{neu,non}$	$M_{neu,mod}$	$M_{neu,high}$
		$\wedge$	$\wedge$	$\wedge$
	Negative	$M_{neg,non}$	$M_{neg,mod}$	$M_{neg,high}$
$H_{B,host}$		$M_{neg,non} - M_{neu,non} < M_{neg,mod} - M_{neu,mod} < M_{neg,high} - M_{neu,high}$		
$H_{C,host}$	Neutral	$M_{neu,non}$	$M_{neu,mod}$	$M_{neu,high}$
		$\wedge$	$\wedge$	$\wedge$
	Negative	$M_{neg,non}$	$M_{neg,mod}$	$M_{neg,high}$
			&	
		$M_{neg,non} - M_{neu,non} < M_{neg,mod} - M_{neu,mod} < M_{neg,high} - M_{neu,high}$		

Note. M indicates a mean score for an aggression level within a condition, e.g., $M_{neu,non}$ is the mean score for non-aggressive boys in the neutral condition.

the data. We try to answer this research question using traditional frequentist analysis, thoughtful frequentist analysis and Bayesian model selection.

2.2 Traditional Frequentist Analysis

The traditional frequentist approach, which is most often used in practice, is to analyse data like ours using tractional null hypothesis testing. In our example, aggressive responses and hostile intentions were used as dependent variables in two 3×2 analyses of variance (ANOVA) with level of aggression (high, moderate, and no aggression) as a between-participants factor and the condition (neutral/negative) as a within-participants factor. Three null hypotheses could be tested for both hostile intentions and aggressive responses:

- $H_{0,1}$: There is no difference between levels of aggression;
- $H_{0,2}$: There is no difference between the condition means;
- $H_{0,1\times2}$: There is no interaction between levels of aggression and the condition.

The results of these tests are presented in Table 2.2.

It can be seen in this table that for both aggressive responses and hostile intentions there appear to be significant differences between aggression level means and that there were no differences between condition means for both aggressive responses and hostile intentions. However the only significant result for the interaction effect is found for hostile attribution (i.e., level of aggression $\times$ condition). Many researchers would now perform a follow-up analysis, which we also do later on, but we first show what happens if the informative hypotheses H_A, H_B, and H_C are evaluated using the null hypotheses $H_{0,1}, H_{0,2}$ and $H_{0,1\times2}$.

Table 2.2: Results of the Two 3x2 Univariate Analyses of Variance

	Hostile		Aggressive	
	F	p	F	p
Aggressive level (df: 2, 55)	2.91	**.047**	8.82	**<.001**
Condition differences (df: 2, 55)	1.10	.29	0.82	.36
Interaction (df: 2, 54)	3.18	**.049**	1.46	.24

2.2.1 What goes wrong?

Although traditional null hypothesis testing has been the dominant research tool for the latter half of the past century it suffers from serious complications if used in the wrong way. That is when the null hypotheses $H_{0,1}, H_{0,2}$ and $H_{0,1\times2}$ are used to answer the question which informative hypothesis H_A, H_B, or H_C is best supported by the data. Let us elaborate on this.

The first and most vital problem that arises is that there is no straightforward relationship between the informative hypotheses under investigation and the null hypotheses that are actually being tested. (Orobio De Castro et al., 2003) were not interested in testing the hypotheses $H_{0,1}, H_{0,2}$ and $H_{0,1\times2}$ that were tested in the ANOVA. Although Wainer (1999) argues in "One Cheer for Null Hypothesis Significance Testing" that the null hypothesis can be useful in some cases, many researchers have no particular interest in the null hypothesis (see, e.g., Cohen, 1990). So why test the null hypothesis if one is not interested in it?

Furthermore, the informative hypotheses H_A, H_B, and H_C differ from the traditional alternative hypotheses: 'not $H_{0,1}$', 'not $H_{0,2}$' and 'not $H_{0,1\times2}$'. As can be seen in Table 2.2, some of the null hypotheses are rejected in favor of the alternative hypothesis (significant results in bold), but what does this tell us? For example, for hostile attribution there is a main level of aggression difference and an interaction between level of aggression and condition. Does this provide any evidence that one of the three informative hypotheses is more likely than the other? Clearly, the answer is 'no', because neither the null

hypotheses nor the alternative hypotheses resemble any of the informative hypotheses under investigation.

In conclusion, using traditional null hypothesis testing does not result in a direct answer to the research question at hand. This issue is usually solved by a visual inspection of the sample means. When inspecting Table 2.3, which shows the descriptive statistics (i.e. standardized means), it appears there is a violation of expectation A with regard to hostile attribution: the mean of the non-aggressive group is lower in the negative condition than in the neutral condition, rather than higher. Does this imply that expectation A is not supported by the data? Or is this a random deviation? The mean differences for hostile attribution between the neutral and negative condition for non-, moderate- and high-aggressive boys, presented in the lower part of Table 2.3, are in agreement with the constraints of hypothesis B. However, does this imply that H_B is preferred over H_A? What if there would have been a small deviance of the constraints imposed on the mean differences: $-.45, -.46, .45$? Or what if there would have been a larger deviance between the mean differences: $-.45, -.55, .45$? When would the difference be large enough to conclude that the informative hypothesis would be preferred?

2.2.2 Multiple Hypothesis Testing and Power

Alongside the complication of using null hypotheses testing the wrong way, the procedure of traditional null hypothesis testing suffers from a number of complications itself. Two important issues will be discussed here: an increase of type I errors due to multiple analyses and the loss of power that results from the adjustment often used to correct for these errors.

Multiple tests are typically needed to evaluate the informative hypotheses at hand and this can be problematic (e.g., Maxwell, 2004). In our example, six F-tests were performed. In general, multiple testing increases the family-wise error rate, which is the probability of incorrectly rejecting at least one null hypothesis of all hypotheses tested. For example, for two independent tests and an alpha level of .05 per test, the probability of correctly concluding

Table 2.3: Emotion Ratings by Aggression Level and Condition

		Hostile					Aggressive				
	Condition	No Aggression		Moderate		High	No Aggression		Moderate		High
$H_{A,host}$	Neutral	0.15		0.39		-0.27	0.52		1.02		1.12
		$\wedge$		$\wedge$		$\wedge$	$\wedge$		$\wedge$		$\wedge$
	Negative	-0.20		0.43		0.18	0.47		1.08		0.93
$H_{B,host}$		-0.45	$<$	0.04	$<$	0.45	-0.05	$<$	0.06	$<$	-0.19

Note. The numbers presented here are standardized means with higher values indicating more responses with a hostile or an aggressive intent to the questions.

that both null hypotheses are not rejected .95 × .95 = .90 and for six tests this is $.95^6$ = .74. In the latter case, the probability of incorrectly rejecting at least one null hypothesis is 1 -.74 = .26. Note that the six tests in Table 2.2 are not independent, but in this situation the overall alpha level is higher than .05 as well.

A solution to the problem of type I error inflation is to control the overall alpha level by using, for example, the often used Bonferroni correction. For this procedure, the overall alpha level is divided by the number of tests performed. The price for using such a correction is a severe reduction in power (see, Cohen, 1992). If the alpha level is corrected, this also requires a larger sample size to maintain sufficient power, which may not always be realistic. In our running example, ethical and clinical considerations urge us to limit, to an absolute minimum, the number of boys with severe behaviour problems who can be asked to participate in such a taxing manipulation. These sample size restrictions are evident in many studies in our field. Moreover, the Bonferroni correction is not unproblematic, the procedure is rather conservative, meaning that the smaller the alpha level, the lower the power. Improvements of the Bonferroni procedure have been developed, including the false discovery rate (Benjamini & Hochberg, 1995) or the Holm-Bonferroni method (Holm, 1979); for an overview see Hsu (1996). However, larger sample sizes are still needed in these cases, and for it remain difficult to determine how the overall alpha level should be corrected with all of these methods.

For example, when using any form of correction, should the overall alpha be corrected separately for each dependent variable? Or should the overall alpha be corrected by using the total number of tests? The answers to these questions are not clear. If we were to use the Bonferroni correction $\frac{\alpha}{3}$ for our example, then the significant results for hostile attribution disappear and the conclusion should be that there are no group main differences and that there is no interaction between group and condition. The null hypothesis cannot

be rejected, but what does this say about the informative hypotheses H_A, H_B, and H_C?

For aggressive responses, aggression level differences remain significant when using $\frac{\alpha}{3}$, implying that $(M_{non,neg} + M_{non,neu}) \neq (M_{mod,neg} + M_{mod,neu}) \neq (M_{neg,high} + M_{neu,high})$, where M is the mean score of a group within a condition. A significant result would indicate that $(0.52 + 0.47 = 0.99) \neq (1.02 + 1.08 = 1.10) \neq (1.12 + 0.93 = 2.05)$, but what can we learn from this with respect to H_A, H_B, and H_C? Clearly, the answer is 'not much'. Even if we pursue this significant result further using post-hoc comparisons, these comparisons do not provide information about the informative hypotheses A, B, or C.

2.3 Thoughtful Frequentist Analysis

What have we learned so far? Testing the null hypotheses $H_{0,1}, H_{0,2}$ and $H_{0,1\times2}$ followed by a visual inspection of the data is not the appropriate tool for evaluating the informative hypotheses H_A, H_B, and H_C. If a researcher has explicit expectations in the form of inequality constraints between means, he or she might be better off by using alternative procedures. In this section, we use thoughtful frequentist analysis, i.e. planned comparisons, to evaluate H_A, H_B, and H_C.

First, three one-sided t-tests could be performed to evaluate H_A:

- $M_{neu,non} < M_{neg,non}$ $(p_{hostile} = .22/2; p_{aggr} = .60/2)$;
- $M_{neu,mod} < M_{neg,mod}$ $(p_{hostile} = .88/2; p_{aggr} = .60/2)$;
- $M_{neu,high} < M_{neg,high}$ $(p_{hostile} = .02/2; p_{aggr} = .06/2)$.

To evaluate H_B planned comparisons could be used, a good primer is presented in Rosenthal, Rosnow and Rubin (2000), where several types of contrasts are introduced. In our example, H_B could be evaluated using the

linear contrast

$$-1\times[M_{neu,non}-M_{neg,non}]+0\times[M_{neu,mod}-M_{neg,mod}]+1\times[M_{neu,high}-M_{neg,high}].$$

A researcher who expects a monotonic relationship can create lambda weights that represent that hypothesis (see, Rosenthal et al., 2000). For now, we will use a linear increase and since this hypothesis is also directional, we expect an increase in the difference between conditions; the resulting p-value can be divided by two. The results are a significant increase for hostile ($p = .008/2$), but a non-significant result for aggression ($p = .32/2$). Both pieces of information (i.e. the results of the one-sided t-tests and planned comparison) need to be combined to evaluate H_C, but it is unclear how to do so.

Although the above procedure generates better results than the naive procedure presented in the previous section, there is still one major problem related to thoughtful frequentist analysis. Recall that we wanted to evaluate H_A, H_B, and H_C. Using planned comparisons, in whatever form, results again in testing the null hypothesis. These tests are clearly not the same as evaluating H_A, H_B, and H_C. A different approach is called for and this is what we do in the next section.

2.4 Bayesian Evaluation of Informative Hypotheses

As put forward by Walker, Gustafson and Frimer (2007) "the Bayesian approach offers innovative solutions to some challenging analytical problems that plague research in [...] psychology" (see also, M. D. Lee & Pope, 2006; M. D. Lee & Wagenmakers, 2005). The core idea of Bayesian inferences is that a priori beliefs are updated with observed evidence and both are combined in a so-called posterior distribution. In the social sciences, however, only few applications of Bayesian methods can be found; one good example is presented in Walker, Gustafson and Hennig (2001). The authors used standard statistical techniques as well as a Bayesian approach to investigate

consolidation and transition models in the domain of moral reasoning. The posterior distribution of reasoning across stages of moral reasoning was used to predict subsequent development. Another example is the study of Schultz, Bonawitz and Griffiths (2007) about causal learning processes in preschoolers. Bayesian inference was used in this article to provide a rationale for updating children's beliefs in light of new evidence and was used to explore how children solve problems.

2.4.1 Bayes in the Social Sciences

An important contribution Bayesian methods can offer to the social sciences is the evaluation of informative hypotheses formulated with inequality constraints using Bayesian model selection. Many technical papers have been published about this method in statistical journals (Hoijtink, 1998, 2001; Hoijtink & Klugkist, 2007; Hoijtink, Klugkist & Boelen, 2008; Kato & Hoijtink, 2006; Klugkist et al., 2005, 2010; Kuiper & Hoijtink, 2010; Laudy, Zoccolillo et al., 2005; Laudy & Hoijtink, 2007; Mulder, Hoijtink & Klugkist, 2009; Mulder, Klugkist et al., 2009). Applied psychology/social science articles that use this method to evaluate hypotheses have been published as well. For example, in a study by Van Well et al. (2009), the authors investigated whether a possible match between sex or gender role identification on the one hand and gender relevance of a stressor on the other hand would increase physiological and subjective stress responses. A first expectation represented a sex match effect; participants were expected to be most reactive in the condition that matches their sex. In a similar way, gender match, sex mismatch, and gender mismatch effects were evaluated using Bayesian model selection software. Another example is the study by Meeus, Van de Schoot, Keijsers et al. (2010). In this study, Bayesian model selection was used to evaluate the plausibility of certain patterns of increases and decreases in identity status membership on the progression and stability of adolescent identity formation. Moreover, expected differences in prevalence of identity statuses between early-to-middle and middle-to-late

adolescents and males and females were evaluated. In sum, Bayesian model selection as described in, for example Hoijtink, Klugkist and Boelen (2008), is gaining attention and is a flexible tool that can deal with several types of informative hypothesis.

The major advantage of evaluating a set of informative hypothesis using Bayesian model selection is that prior information can be incorporated into an analysis. As was argued by Howard, Maxwell and Fleming (2000), replication is an important and indispensible tool in the social sciences. Evaluating informative hypotheses fits within this framework because results from different research papers can be translated into different informative hypotheses. The method of Bayesian model selection can provide each informative hypothesis with the degree of support provided by the data. As a result, the plausibility of previous findings can be evaluated in relation to new data, which makes the method described in this paper an interesting tool for replication of research results. Another advantage of evaluating informative hypotheses is that more power is generated with the same sample size. An increase in power is achieved because using the data to directly evaluate H_A, H_B and H_C by evaluating H_A versus H_B versus H_C is more straightforward than testing several null hypotheses that are not related to the hypotheses of interest.

2.4.2 Software

In this paper we analysed the informative hypotheses of our example using the software presented in (Mulder, Klugkist et al., 2009). The method described in Mulder et al. can deal with many complex types of (in)equality constraints in multivariate linear models, e.g. MANCOVA, regression analysis, repeated measure analyses with time varying and time in-varying covariates. A typical example of an informative hypothesis in the context of regression analysis can be found in Deković, Wissink and Meijer (2004). It was hypothesized that adolescent disclosure is the strongest predictor of

antisocial behaviour, followed by either a negative or positive relation with the parent.

Software is also available for evaluating informative hypotheses in AN(C)OVA models (Klugkist et al., 2005; Kuiper & Hoijtink, 2010), latent class analysis (Hoijtink, 1998, 2001; Laudy, Zoccolillo et al., 2005) as well as order restricted contingency tables (Laudy & Hoijtink, 2007; Klugkist et al., 2010). Readers interested in this software can visit *www.fss.uu.nl/ms/informativehypothesis*. Users of the software need only provide the data and the set of constraints; the Bayes factors are computed automatically by the software. A first attempt in analysing data can best be made by using the software program 'confirmatory ANOVA' (Kuiper et al., 2010). We refer to the book of Hoijtink, Klugkist and Boelen (2008) as a first step for interested readers.

2.5 Introduction to Bayesian Model Selection

In this section we provide a brief introduction to the evaluation of informative hypotheses formulated with inequality constraints using Bayesian model selection. The main ideas are introduced below, and we refer interested readers to S. Lynch (2007) for a general introduction to Bayesian analysis and we refer to Gelman, Meng and Stern (1996) for a technical introduction to Bayesian analysis. For incorporating inequality constraints in the context of Bayesian model selection, we refer interested readers to Hoijtink, Klugkist and Boelen (2008).

2.5.1 Bayes Factor

As was shown by Klugkist et al. (2005) informative hypotheses can be compared using the ratio of two marginal likelihood values, which is a measure for the degree of support for each hypothesis provided by the data (see, e.g., Hoijtink, Klugkist & Boelen, 2008). This ratio results in the Bayes factor, see Kass and Raftery (1995) for a statistical discussion of the Bayes

factor. The outcome represents the amount of evidence in favour of one hypothesis compared with another hypothesis.

Returning to our example of Orobio de Castro et al. (2003), the informative hypotheses H_A, H_B and H_C can be evaluated using Bayesian model selection. To do so, we first compare these informative hypotheses to a so-called unconstrained hypothesis, denoted by H_{unc}. A hypothesis is unconstrained if no constraints are imposed on the means. The comparison with H_{unc} is made because it is possible that all informative hypotheses under investigation do not provide an adequate description of the population from which the data were sampled. In that case, the unconstrained hypothesis will be favored by Bayesian model selection. Hence, Bayesian model selection protects a researcher against incorrectly choosing such a 'bad' hypothesis.

As was shown by Klugkist et al. (2005), the Bayes factor (BF) of H_A versus H_{unc} can be written as

$$BF_{A,unc} = \frac{f_i}{c_i} \ , \tag{2.1}$$

where f_i can be interpreted as a measure for model fit and c_i as a measure for model complexity of H_A. The Bayes factor of H_A versus H_B can then be written as:

$$BF_{A,B} = \frac{BF_{A,unc}}{BF_{B,unc}} \ . \tag{2.2}$$

The Bayes factor in Equation (2.1) combines model fit and complexity and represent the amount of evidence, or support from the data, in favor of one hypothesis (say, H_A) compared to another hypothesis (say, H_B).

The results may be interpreted as follows: $\text{BF}_{A,B} = 1$ states that the two hypotheses are equally supported by the data; $\text{BF}_{A,B} = 10$ states that the support for H_A is 10 times stronger than the support for H_B; $\text{BF}_{A,B} = 0.25$ states that the support for H_B is 4 times stronger than the support for H_A. Note that there is no cut-off value provided; we return to this issue in the next section, but let us first re-analyse our example, and after that we elaborate on f_i and c_i.

2.5.2 Example Reconsidered

To re-analyse the data of Orobio De Castro et al. (2003) we computed the Bayes factors using two analysis of variance models, one for hostile attribution and one for aggressive response. The results are presented in Table 2.4.

For hostile attribution the $\text{BF}_{A,unc}$ of H_A compared to H_{unc} is 0.24. This implies that H_A is not better than the unconstrained hypothesis and is consequently not supported by the data (accounting for model fit and complexity). The $\text{BF}_{B,unc}$ of H_B compared to H_{unc} is 4, indicating that support from the data is 4 times stronger for H_B than for H_{unc}. The $\text{BF}_{C,unc}$ indicates that support from the data is 1.5 times stronger for H_C than for H_{unc}. In sum, only H_B and H_C are supported by the data.

Using these results, one can compute a Bayes factor between two informative hypotheses. The resulting Bayes factor is equal to the ratio of the Bayes factor for each informative hypothesis with the unconstrained hypothesis by using Equation (2.2). The $\text{BF}_{B,C}$ for hostile attribution between H_B and H_C is $\frac{4}{1.5} = 2.66$, which means that the support for H_B is 2.66 times stronger than the support for H_C. A comparison with H_A is not necessary since the constraints of this hypothesis are not supported by the data anyway. In conclusion, there is no support for the expectation that an increase in hostile intentions takes place for all three groups following emotion manipulation, but there is support for the expectation that the increase in

Table 2.4: Estimates for Bayes Factors against H_{unc}, model fit and model complexity for H_A, H_B, and H_C

	Hostile			Aggressive		
	f_i	c_i	BF	f_i	c_i	BF
H_A	.06	.25	0.24	.23	.25	0.92
H_B	.64	.16	4.00	.02	.16	0.12
H_C	.03	.02	1.50	1 e^{-6}	.02	0.00
H_{unc}	1	1	1	1	1	1

hostile intentions becomes larger when the groups consist of more aggressive boys.

Similar computations can be performed for the aggressive response, see Table 2.4. However, none of the hypotheses under investigation is better than an unconstrained hypothesis. Consequently, none of the hypotheses give an adequate description of the population from which the data were sampled. As a result, there is no increase in aggressive response following emotion manipulation and there is no support for the expectation that the increase in aggressive response becomes larger when the groups consist of more aggressive boys. A combination of both hypotheses, H_C, receives even less support.

2.5.3 Complexity and Fit

For a better understanding of the Bayes factor and its relation with model fit and model complexity we elaborate on f_i and c_i. As was shown before, Bayesian model selection provides the degree of support for each hypothesis under consideration and combines model fit and model complexity. It has a close link with classical model selection criterion such as AIC (Akaike, 1981) and BIC (Schwarz, 1978) that also combine fit and complexity to determine the support for a particular model. However, in contrast to Bayesian model selection these classical criteria are as of yet unable to deal with hypotheses specified using inequality constraints (Van de Schoot, Romeijn & Hoijtink, 2010). In the specific application of Bayesian model selection used in this paper, the Bayes factors selection criteria also combine model fit and complexity, but is able to account for inequality constraints. As will now be illustrated, complexity and fit are (although implicitly) also important parts of Bayesian model selection.

Model Complexity

The first component of the Bayes factor is model complexity, c_i, which can be computed before observing any data. The Bayes factor incorporates the

complexity of a hypothesis by determining the number of restrictions imposed on the means. Note that model complexity is independent of the data because it is the proportion of the prior distribution in agreement with the constraints. Let us elaborate on this using our running example.

According to Sober (2002), the simplicity of a hypothesis can be seen as an indicator of the amount of information the hypothesis provides. Classical model selection tools favor models that allow for fewer possibilities, and call such models simpler. Relating this to, for example, the AIC and BIC, where complexity is measured as the number of parameters in a model, the more dimensions are 'shaved' away the simpler the model becomes. We maintain that there is also such a natural relation between introducing inequality constraints and ruling out possibilities, that is, when specifying such inequality constraints, a researcher also 'shaves' away parameter space volume. In sum, a simple hypothesis contains more restrictions and contains more information and as such, is more specific and should be favoured by the model selection procedure.

Returning to our example, the most complex hypothesis is always H_{unc}, in the sense that all combinations of means are allowed and no constraints are imposed. Therefore, c_i for H_{unc} is equal to 1, see Table 2.4. Let us consider the hypotheses specified for hostile attribution. There are two constraints specified for H_B (see Table 2.1). Consequently, not all combinations of means are possible. H_B is therefore considered simpler than H_{unc}. Three constraints are specified for H_A and this hypothesis is even simpler than H_B. The simplest hypothesis is H_C because here the most information is added: the constraints of H_A in addition to the constraints of H_B. With respect to complexity, the hypotheses can be ordered from simplest to most complex: H_C, H_B, H_A.

In Table 2.4 estimates for model complexity are displayed and our expected ordering for both hostile attribution and aggression is confirmed. That is, the proportion of the prior distribution in agreement with the constraints for H_A is .25 and for H_C only .02, making this latter hypothesis

less complex because more information is specified in term of the number of inequality constraints.

Model Fit

After observing some data, the second component of the Bayes factor is model fit, f_i. Loosely formulated, it quantifies the amount of agreement of the sample means with the restrictions imposed by a hypothesis.

Consider the sample means in Table 2.3. The observed sample means fit perfectly with an unconstrained hypothesis because no constraints are imposed on the means. Consequently, H_{unc} always has the best model fit compared to any other informative hypothesis and $f_i = 1$, see Table 2.4. With respect to the informative hypothesis on hostile attribution, it appears that one constraint is violated for H_A: the sample mean of the non-aggressive group for the neutral condition is higher for the negative condition rather than lower. As a result, the model fit of H_A is worse than the model fit for H_{unc}. For H_B there appeared to be no violations of the constraints. Since H_C is a combination of the constraints of H_A and H_B, there is one violation of the constraints imposed by this hypothesis. In sum, with regard to model fit, H_B performs better than H_A and H_C, respectively. In Table 2.4 estimates for model fit for the three informative hypotheses on hostile attribution are displayed and as can be seen the expected ordering is confirmed. With regard to three informative hypotheses on aggression the fit is rather low for all three hypotheses. After computing f_i and c_i, the Bayes factor shown in Equation (2.1) can be computed, for example for hostile attribution

$$BF_{B,unc} = \frac{f_i}{c_i} = \frac{.64}{.16} = 4 \ . \tag{2.3}$$

As was correctly noticed by one of the reviewers, it can be illustrative to provide more information than just the Bayes factors in terms of model fit and model complexity. Information about the posterior distributions of the means and their credibility intervals can be found in Figure 2.1. The interpretation of a Bayesian 95% credibility interval is that, for example,

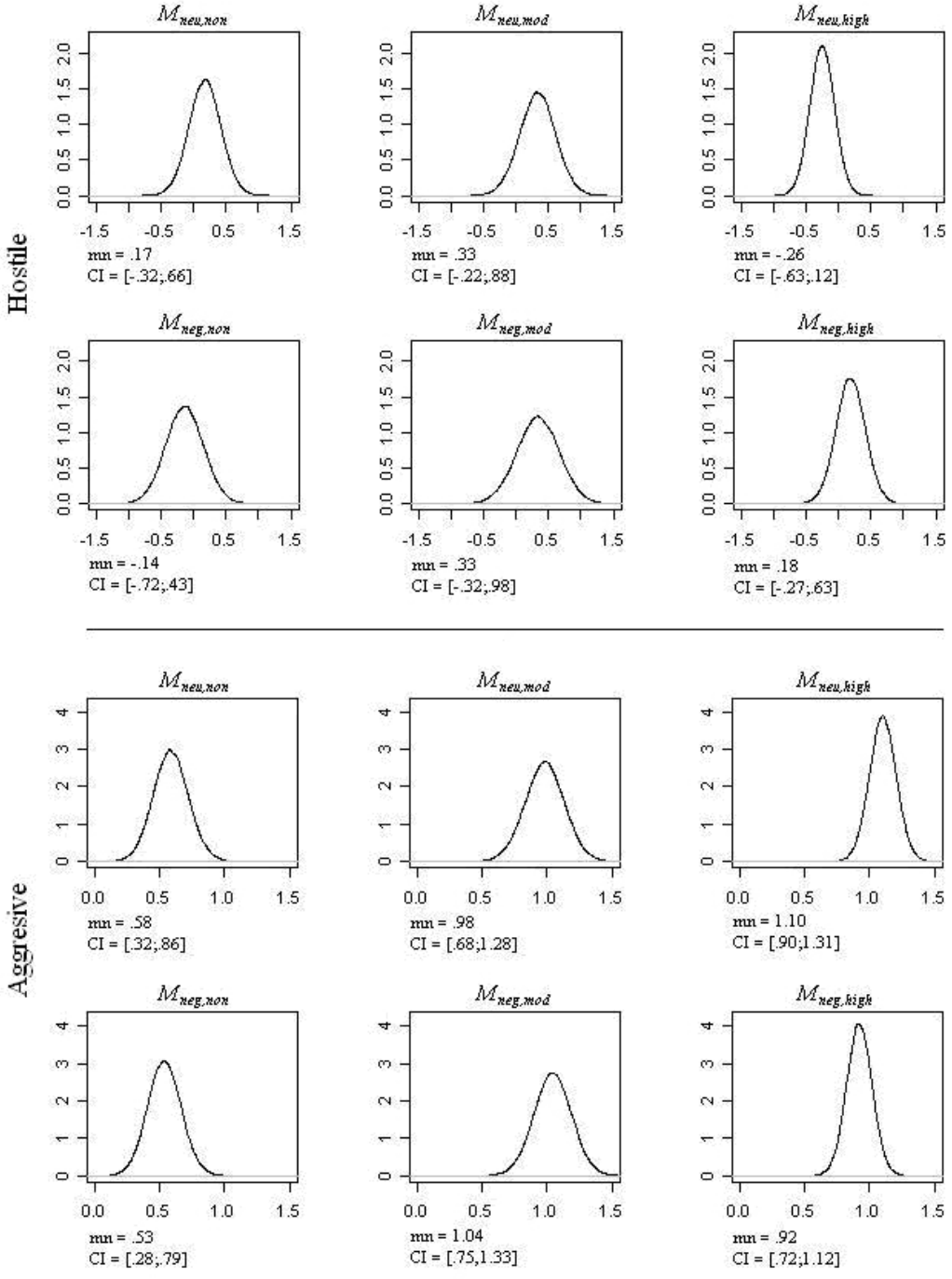


Figure 2.1: Posterior distributions for all groups on the dependent variables hostile attribution and aggressive responses. Note that 'mn' denotes posterior mean and 'C.I.' denotes the Bayesian credibility interval.

the posterior probability that $M_{neu,non}$ for hostile lies in the interval from -.32 to .66 is 0.95 (see, e.g., S. Lynch, 2007). These intervals are often used in practice to decide whether means differ from zero or from other means. It can for example be seen that the posterior mean $M_{neu,non}$ for aggression is .58 and there is a .95 probability that it is between .32 and .86. This credibility interval does not include zero and consequently the null hypothesis $M_{neu,non} = 0$ is rejected. Furthermore, it can be seen that the credibility intervals for $M_{neu,non}$ and $M_{neg,non}$ for aggression show an overlap, so the constraint $M_{neu,non} < M_{neg,non}$ is not supported by the data. Suppose we would observe the same results (i.e. posterior means) but with a larger sample size, the posterior distributions would be more peaked. Hence, the overlap of the credibility intervals for $M_{neu,non}$ and $M_{neg,non}$ will disappear. Consequently, the fit of the model would increase. In the next section we elaborate on the relation between model fit on the one hand and effect size, and sample size on the other hand.

2.6 Bayes Factors versus p-Values

Recall that a Bayes factor provides a direct quantification of support as evidenced in the data for two competing hypotheses. Most researchers would agree that 100 times more support seems to be quite a lot and, for example, 1.04 times more support is not that much. However, clear guidelines are not provided in the literature and we do not provide these either. We refrain from doing so because we want to avoid creating arbitrary decision rules. Remember the famous quote about p-values: "[...] surely, God loves the .06 nearly as much as the .05" (Rosnow & Rosenthal, 1989, p. 1277).

To gain insight into the interpretation of Bayes factors in comparison to p-values, consider the following imaginary example. Suppose there are six means, denoted by $M_1, \ldots, M_6$, and that the informative hypothesis of interest is $H_D : M_1 < M_2 < M_3 < M_4 < M_5 < M_6$. We created data in such

a way that the sample means and variance correspond exactly to population values as are shown in the footnote of Table 2.5. Now let us compare:

1. The F-test for traditional frequentist analysis;
2. Planned comparisons for thoughtful frequentist analysis assuming a linear increase ($-2.5 \times M_1 + -1.5 \times M_2 + -.5 \times M_3 + .5 \times M_4 + 1.5 \times M_5 + 2.5 \times M_6$)
3. Bayesian evaluation of informative hypotheses using BFs as described above for $\text{BF}_{D,unc}$.

We ran these analyses for different populations with a small and medium effect, a small and large sample size, and with zero, one and two violations of the ordering; see Table 2.5. Comparison of the resulting p-values with the Bayes factors will provide insight in the interpretation.

As can be seen in Table 2.5, for some of the data the classical F-test is not significant, although there are differences between the means within the population (i.e. population 2, 6, 8). This result indicates a power problem that is not shared by the planned comparison and the Bayes factor. The results for the planned comparison indicate that for all populations, apart from the null population 1, there is a significant linear increase in the six means even with 1 or 2 violations of the constraints.

Inspection of the Bayes factors indicates that its value is dependent on, firstly, effect size. Compare for example population 2 with population 4, with Bayes factors 29 versus 91, respectively. Second, sample size. Compare for example population 2 with population 3, with Bayes factors 29 versus 470, respectively. Finally, the number of violations. Compare for example populations 2, 6 and 8 with 0, 1 and 2 violations and with Bayes factors of 29, 20 and 4, respectively. In the latter population there is still support for the informative hypothesis, but 4 is clearly not a great deal of support in comparison to the other, much larger, results.

Recall the posterior means presented in Figure 2.1. Suppose the sample size is increased, then the posterior distributions will become more peaked

Table 2.5: Results of the Comparison between a Classical F-Test, Planned Comparison, and Bayes Factors

Population	Small/medium effect[1]	Small/large sample size	0/1/2 violations[2]	Classical F-test; $F_{1,5} =$	Linear increase $F_{1,5} =$	Bayes factor $BF_{/,unc} =$
1	No effect[3]	100	0	0; p = 1	0; p = .50	1.05
2	Small	10	0	1.40; p = .15	0.84; p = .005	29.51
3	Small	100	0	14.0; p < .001	0.84; p < .001	470.29
4	Medium	10	0	3.58; p = .007	1.34 p < .001	91.55
5	Medium	100	0	35.84; p < .001	1.34; p < .001	694.27
6	Small	10	1	1.40; p = .24	0.79; p = .005	20.52
7	Small	100	1	14.0; p < .001	0.79; p < .001	48.22
8	Small	10	2	1.40; p = .23	0.74; p = .005	12.74
9	Small	100	2	14.0; p < .001	0.74; p < .001	4.75

[1] Effect size according to definition of Cohen (1992) with population means for the small effect: $M_1 = -.50,\ M_2 = -.30,\ M_3 = -.10,\ M_4 = .10,\ M_5 = .30,\ M_6 = .50$ (SD = 1) and for the medium effect: $M_1 = -.80,\ M_2 = -.48,\ M_3 = -.16,\ M_4 = .16,\ M_5 = .48,\ M_6 = .80$ (SD = 1)

[2] With 1 violation two means are reversed ($M_1 = -.50,\ M_2 = \mathbf{-.10},\ M_3 = \mathbf{-.30},\ M_4 = .10,\ M_5 = .30,\ M_6 = .50$) and with 2 violations four means are reversed ($M_1 = -.50,\ M_2 = \mathbf{-.10},\ M_3 = \mathbf{-.30},\ M_4 = .10,\ M_5 = \mathbf{.50},\ M_6 = \mathbf{.30}$).

[3] All means are zero in the population.

and the overlap between distributions will disappear. Stated otherwise, the Bayes factor will increase with increasing sample size because an increase in model fit. The same holds for an increase in effect size, that is, the further the posterior means are away from each other, the less overlap between distributions.

What can be learned from this exercise? First, Bayes factors are sensitive for effect size, the number of violations, and sample size. When comparing informative hypotheses the complexity of each hypothesis under investigation is independent of these three concepts, as was shown before. It is the fit of the model that is influenced by the three concepts, namely the fit of a model will increase with higher effect sizes, a decrease in the number of violations and an increase in sample size. Second, in this section we specified only one single informative hypothesis which we evaluated with Bayes factors and p-values. It is interesting to note that the Bayes factor tells us exactly how much better a certain informative hypothesis is against another hypothesis. In comparison, a p-value tells us the probability, given that the null hypothesis is true, of observing the same data or more extreme data than those actually observed. The p-value, however, is often misinterpreted as the probability that the null hypothesis is true (see, e.g., Balluerka, Gómez & Hidalgo, 2005). Recall that if we would specify more informative hypotheses it is difficult, or even impossible, to use p-values as was shown before.

2.7 Conclusion

In the current paper we showed how subjective beliefs influence analyses in hidden ways and how they might be incorporated explicitly. Researchers in developmental psychology often have explicit expectations about their research questions, or as M. D. Lee and Pope (2006) say "In the real-world much is usually already known about a problem before data are collected or observed." As we showed in the current paper, these expectations can be translated into informative hypotheses. However, as we demonstrated

with a case study, the average researcher wants to evaluate such informative hypotheses, but tests a set of null hypotheses. We argued that researchers should not use traditional frequentist analysis, not even thoughtful frequentist analysis, if they are not interested in the conclusion that the observed data either are or are not in agreement with the null hypothesis. Rather researchers should directly evaluate all the informative hypotheses under investigation without relying on testing the null hypothesis. This can be done using Bayesian model selection. This way researchers can use all the knowledge available from previous investigations and can learn more from their data than traditional null hypotheses testing. All criticisms of null hypothesis testing aside, the best argument for evaluating informative hypotheses is probably that, like Orobio De Castro et al. (2003), many researchers want to evaluate a set of hypotheses formulated with inequality constraints, but have been unable to do so because the statistical tools were not yet available. As this paper has illustrated, these tools are available to any researcher.

3

Background Knowledge in Model Selection Procedures

Van de Schoot, R., Romeijn, J.-W. & Hoijtink, H.

Manuscript under review

Abstract

This paper concerns the interplay between simplicity and model fit in statistical model selection, particularly the use of background knowledge as expressed in order constraints between the parameters of interest. Extant model selection procedures do not manage to accommodate such order constraints. We will present two revised model selection criteria that are being proposed in statistical literature and then argue that these revised criteria give rise to a refinement of the notion of model complexity. Rather than taking the number of parameters as an expression of complexity, we maintain that complexity is captured by model size.

While analyzing their data, scientists want to select a simple statistical model *and* they want to learn as much as possible from their data. In order to do so, they often use background knowledge, based on previous research, literature reviews, and the current academic debate. The background knowledge is typically translated in a set of candidate models, that each generate specific statistical hypotheses (see, e.g., Henderson, Goodman, Tenenbaum & Woodward, 2010). Finally, model selection criteria are used to choose between these models.

This paper concerns the question of how a specific kind of background knowledge can be introduced into model selection procedures, namely inequality constraints between the parameters of interest. Many scientists have explicit expectations about the ordering between statistical parameters. Phrases like "The mean outcome in both experimental groups is expected to be larger than in the control group" and "Women score higher than men in each condition" can be found in many papers. Evaluating such expectations, formulated with inequality constraints between the parameters of interest, is at the very forefront of research in statistics (Anraku, 1999; Hoijtink, Klugkist & Boelen, 2008; Klugkist et al., 2005; Kuiper & Hoijtink, 2010; Mulder, Klugkist et al., 2009; Silvapulle & Sen, 2004; Van de Schoot, Hoijtink & Deković, 2010; Van de Schoot, Hoijtink, Mulder et al., 2010).

There is a small variety of model selection procedures commonly used in practical applications, most notably Akaike's Information Criterium (AIC; Akaike, 1973), and the Bayesian Information Criterium (BIC; Schwarz, 1978). All model selection tools boil down to a trade-off between model fit and model complexity, or conversely, model simplicity. The philosophical story about this trade-off has been told in the literature (Forster & Sober, 1994; Forster, 2002; Kieseppä, 2001; Sober, 2006, 2002). However, a philosophical understanding of the interplay between model fit and model complexity in

relation to the use of inequality constraints is lacking. That project will be undertaken in the current paper.

The introduction of background knowledge in model selection procedures provides us with a new challenge. The problem is that we can not tell the same story about the trade-off between fit and complexity as has been told in the philosophical literature. First, Forster and Sober argue that model dimensionality is a expression of model complexity (Forster & Sober, 1994; Forster, 2002; Sober, 2006, 2002). We argue, however, that this notion of complexity needs refinement when inequality constraints are specified between the parameters of interest. In this paper we show that using model size (i.e. admissible parameter space) rather then model dimensionality (i.e. the number of parameters in a model) should be used as a measure of complexity.

Another important philosophical perspective on model selection is proposed by Kieseppä (2001) who described it as a three step procedure: (1) choose a family of models; (2) choose a statistical model which belongs to a specific family of models; (3) choose an element of the best statistical model. In this paper we propose a intermediate step in this procedure: Step (2'), where a set of statistical models is specified using inequality constraints between statistical parameters. In the end the model with the best trade-off between model fit and model *size* is selected.

The plan of this paper is as follows. After introducing two standard model selection techniques (AIC, BIC), we show in Section 3.2 why these selection procedures are equipped to select the best model in Step (2) of Kieseppä (2001), but are not equipped to select the best model in Step (2'). In Section 3.3 we elaborate on revisions of the AIC and BIC (i.e. the ORIC and the *prior adapted* BIC, respectively) that are able to select the best model in Step (2'). Finally, in Section 3.4 we show how the notion of simplicity in model selection procedures can be refined, but first we define in Section 3.1 the aforementioned specific kind of background knowledge.

3.1 Background Knowledge

There are many kinds of background knowledge that scientist might want take into account when analyzing their data. In the current paper, we are going to focus on one particular type of background knowledge: expectations about the ordering of statistical parameters that researchers might consider.

3.1.1 Practical Example of Background Knowledge

We will illustrate the type of background knowledge at issue with a study in the field of developmental psychology taken from Jongmans, Smits-Engelsman and Schoenmaker (2003). The authors investigated differences in the severity of perceptual motor problems encountered by children with developmental coordination disorder (DCD) for two different groups, with and without concomitant learning disabilities (LD). DCD entails the partial loss of the ability to coordinate and perform certain purposeful movements and gestures. Children with LD have trouble performing specific types of skills or completing tasks if left to figure things out by themselves. The researchers constructed four groups of children and computed the mean for each group on perceptual motor problems: μ_{00} pertains to the subgroup of having neither DCD nor LD; μ_{10} pertains to the subgroup having LD only; μ_{01} pertains to the subgroup having DCD only; and finally μ_{11} pertains to the subgroup of having both DCD and LD.

Based on previous research the authors have two different expectations about the ordering of these four means and this is exactly where background knowledge is introduced. A first expectation is that children having neither DCD nor LD have a lower score on perceptual motor dysfunction than the other three groups. The statistical hypothesis for this expectation is $\mu_{00} < \{\mu_{10}\ ,\ \mu_{01}\ ,\ \mu_{11}\}$. Another batch of research articles leads to an additional expectation, namely that children having only one disorder are better off than children with DCD who also suffer from LD. Suffering from both disorders would decrease the severity of perceptual motor problems

more then when suffering from only one disorder. The statistical hypothesis for this expectation is $\mu_{00} < \{\mu_{10}\ ,\ \mu_{01}\} < \mu_{11}$.

As can be seen from this example, the background knowledge here concerns rivalling expectations about the ordering of means. These are expressed in two different statistical models specified with inequality constraints among the parameters. Model selection tools can now be used to choose between the statistical models.

3.1.2 Restricting the Notion of Background Knowledge

In this section we will take a closer look at the notion of background knowledge used in this paper. First of all, we limit our discussion about background knowledge exclusively to information about the statistical parameters. It can consist of two parts: (i) certain parameters may be ruled out at the outset or equated with another parameter, thereby decreasing the dimension of the parameter space, or (ii) a specific part of the parameter space may be ruled out because of the use of inequality constraints. In the latter case the number of dimensions of the parameter space will typically be the same, but the range of possible statistical hypotheses is made smaller. In other words, the size of the parameter space is reduced.

To bring out the salient parts of our arguments, we present a simple example based on Jongmans et al. (2003) which we will use throughout the remainder of this paper. Say that we have some data E and that we think these data are sampled from a distribution characterized by two parameters of the Jongmans et al. (2003) example, μ_{01} and μ_{11} (i.e. the subgroup having DCD only, and the subgroup of having both DCD and LD). For reasons of simplicity we will call them μ_1 and μ_2, respectively. Furthermore assume that $\mu_1 \in [0, 1]$ and $\mu_2 \in [0, 1]$. In this case each pair of values for μ_1 and μ_2 represents a specific statistical hypothesis concerning the data. By contrast, a statistical model consists of a set of such hypotheses. Such models sometimes are referred to as composite hypotheses. We can now specify a set of T different statistical models denoted by $\mathcal{M}_t (t = 1, \ldots, T)$.

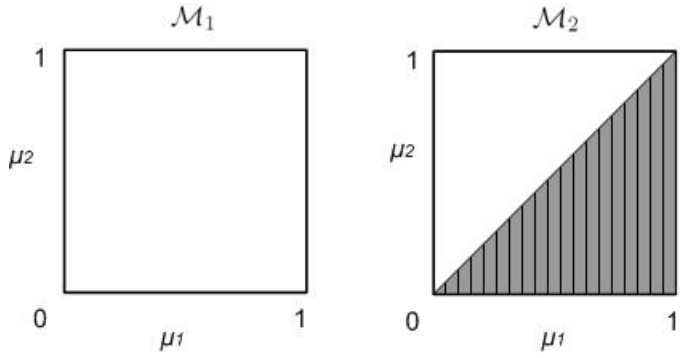


Figure 3.1: The squares in this figure represent the parameter space for models $\mathcal{M}_1$ and $\mathcal{M}_2$. Say that the square consists of points and that each point represents a combination of two values, one for μ_1 and one for μ_2. For $\mathcal{M}_1$ all points, or combinations of μ_1 and μ_2, are a-priori allowed and as such the admissible parameter space is equal to the total parameter space. For $\mathcal{M}_2$ some part of the parameter space is a-priori not allowed, which is graphically displayed by the gray area.

One possible statistical model for the data allows for all possible values of both parameters; that is, $\langle \mu_1, \mu_2 \rangle \in [0, 1]^2$. Call this model $\mathcal{M}_1$. It consists of the entire range of hypotheses $H_{\mu_1 \mu_2}$ and is referred to as the unconstrained model. In the left panel of Figure 3.1 a graphical representation is given for $\mathcal{M}_1$. The square represents the parameter space and that each point in that space represents a combination of two values, one for μ_1 and one for μ_2. For $\mathcal{M}_1$ all points, or combinations of μ_1 and μ_2, are a-priori admissible. As such, the admissible parameter space is equal to the total parameter space.

Now, consider another possible model, $\mathcal{M}_2$, which imposes the restriction that $\mu_2 > \mu_1$. The right panel of Figure 3.1 shows the parameter space of $\mathcal{M}_2$. In contrast with $\mathcal{M}_1$, not all combinations of μ_1 and μ_2 are a-priori admissible; this is represented by the shaded area in Figure 3.1. The white area is what we will call the 'admissible' parameter space of the range of hypotheses for $H_{\mu_2 > \mu_1}$. Stated otherwise, the subgroup of children having DCD only is expected to have higher scores on the severity of perceptual motor problems compared to the subgroup of children having both DCD and LD.

Models with inequality constraints are nested in the unconstrained model: $\mathcal{M}_2 \subseteq \mathcal{M}_1$. Note that all models are a particular set of statistical hypotheses, each of which fixes a fully specified distribution for the data. Their difference is that the last model restricts the possible values for the parameters μ_1 and μ_2. Before observing E, the size of the admissible parameter space differs between the models under investigation: the volume of admissible parameter space is smaller for $\mathcal{M}_2$. After observing E, it could be that E is either located in that part of the parameter space where both $\mathcal{M}_1$ and $\mathcal{M}_2$ are admissible, or that it is located where $\mathcal{M}_1$ is admissible but $\mathcal{M}_2$ is not. Model selection procedures are developed to choose among statistical models after observing E.

3.1.3 Background Knowledge and Model Selection Procedures

Kieseppä (2001) describes a three step procedure for model selection procedure:

1. Choose between a number of sets of models.
2. Model selection: choose among the chosen set of models, which are denoted by $\mathcal{M}_s$ $(s = 1, \ldots, S)$.
3. Estimation: choose an element of the selected model.

He argues that when a reasonable choice has been made in the first step, statistical model selection techniques can help to make a reasonable choice in the second step. In the current paper we will not focus on Step (3).

Say that in Step (1) we have chosen a set of models in which four statistical parameters feature: $\mu_1, \ldots, \mu_4$. This set consists of models differing in the exact parameters included, so that these models differ in dimensionality. In the second step we now have to choose among those models. For example, we may have a set of three models $\{\mathcal{M}_0 : H_{\mu_1,\mu_2,\mu_3,\mu_4}\}$, $\{\mathcal{M}_1 : H_{\mu_1,\mu_2,\mu_a}\}$, where $\mu_a = \mu_3 = \mu_4$ and $\{\mathcal{M}_2 : H_{\mu_a,\mu_b}\}$, where $\mu_a = \mu_3 = \mu_4$ and $\mu_b = \mu_1 = \mu_2$. Note that these three models differ in model dimensionality, that is, the

number of parameters differ between the models. Model selection tools can choose between these models using a trade-off between model fit and model complexity, as further described in Section 3.2.

In Step (2) the models may differ in dimensionality, but returning to our example shown in Figure 3.1 there is no difference in the number of dimensions between the models under investigation. Also in the example of Jongmans et al. (2003) the two statistical models do not differ in dimensionality. Therefore, in this paper we propose another step in the model selection procedure of Kieseppä (2001), called Step (2'). In this new step, a set of models is specified that differ in the inequality constraints formulated between the parameters of interest. In this way, the models under investigation at Step (2) may differ in dimensionality and at Step (2') they may differ in admissible parameter space. Note that both dimensionality and specifying inequality constraints reduces the admissible parameter space volume. This implies that for model selection procedures in both Step (2) and Step (2') we are selecting on the basis of the size of the parameter space. However, in Step (2') the notion of size includes the size restrictions following inequality constraints between statistical parameters. As we show in the remainder of this paper, standard model selection procedures are equipped to select the best model in Step (2), but they are not equipped to select the best model in Step (2').

3.2 What Goes Wrong?

In the previous sections it has become apparent that background knowledge can be expressed in a set of relevant statistical models that differ in the size of the parameter space, either by changing the dimensionality of a model, or by restricting the parameter space using inequality constraints. In this section we inspect the bahaviour of two classical model selection criteria for both situations.

A number of so-called information criteria (ICs) is available, most notably Akaike's IC, or AIC, (Akaike, 1973, 1981), and the Bayesian IC, or BIC, (Schwarz, 1978). These ICs optimize entirely different things (see Appendix A for a technical introduction): AIC minimizes the expected Kullback-Leibler divergence to the true distribution; and BIC is focused on maximizing the marginal likelihood of the model. So, both criteria are designed to pursue an entirely different goal. Also, each IC is derived from a different starting point (the Kullback-Leibler divergence for the AIC, and the marginal likelihood for the BIC). But, interestingly enough, the result always consist of two parts,

$$\text{IC}_t = -2 \log f(y|\hat{\theta}_t) + \lambda_t \ , \tag{3.1}$$

where the subscript t refers to model $\mathcal{M}_t (t = 1, \ldots, T)$, $f(y|\hat{\theta}_t)$ denotes the likelihood of the model parameters of the data y evaluated at the maximum likelihood estimate of $\hat{\theta}_t$. The parameter λ_t differs across the IC's (see Appendix A for more details):

- $\lambda_t = 2d$ for the AIC_t,
- $\lambda_t = d \log(n)$ for the BIC_t,

where n is the sample size, and d is the number of parameters in the model. Both expressions somehow include the number of parameters in the model; we use the number of parameters and the dimensionality of the model interchangeably.

It is interesting to note that λ in Equation (3.1) is primarily an expression for the asymptotic bias in the log-likelihood as an estimate for the target function, and not, as is stated in many articles, deliberately put in the formula as an expression of simplicity (see e.g., Burnham & Anderson, 2004). The strength of the ICs lies in the fact that the appearance of dimensionality is a result of the derivation. The interpretation of the Λ term as being a measure for simplicity arises only post-hoc. The result, however, is intuitively meaningful: the likelihood of the best fitting hypothesis within the model is a measure of model fit, and the number of parameters of the model is a measure of complexity. The greater the number of dimensions, the greater the compensation for model complexity becomes. So, adding a parameter should be accompanied by an increase in model fit to accommodate for the increase in complexity.

Reconsider the example of Section 3.1.2 where the number of dimensions is the same for $\mathcal{M}_1$ and $\mathcal{M}_2$. The second term on the right hand side of Equation (3.1) will not make the difference between both models. Now suppose the maximum of the likelihood is completely located in the lower triangle of Figure 3.1, consequently the constraints of $\mathcal{M}_2$ are not supported by the data. In this situation the first term on the right hand side of Equation (3.1) will be higher for $\mathcal{M}_2$ than for $\mathcal{M}_1$ and the model selection procedure is able to distinguish both models.

But now consider the situation that the maximum of the likelihood is completely located in the upper triangle of Figure 3.1. In this situation both terms on the right hand side of Equation (3.1) will be the same for $\mathcal{M}_1$ and $\mathcal{M}_2$. In this latter situation neither AIC, nor BIC can distinguish $\mathcal{M}_1$ from $\mathcal{M}_2$ because $\mathrm{AIC}_1 = \mathrm{AIC}_2$ and $\mathrm{BIC}_1 = \mathrm{BIC}_2$. This result is counterintuitive and unwanted because $\mathcal{M}_2$ is clearly more restricted than $\mathcal{M}_1$ in virtue of the inequality constraint. So, a model selection procedure should select $\mathcal{M}_2$ as the best model if there is no difference in the maximum of the likelihood, or conversely model fit. Unfortunately, neither of the elements in Equation

Equation (3.1) manages to accommodate background knowledge as expressed in inequality constraints between the parameters of interest.

To sum up, classical ICs are equipped to choose between a number of different statistical models if the model fit differs between models, or if the dimensionality differs (Step (2) of Kieseppä, 2001). However, if we want the ICs to bring out the differences between statistical models in Step (2'), we need to revise the classical ICs and this is what we do in the next section.

3.3 Model Selection Criteria Revised

In this section we introduce a revision for subsequently the AIC and BIC: the order restricted IC, or ORIC (Anraku, 1999), and the *prior adjusted* BIC (Romeijn, Van de Schoot & Hoijtink, 2010). In Appendix B more details are provided about these revised ICs. The revision of the AIC is only described for a certain class of statistical models, namely analysis of variance models (ANOVA)

$$y_i = \sum_{j=1}^{J} \mu_j d_{ij} + \varepsilon_i, \tag{3.2}$$

where y_i is the observation of the dependent variable of person i ($i = 1, \ldots, N$), μ_j is the mean of group j ($j = 1, \ldots, J$), and d_{ij} denotes the group membership of a person, with 0 denoting not being a member of the group and 1 denoting being a member of the group. The residuals, ε_i, for each group are assumed to be normally distributed with mean zero and variance σ^2.

Using $\theta = \{\boldsymbol{\mu}, \sigma^2\}$ for ANOVA models the likelihood is given by

$$f(y|\boldsymbol{\mu}, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{(y_i - \sum_{j=1}^{J} \mu_j d_{ij})^2}{2\sigma^2} \right\}. \tag{3.3}$$

Concerning the revised ICs, it is important to realize that the ORIC and the *prior adjusted* BIC are not post hoc solutions. In fact, these alternative model selection criteria are developed from the same starting points as the original

criteria: the Kullback-Leibler divergence for the ORIC, and the marginal likelihood for the *prior adjusted* BIC. At some point the derivations of these revised ICs have been altered to account for inequality constraints between the parameters (see Appendix B for more details) and interestingly enough, the results are quite intuitive. While both revised IC is derived from a different starting point, again the result always consists of two parts,

$$\mathrm{IC}_t = -2 \log f(y|\hat{\theta}_t) + \lambda_t \ , \tag{3.4}$$

where $f(y|\hat{\theta}_t)$ denotes the likelihood of the data y evaluated at the maximum likelihood estimates of the model parameters, denoted $\hat{\theta}_t$, and λ_t is the estimated bias for the maximized log-likelihood value. The last term of equation Equation (3.4) again differs across the IC's:

- $\lambda_t = 1 + \sum_{l=1}^{q_m} \mathrm{LP}_l \cdot l$, for the ORIC, where q_m is the number of distinct mean values ($l = 1, \ldots, q_m$) and LP is a level probability (see Appendix B.1 for a more detailed description) and '1' refers to the unknown variance term;
- $\lambda_t = d \ \log(n) - 2 \log h_t(\hat{\theta})$, for the *prior adapted* BIC, where $h_t(\hat{\theta})$ is the prior evaluated at the maximum likelihood estimates (see Appendix B.2 for a more detailed description). Note that $h_1(\hat{\theta}) = 1$ for the unconstrained model and $h_t(\hat{\theta}) = c_t \times h_1(\hat{\theta})$ for model t as we will show below in Equations (3.5) and (3.6).

In their own way both λ-terms express the size of the admissible parameter space, because each can incorporate the inequality constraints that might be imposed on the parameters. Let us elaborate on both revised ICs and their behaviour in Step (2') of Section 3.1.3.

First, the literature contains one modification of Akaikes information criterion that can be used in the context of inequality constrained analysis of variance models with equal group sizes. It is called the order restricted information criterion, or ORIC (Anraku, 1999). It can be used for the evaluation of models differing in the order restrictions among a set of

means. Inequality constraints are taken into account in the estimation of the likelihood and in the penalty term of the ORIC.

Anraku (1999) showed that the number of parameters in an order restricted model can be expressed as a function of the level probabilities that are used when testing an inequality constraints hypothesis. A level probability is the probability that that the order restricted maximum likelihood estimator of the means for the hypothesis at hand consists of l distinct value given that for each group the data come from a distribution with the same mean and variance. Computation of level probabilities can be done via simulation, see pp. 78-81, Silvapulle and Sen (2004), see also Kuiper and Hoijtink (2010), or Kuiper et al. (2010) for software. In short, for inequality constrained hypotheses the penalty term of the ORIC can computed by replicating data sets from a population where for each group the data comes from a distribution with the same mean and variance. The penalty term is then easily computed by estimating the percentage of replicated data sets that satisfy the constraints of the model under investigation, see Appendix B.1 for more details and a description how to compute the ORIC for our simple example.

Secondly, as indicated in Section A, the BIC is an approximation of the marginal likelihood of the model. Since the BIC is a Bayesian model selection procedure, a prior density over the statistical hypotheses in the model naturally shows up in the derivation. In the derivation of the ordinary BIC, it is shown that the influence of that prior is relatively small, and for increasing data sets becomes negligent. In the derivation of the *prior adjusted* BIC (Romeijn et al., 2010), however, the influence of the priors is retained, and factored into the term expressing the simplicity of the model. The prior distribution used is the encompassing prior approach, further explored in Klugkist et al. (2005) in combination with a uniform prior distribution. In short, the method employs an encompassing prior, $h_t(\hat{\theta})$ for the model parameters θ and is specified for the unconstrained model, for our example $\mathcal{M}_1$ in Figure 3.1.

Then, the prior distribution of each nested model, $\mathcal{M}_t$ $(t = 2, \ldots, T)$, can be derived from the prior of the unconstrained model, $h_1(\hat{\theta})$. If $h_1(\hat{\theta})$ is defined for the total parameter space, then the prior distribution for a constrained model, $h_t(\hat{\theta})$, is then proportional to $h_1(\hat{\theta})$ and zero elsewhere:

$$h_t(\hat{\theta}) : \begin{cases} c_t^{-1} h_1(\hat{\theta}) & \text{if} \hat{\theta} \in \mathcal{M}_t \\ 0 & \text{otherwise} , \end{cases} \tag{3.5}$$

where c_t is a normalization constant given by

$$c_t = \int_{\mu} \in \mathcal{M}_t h_1(\hat{\theta}) d\theta \ . \tag{3.6}$$

If $h_1(\theta)$ is chosen constant, for example $h_1(\theta) = 1$, in that case the *prior adjusted* BIC is able to distinguish inequality constrained hypotheses.

Consider the example of Section 3.1.2, where the number of dimensions is the same for $\mathcal{M}_1$ and $\mathcal{M}_2$. Whenever the first term on the right hand side of Equation (3.4) is equal for $\mathcal{M}_1$ and $\mathcal{M}_2$ it should be $h_t(\hat{\theta})$ that makes the difference between both models. Because of the truncated prior distribution, as is argued in Romeijn et al. (2010), $h(\hat{\theta})$ is the same as the inverse of the volume of the model under investigation.

For our example $h_1(\hat{\theta}) = 1$ and $c_t = .5$, since only halve of the parameter space is admissible, then because of Equation (3.5) $h_2(\hat{\theta}) = 2$. Consequently, for the situation that the first term on the right hand side of Equation 3.4 does not differ between $\mathcal{M}_1$ and $\mathcal{M}_2$, *prior adapted* BIC_2 < *prior adapted* BIC_1 and the *prior adapted* BIC will select $\mathcal{M}_2$ as the best model.

Note that, again, these revisions are not adjustments of equation Equation (3.1) to account for inequality constraints, but that they are derived from scratch. The result is surprisingly intuitive: the parameter space volume shows up as part of the selection criteria. In the next section we argue that the adjustments of the model selection criteria are refinements of the post-hoc interpretation of simplicity.

3.4 Refinement of the Notion of Simplicity

What have we learned so far? Scientists have expectations in the form of (i) the number of parameters and (ii) inequality constraints between the parameters. Classical ICs (i.e., AIC and BIC) can select among models when they differ in fit or dimensionality, but not when the only difference is in the volume of the admissible parameter space as affected by inequality constraints between the parameters of interest. We saw that the revised ICs (ORIC, and *prior adjusted* BIC) do not suffer from this problem. We have also seen that classical ICs are a combination between the fit and complexity. The likelihood of the best fitting hypothesis within the model is considered as a measure of model fit, and the dimensionality of a model is considered as a measure of complexity. Finally, we saw that revised IC's include each in their own way the size of the model.

Forster and Sober, among many others, build a good case for interpreting the dimensionality, appearing in the classical ICs, as a measure for complexity (Forster, 2002; Sober, 2006). In this section we argue that for the very same reasons the λ-terms from equation Equation (3.4) of the revised ICs can be interpreted as a measure for complexity as well. In what follows, we draw a parallel between dimensionality and complexity on the one hand and model size and complexity on the other, so that the derivations of the revised ICs provide us with a refined notion of complexity.

To motivate this parallel, we need to elaborate on the exact relation between simplicity and the number of parameters. To do so, consider model selection from the falsificationist perspective of Popper (1963). A model that deems fewer possibilities admissible is easier to falsify and if such a model is supported by the data it should be rewarded for its risky prediction. As Popper wrote, "Confirmations should count only if they are the result of risky predictions; that is to say, if, unenlightened by the theory in question, we should have expected an event which was incompatible with the theory an event which would have refuted the theory" (*ibid.*, pp. 33-39). Thus, the

more possibilities a researcher is willing to rule out, the more attractive a model becomes.

Model selection procedures connect this idea of ruling out possibilities with the notion of simplicity. They favor models that allow for fewer possibilities, and call such models simpler. Relating this to the AIC and BIC, where complexity is measured as the dimensionality of the model- see λ in equation Equation (3.1)- the more dimensions are 'shaved' away using the principle of Ockham's razor,[3] the simpler the model becomes. The question now is whether the same holds for the revised ICs. That is, can we interpret λ in equation Equation (3.4) as a measure for complexity as well?

In Section 3.1 we showed that researchers do not only specify the number of parameters in their statistical models, but that they also use inequality constraints between the statistical parameters. As indicated, there is a natural relation between cutting down the number of parameters and ruling out possibilities. But there is also such a natural relation between introducing inequality constraints and ruling out possibilities. When specifying such constraints, a researcher also rules out possibilities and hence 'shaves' away parameter space volume. Compare, for example, the models in Figure 3.1, which differ in admissible parameter space but not in model dimensionality. Here the parameter space is not restricted by the number of parameters, but by inequality constraints between the parameters of interest.

Returning to Popper and his "risky predictions", adding more constraints to a model leads to more risky models. Looking again at the three models in Figure 3.1, we can easily see that a researcher who specifies the model $\{\mathcal{M}_2 :$

[3] In the 14th-century the logician William of Ockham formulated a principle later referred to as Ockham's razor and is often summarized as follows "entities should not be multiplied beyond necessity". This quote, however, is (so far) not found in his writings. What has been found is: "a plurality should not be assumed without necessity, as has often been said" (Derkse, 1993). In other words, any model should make as few assumptions as possible. The interpretation is often formulated like "all things being equal, the simplest solution tends to be the best one".

$H_{\mu_2>\mu_1}\}$ takes more risks than a researcher who specifies an unconstrained model $\{\mathcal{M}_1 : H_{\mu_1,\mu_2}\}$ since less parameter space is admissible. Recall that in the falsificationist perspective, the ICs select those models that have a good fit and at the same time give off riskier predictions, and that this quality of generating risky predictions was associated with the complexity of the model. We have seen that risky predictions are associated with model size, both in terms of number of dimensions and in terms of admissible parameter space volume, and we have also seen that the revised ICs select models on the basis of both dimensionality and this admissible volume. Our suggestion is that we can therefore interpret both dimensionality and admissible parameter space volume as components of model complexity.

In sum, our proposal is to refine the notion of complexity along the foregoing lines. It is not the dimensionality of a model that determines the simplicity of a model, but it is parameter space volume, which is a combination of model dimensionality and the volume of the admissible parameter space. In our exposition the latter is a result of inequality constraints between the parameters of interest. The revised model selection criteria, each in their own way, give expression to this new notion of complexity. Moreover, just like for the original ICs, these expression of complexity are not put in by hand, but drop out of the independently motivated derivations.

3.5 Conclusion

In this paper we discussed the role of a particular kind of background knowledge in statistical model selection. We showed that when models only differ in volume of admissible parameters space expressed by inequality constraints, classical model selection criteria (AIC, and BIC) fail to choose between these models. Drawing on recent developments in statistics, we presented revisions of these classical criteria that manage to cope with such constraints: the order restricted information criterium, or ORIC (Anraku,

1999), and the *prior adjusted* BIC (Romeijn et al., 2010). Finally, we argued that these revisions provide us with a refined notion of model complexity. This is good news for scientists because it provides them with an interpretation of the new model selection tools, and it is good news for philosophers because, we claim, they are given a more nuanced view of what the complexity of a model amounts to.

APPENDICES

A Two Often Used Model Selection Criteria

In this appendix we provide a short technical introduction to AIC (Akaike, 1973), and BIC (Schwarz, 1978). For a detailed comparison between the two ICs, we refer to Hamaker, van Hattum, Kuiper and Hoijtink (2009).

A.1 AIC

The AIC (Akaike, 1973, 1981) is an information-theoretic model selection method based on the Kullback-Leibler (K-L) distance (Kullback & Leibler, 1951). The K-L distance quantifies the discrepancy between two probability distributions. Note that by 'distance' we mean divergence and not a simple Euclidean distance. For one, the 'distance' relation between two points is not symmetric.

Let $f(\cdot|\theta^*)$ be the true model with a probability distribution over the sample space. Furthermore, let $f(\cdot|\theta)$ denote an approximating model which is a probability distribution. Note that both $f(\cdot|\theta^*)$ and $f(\cdot|\theta)$ are probability distributions, but the main difference is that while $f(\cdot|\theta^*)$ is infinitely complex with an infinite number of parameters, $f(\cdot|\theta)$ is characterized by a finite set of parameters θ. The K-L distance is then the amount of information lost when $f(\cdot|\theta)$ is used to approximate $f(\cdot|\theta^*)$, denoted by $\delta(\theta^*;\theta)$:

$$\delta(\theta^*;\theta) = \mathrm{E}_{f(y|\theta^*)}\Big[\log f(y \mid \theta^*) - \log f(y \mid \theta)\Big] \,, \tag{3.7}$$

where y is observed data.

The criterion is that this K-L distance needs to be minimized in order to obtain a model selection criterion. To do so Akaike (1973) considered the data generating mechanism to be fixed, and only $f(\cdot|\theta)$ varies over a space of models, making the K-L information criterion equivalent to maximizing the second term on the right hand side of (3.7). The K-L information criterion cannot be used directly in model selection because it requires knowledge of the true parameter values, θ^*, and clearly we do not have such knowledge. Using a hypothetical cross-validation data set x, which, just like the observed data set y, arises from $f(\cdot|\theta^*)$, Akaike (1973) found that the maximized log-likelihood value is a biased estimate of the expected estimated relative K-L distance:

$$\mathrm{E}_{f(x|\theta^*)}\left[\mathrm{E}_{f(y|\theta^*)}\left\{\log f(x \mid \hat{\theta}_y)\right\}\right] \approx \log f(y \mid \hat{\theta}_y) - 2d \ , \tag{3.8}$$

where $\hat{\theta}_y$ is the expected maximum log likelihood estimate of the observed data set and where the bias term d is approximately equal to the number of estimable parameters in the approximating model. The expression on the right hand side of Equation (3.8) is called the AIC. The smaller the numerical expression of the AIC the more attractive the model under investigation becomes.

The AIC, namely $\log f(y \mid \hat{\theta}_y) - 2d$, can be considered as an approximately unbiased estimator of the K-L distance for large samples. The interpretation of the AIC is, among a finite set of models being compared that the model satisfying the lowest AIC value is expected to result in the minimum loss of information if this model were fitted to a future sample of observations from the same underlying data generating mechanism.

A.2 BIC

The Bayesian information criterion (BIC), or Schwarz Criterion (Schwarz, 1978) is a criterion for model selection among a class of parametric models based on a Bayesian model selection approach. In short, Bayes' method says

that statistical models should be compared by their posterior probabilities. That is, we should choose the model M_t $(t = 1, \ldots, T)$, which has the largest posterior model probability given the observed data y. The posterior model probability for model M_t is given by

$$p(M_t|y) = \frac{f(y|M_t)p(M_t)}{\sum_{i=1}^{K} f_i(y|M_t)p(M_i)} \ , \tag{3.9}$$

where k is a model index (with $k = 1, \ldots, K_i$), $f(y|M_t)$ is the marginal model probability of data y for model M_t, and $p(M_t)$ is the prior probability for model M_t. Schwarz (1978) assumed that the prior probabilities of all models are equal and that there is a flat, uniform prior distribution over parameter values in each model containing the same amount of information as one single observation (Raftery, 1995). Then finding the model with the largest posterior probability is the same as choosing the model with the largest marginal model probability. He showed that

$$-2 \log f(y|M_t) \approx -2\log f(y|\hat{\theta}_y) + d \ \log(n) \ , \tag{3.10}$$

where $\hat{\theta}_y$ is the maximum likelihood estimator, and d is the dimensionality of the model.

The expression on the right hand side of Equation (3.13) is called the BIC. Again, the smaller the numerical expression of the BIC the more attractive the model under investigation becomes. The BIC, namely $-2\log f(y|\hat{\theta}_y) + d \ \log(n)$, can be considered an approximately unbiased estimator of the marginal likelihood for large samples and a specific prior distribution. Raftery (1995) indicates that the BIC may be viewed as a predictive score and can be used for making out-of-sample predictions, i.e. inferences to observations from different populations.

B Revised Model Selection Criteria

In this appendix we provide a short technical introduction to the ORIC (Anraku, 1999), and the *prior adjusted* BIC (Romeijn et al., 2010).

B.1 Revision of AIC

The literature contains one modification of Akaike's information criterion that can be used in the context of inequality constrained ANOVA models. It is called the order restricted information criterion (ORIC), see Anraku (1999); see also Kuiper and Hoijtink (2010). The ORIC can be used for the evaluation of models differing in the order restrictions among a set of means.

Like other information criteria, the ORIC is based on the likelihood and a penalty term, equal to

$$\text{ORIC} = -2\log f(y|\hat{\theta}) + 2\Big[1 + \sum_{l=1}^{q_m} \text{LP} \cdot l\Big], \tag{3.11}$$

where the term between square brackets denotes the penalty term for the constrained means where the '1' refers to the unknown variance term in ANOVA models. Furthermore, q_m is the number of distinct mean values ($l = 1, \ldots, q_m$), and LP is a level probability, explained below. In the AIC, the maximum likelihood estimators of $\hat{\theta}$ are the values that maximize the log likelihood for the statistical model under investigation (Burnham & Anderson, 1998), but in the ORIC the order-restricted maximum likelihood estimator for the means must be found.

Anraku (1999) showed that the number of parameters in an order restricted model can be expressed as a function of the level probabilities that are used when testing an inequality constrained hypothesis. A level probability is the probability that that the order restricted maximum likelihood estimator of the means for the hypothesis at hand consists of l distinct value given that for each group the data come from a distribution with the same mean and variance.

Consider the example of Section 3.1.2, where the number of dimensions is the same for $\mathcal{M}_1$ and $\mathcal{M}_2$ and suppose the maximum of the likelihood is completely located in the upper triangle of Figure 3.1, then it should be the penalty term which differentiates $\mathcal{M}_1$ from $\mathcal{M}_2$. With no inequality constraints imposed on the means the ORIC is equal to the AIC, so for $\mathcal{M}_1$ the number of parameters is equal to 2 (plus 1 because of the unknown variance term). For inequality constrained hypotheses the penalty term of the ORIC can computed by replicating data sets from a population where for each group the data comes from a distribution with the same mean and variance (see Appendix B.1). For our example, a number of data sets are generated with $\mu = \mu_1 = \mu_2$ and $\sigma^2 = \sigma_1^2 = \sigma_2^2$. For each data set the mean for both groups is then computed $(\bar{x}_1, \bar{x}_2)$. Now for $\mathcal{M}_2$, q_m is equal to 2, because there are two levels: (1) $\bar{x}_2 < \bar{x}_1 \Rightarrow \mu_1 = \mu_2$, and (2) $\bar{x}_2 > \bar{x}_1 \Rightarrow \mu_2 > \mu_1$. The level probability (LP) is then computed by estimating the percentage of replicated sample means that satisfy the constraints of the model under investigation for each level q_m, in our example $LP_1 = P(\bar{x}_2 < \bar{x}_1) = .5$ and $LP_2 = P(\bar{x}_2 > \bar{x}_1) = .5$. Now, $\lambda_1 = 2$ and $\lambda_2 = .5 \cdot 1 + .5 \cdot 2 = 1.5$ (both plus 1 for λ_1 and λ_2 because of the unknown variance term). Consequently, for the situation that the first term on the right hand side of Equation (3.4) does not differ between $\mathcal{M}_1$ and $\mathcal{M}_2$, $\text{ORIC}_2 < \text{ORIC}_1$ and the ORIC will choose $\mathcal{M}_2$ as the best model.

In general, for inequality constrained hypotheses the penalty term of the ORIC can computed by replicating data sets from a population where for each group the data comes from a distribution with the same mean and variance. The penalty term is then easily computed by estimating the percentage of replicated data sets that satisfy the constraints of the model under investigation. Computation of level probabilities can be done via simulation, see pp. 78-81, Silvapulle and Sen (2004), see also Kuiper and Hoijtink (2010), or Kuiper et al. (2010) for software.

B.2 Revision of BIC

The derivation of the *prior adapted* BIC runs along the same lines as the derivation of the original BIC by Schwarz (1978) (see also, Raftery, 1995): the BIC of a model M_t is an approximation of the marginal probability of the data y given model M_t, see Equation (3.9). In the derivation it is shown that

$$\begin{aligned}\log f(y|M_t) \approx \log f(y|\hat{\theta}_y) - (d/2)\log(n) \\ + \log h(\hat{\theta}) + (d/2)\log(2\pi) - \frac{1}{2}\log|\mathbf{i}|,\end{aligned} \tag{3.12}$$

where $h(\hat{\theta})$ is the prior probability density at $\hat{\theta}_y$, which is the maximum likelihood estimate given the data y. The parameters $\hat{\theta}_y$ is the maximum likelihood estimate of the model given the data y. Furthermore, d is the number of parameters in the model, n is the sample size, and $|\mathbf{i}|$ denotes the expected Fisher information matrix for one observation.

Now, the first term on the right hand side of (3.12) is of order $O(n)$, the second term on the right hand side of (3.12) is of order $O(\log n)$, while the last three terms in (3.12) are of order $O(1)$ or less if $n \to \infty$. Removing the terms with order $O(1)$ or less and multiplying with -2 gives the BIC

$$-2\log f(y|M_t) \approx -2\log f(y|\hat{\theta}_y) + d\ \log(n)\ , \tag{3.13}$$

The terms of order $O(1)$ or less can be considered as an error of the estimation of $\log f(y|M_t)$, but arguably, they can be ignored because the first two terms will dominate the equation as n tends to infinity. Moreover, it was shown by Raftery (1995) that the error is much smaller for a reasonable choice of the prior distribution, namely a normal mean with $\hat{\theta}$ and variance matrix $\mathbf{i}^{-1}$.

So far for the BIC, let us focus on inequality constrained models; see Romeijn et al. (2010). There are roughly three cases when comparing an inequality constrained models: (1) the models may differ in dimensionality; (2) they may differ in maximum likelihood; or (3) they merely differ in the

volume of admissible parameter space thereby implying that they neither differ in dimensionality nor in maximum likelihood. In the first two of these cases the *prior adapted* BIC boils down to the original BIC. For the last case, it can be shown that none of the terms in Equation (3.12) differ when we compare such inequality constrained models, except for $h(\hat{\theta})$. Even if it is only of $O(1)$, it makes the difference between the models under investigation and $h(\hat{\theta})$ should therefore be left in the equation. Ignoring terms of this order is unwanted.

Instead, we must include the term pertaining to the prior probability density over the model in the BIC, thus creating the *prior adapted* BIC:

$$-2\log f(y|M_t) \approx -2\log f(y|\hat{\theta}_y) + d\ \log(n) - 2\log h_t(\hat{\theta})\ . \tag{3.14}$$

Restricting ourselves to a uniform prior probability density over the models, and fixing the volume M_1 to 1, we obtain that $h_1(\hat{\theta}) = 1$. In that case $h_t(\hat{\theta})$ can be interpreted as the inverse of the volume of the model M_k, as argued in Romeijn et al. (2010). Whenever the first two terms in the equation are equal for the models of interest, the third term in *prior adapted* BIC makes the difference.

PART II

Statistics

4

Psychological Functioning, Personality and Support from Family: An Introduction to Bayesian Model Selection

Van de Schoot, R., Hoijtink, H., Mulder, J., Van Aken, M., Dubas, J.S., Orobio de Castro, B., Meeus, W. & Romeijn, J.-W.

Manuscript under review

Abstract

Most researchers have specific expectations concerning their research questions. These may be derived from theory, empirical evidence, or both. Yet despite these expectations, most investigators still use null hypothesis testing to evaluate their data, that is, when analyzing their data they ignore the expectations they have. In the present article, Bayesian model selection is presented as a means to evaluate the expectations researchers have, that is, to evaluate so called informative hypotheses. Although the methodology to do this has been described in previous articles, these are rather technical and have mainly been published in statistical journals. The main objective of the present article is to provide a basic introduction to the evaluation of informative hypotheses using Bayesian model selection. Moreover, what is new in comparison to previous publications on this topic is that we provide guidelines on how to interpret the results. Bayesian evaluation of informative hypotheses is illustrated using an example concerning psychosocial functioning and the interplay between personality and support from family.

Statistical hypothesis evaluation has moved beyond simply testing the traditional null hypothesis: 'nothing is going on'. New developments allow researchers to learn more from their data than merely that the null hypothesis is rejected. In this paper we will introduce one such development: the evaluation of informative hypotheses using Bayesian model selection (Hoijtink, 1998, 2000, 2001; Hoijtink, Klugkist & Boelen, 2008; Klugkist & Hoijtink, 2007; Klugkist et al., 2005; Kuiper et al., 2010; Laudy, Boom & Hoijtink, 2005; Laudy & Hoijtink, 2007; Mulder, Hoijtink & Klugkist, 2009; Mulder, Klugkist et al., 2009)

In practice, researchers have specific expectations about their research questions which may be derived from theory, empirical evidence, or both. For example, suppose that most previous studies find that resilient adolescents (R) score lower on internalizing problems than under-controlled adolescents (U) who, in turn, score lower on internalizing problems than over-controlled adolescents (O): $H_1 : R < U < O$. Suppose that a new article reports the opposite result: $H_2 : O < U < R$. Hypotheses such as H_1 and H_2 will be called informative hypotheses because they contain information about the ordering of the means. After obtaining new data, the research question for this example could be: Which of both informative hypotheses receives more support from the data?

Bayesian model selection can be used to provide an answer to this question. The result might be that there is 40 times more support in the data for H_1 than for H_2. As a consequence we not only have a direct answer to our research question, but we also have an indication of how much better one hypothesis is, compared with another hypothesis.

Note that neither H_1, nor H_2 resemble the traditional null hypothesis (nothing is going on, $H_0 : O = U = R$), nor the traditional alternative hypothesis (something is going on, but it is not specified what). We argue that many researchers are not particularly interested in these traditional

hypotheses (Cohen, 1990, 1994; M. D. Lee & Pope, 2006; M. D. Lee & Wagenmakers, 2005; Trafimow, 2003). It may already be known from previous research that the means are not equal, so why not use this information? Furthermore, rejecting the null hypothesis does not imply that either $H_1 : R < U < O$ or $H_2 : O < U < R$ is going on. For a more detailed comparison of traditional null hypotheses testing and Bayesian evaluation of informative hypotheses we refer to Hoijtink, Huntjes, Reijntjes, Kuiper and Boelen (2008); Hoijtink and Klugkist (2007); or Kuiper and Hoijtink (2010).

Recent developments in statistics, rendered tools that enable the direct evaluation of predetermined informative hypotheses. Although applied articles are emerging in the field of the social sciences (Kammers et al., 2009; Laudy, Zoccolillo et al., 2005; Meeus, Van de Schoot, Keijsers et al., 2010; Van de Schoot et al., 2009; Van de Schoot & Wong, 2010; Van Well et al., 2009), an easy-to-read introduction to the evaluation of informative hypotheses and general guidelines on how to interpret the results are still lacking. The purpose of the current paper is therefore (i) to present an introduction to the evaluation of informative hypotheses using Bayesian model selection, and (ii) to provide guidelines on how to interpret the results. The methodology is illustrated using two examples: a simple example to introduce the components of Bayesian model selection and an example evaluating whether psychosocial functioning is the result of the interplay between personality and support from family (Van Aken & Dubas, 2004). Before introducing the methodology, let us first elaborate on what exactly is meant by an informative hypothesis.

4.1 What are Informative Hypotheses?

Informative hypotheses contain information about the ordering of means, regression coefficients or any other statistical parameter and can be constructed using the following constraints:

1. Larger than, denoted by $'>'$;
2. Smaller than, denoted by $'<'$;
3. Equal to, denoted by $'='$.

Such expectations about the ordering of parameters can stem from previous studies, a literature review or even academic debate. If no information is available about the ordering of two parameters, they are separated by comma. An informative hypothesis can consist of combinations of constraints among, for example, a set of means (note that a mean will be denoted by the symbol μ). An example is the hypothesis $H_1 : \{\mu_1, \mu_2\} < \mu_3 = \mu_4$, where group 1 and 2 are both expected to have smaller mean scores than group 3 and 4. Also, group 1 and 2 are not, but group 3 and 4 are restricted to have the same value. (In)equality constraints can also be used between (combinations of) means and a threshold, for example, $H_2 : \mu_1 - \mu 2 > .20; \mu_3 - \mu_4 < .50$, where the difference between the means of group 1 and 2 is expected to be larger than .20 and where the difference between group 3 and 4 is expected to be smaller then .50. If no constraints are imposed on any of the means, and any ordering is equally likely, the unconstrained hypothesis $H_3 : \mu_1, \mu_2, \mu_3, \mu_4$ is obtained.

The major advantage of evaluating a set of informative hypotheses is that prior information can be incorporated into an analysis. As argued by Howard et al. (2000), replication is an indispensable tool in the social sciences. Evaluating informative hypotheses fits within this framework because results from different research papers can be translated into different informative hypotheses. The method of Bayesian model selection can provide each informative hypothesis with the degree of support supplied by the data. As a result, the plausibility of previous findings can be evaluated in relation to new data, which makes the method described in this paper an interesting tool for replication of research results.

4.2 Bayesian Statistics

An important contribution Bayesian statistics can make to the social sciences is the evaluation of informative hypotheses using Bayesian model selection (for an introduction see Hoijtink, Klugkist & Boelen, 2008, and for a general introduction to Bayesian statistics see S. Lynch, 2007. It has proved to be a flexible tool that can deal with many types of constraints. Gelfand, Smith and Lee (1992) first showed how inequality constraints can be accounted for using Markov chain Monte Carlo methods (see also Hoijtink, 2000). By now, many researchers showed that this same approach also works in more complicated models. In this paper we show how to analyze informative hypotheses using a MANOVA as described in Mulder, Klugkist et al. (2009) (see also Mulder, Hoijtink & Klugkist, 2009). There is software corresponding to these papers that can deal with many types of (in)equality constraints in multivariate linear models: (M)AN(C)OVA, regression analysis, repeated measure analyses with time-varying and time-invariant covariates. Software is also available for ANCOVA (Klugkist et al., 2005); latent class analyses (Laudy, Boom & Hoijtink, 2005; Hoijtink, 1998, 2001) and order restricted contingency tables (Laudy & Hoijtink, 2007). A first attempt can best be made using the software programme 'confirmatory ANOVA' (Kuiper & Hoijtink, 2010) (see also Kuiper et al., 2010). Readers interested in the software can visit *www.fss.uu.nl/ms/informativehypothesis*.

What follows is a general description of the methodology used. We do not provide a full explanation of the analyses. Instead, references to more technical papers are provided for interested readers throughout this section.

4.2.1 Simple Example

To understand the methodology, consider the following simple example based on the data of Van Aken and Dubas (2004). Suppose the research question is whether the mean score on externalizing behavioural problems differs for over- (denoted by μ_O; $n = 158$) and under-controlled adolescents (denoted by

μ_U; $n = 207$). Furthermore, suppose the first hypothesis (H_A) postulates that there is no restriction between the means (that is, any combination of means is admissible). The second hypothesis (H_B) postulates that the externalizing behavioural problems of both groups are equal. The third hypothesis (H_C) postulates that over-controlled adolescents score lower on externalizing problem behaviour than under-controlled adolescents. Formally, the three hypotheses in this simple example are:

$$\begin{aligned} H_A &: \mu_O \ , \ \mu_U; \\ H_B &: \mu_O = \mu_U; \\ H_C &: \mu_O < \mu_U. \end{aligned} \tag{4.1}$$

Of course, these hypotheses can be evaluated using classical null hypothesis testing, or one-sided hypothesis testing. However, when there are more groups, more variables, or more constraints, null hypothesis testing is not the appropriate tool for evaluating which hypothesis (H_A, H_B, or H_C) receives most support from the data. To evaluate the set of hypotheses with Bayesian model selection, four components are needed that will be explained successively:

1. 'Admissible parameter space', or the expectations of the researcher;
2. The 'likelihood' of specific values of the parameters, representing the information in the data set with respect to μ_O and μ_U;
3. The 'marginal likelihood', which represents the support from the data for each hypothesis, combining model fit and model size;
4. This latter component is converted into the Bayes factor which is the model selection criteria.

4.2.2 Admissible Parameter Space

The first component is something we call the 'admissible parameter space' which results from the (in)equality constraints imposed on the means (see,

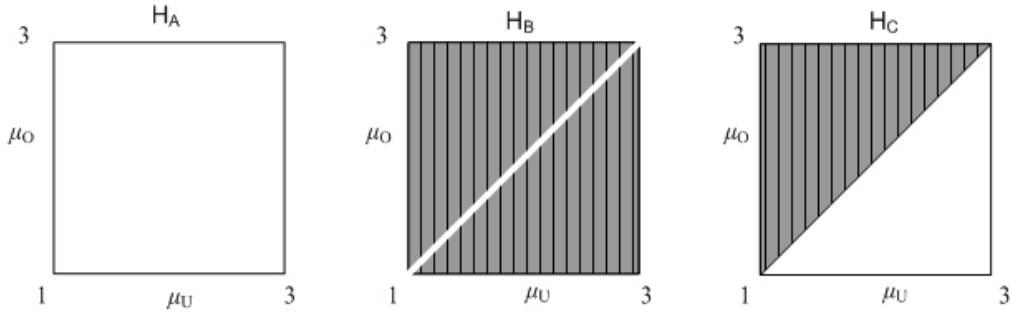


Figure 4.1: Admissible parameter space

e.g. Halpern, 2003, p. 12), for a philosophical introduction to parameter space, logical space, etc.). Let the squares in Figure 4.1 represent the total parameter space for all possible combinations of μ_O and μ_U in the population. The boundaries are determined by the scale of the variable, in this case the mean score on externalizing behavioural problems range between 1 trough 3 (mean = 1.56, SD = 0.37).

Now, let the *admissible* parameter space be the total of all possible combinations of μ_O and μ_U that satisfy the restrictions of each of the hypotheses (i.e. H_A, H_B, H_C). For H_A, every combination of μ_O and μ_U is permitted, and therefore, the admissible parameter space of H_A is equal to the total parameter space (left-hand panel of Figure 4.1). For H_B, μ_O and μ_U are assumed to be equal, which implies that only that part of the parameter space in which μ_O is equal to μ_U is admissible. This is represented by the diagonal in the center panel of Figure 4.1. For H_C, only combinations of μ_O and μ_U are permitted in which μ_O is smaller than μ_U, which results in the lower triangle in the right-hand panel of Figure 4.1. Note that, with respect to the admissible parameter space, the hypotheses can be ordered from a small parameter space to a large parameter space: H_B, H_C, H_A.

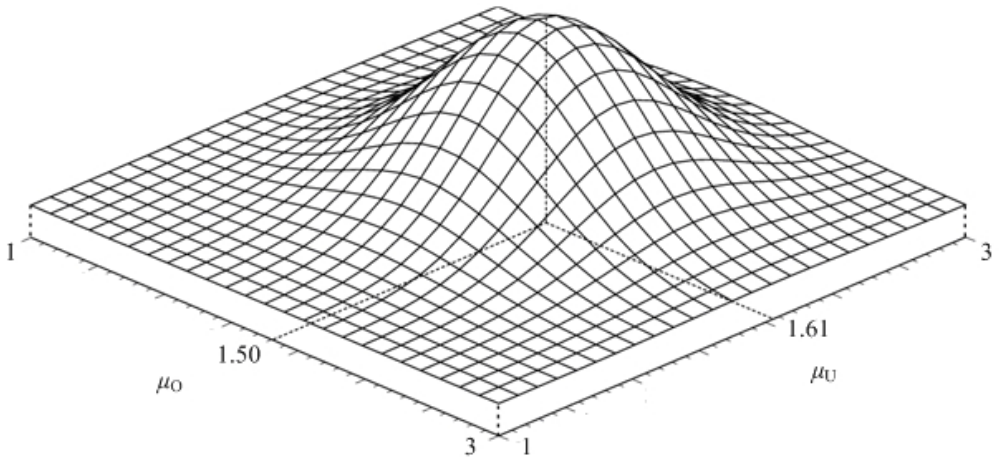


Figure 4.2: The likelihood function plotted as a function of μ_O and μ_U

4.2.3 Likelihood

The second component is the likelihood of specific values of the parameters, which is the representation of the information about the means in the data set (see, e.g. S. Lynch, 2007, pp. 36-37). In Figure 4.2 an illustrative likelihood function is plotted as a function of μ_O and μ_U. The higher this surface, the more likely the corresponding combination of μ_O and μ_U in the population becomes. In this example the sample means are 1.50 (SD = 0.33) for μ_O and 1.61 (SD = 0.39) for μ_U. So, given the data, the combination $\mu_O = 1.50$ with $\mu_U = 1.61$ is the most plausible, or the most likely combination of values for the population means. As can be seen in Figure 4.2, the likelihood function achieves its maximum for this combination. Other combinations of means are less likely. For example, the value of the likelihood function is much lower for the combination $\mu_O = 0.50$ and $\mu_U = 2.10$ and hence this combination of values is less likely to be the population values.

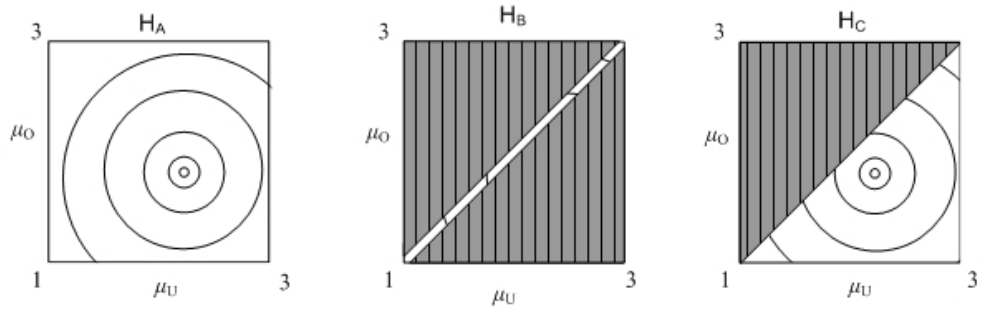


Figure 4.3: Likelihood function of the data within the admissible parameter space

4.2.4 Marginal Likelihood

The third component is the marginal likelihood (e.g. Chib, 1995; Kass & Raftery, 1995), which is a measure for the degree of support for each hypothesis provided by the data. The marginal likelihood is approximately equal to the average height of the likelihood function within the admissible parameter space. Let us elaborate on this.

Recall that Figure 4.1 presents the admissible parameter space for each hypothesis and Figure 4.2 displays the likelihood as a function of μ_O and μ_U. Both pieces of information are combined in Figure 4.3. The likelihood function in Figure 4.2 is now presented as a contour plot in Figure 4.3. The maximum value of the likelihood is located in the center of the smallest circle. Remember that as you move away from this center, the value of the likelihood of the combination of population means of μ_O and μ_U becomes smaller.

Because the admissible parameter space for H_A is equal to the total parameter space, the marginal likelihood of H_A can be computed as the average value of the likelihood in the total parameter space. This value is only meaningful in comparison to the marginal likelihood values of the other hypotheses under investigation. For H_B the average likelihood value is computed with respect to the diagonal in Figure 4.3 and for H_C, the average likelihood value is computed in the lower triangle in Figure 4.3. The marginal

likelihood values are $H_A = 2.83\ e^{-67}$; $H_B = 1.81\ e^{-68}$; $H_C = 5.71\ e^{-67}$. As can be seen, H_C has the highest value, followed by H_A and then H_B.

Note that comparing models can not only be done by comparing marginal likelihood values (e.g., Raftery, 1995), but also using the well known model selection criteria AIC (Akaike, 1981) and BIC (Schwarz, 1978). These selection criteria combine fit and complexity to determine the support for a particular model (Burnham & Anderson, 2004). However, in contrast to Bayesian model selection these classical criteria are as of yet unable to deal with hypotheses specified using inequality constraints (Mulder, Klugkist et al., 2009). If we take a closer look at the plots in Figure 4.3, we can also observe model fit and model size which are important components of the marginal likelihood.

Many high likelihood values are located within the admissible parameter space of H and H_A, but not in H_B. This indicates a good model fit for H_C and H_A, but not for H_B. Moreover, H_C has a smaller admissible parameter space compared with H_A and is therefore less complex. Furthermore, note that the likelihood values in the upper triangle in H_C are low and are not taken into account in the computation of the marginal likelihood for H_C, but are taken into account in the computation of the marginal likelihood for H_A. Consequently, the average likelihood value of H_C, and hence its marginal likelihood, is larger than the average likelihood, marginal likelihood, value of H_A. For H_B, only small likelihood values are within the admissible parameter space which implies a poor model fit. Although the admissible parameter space is smallest for H_B, the marginal likelihood is smaller than that of H_A and H_C because of the 'poor' model fit. In sum, the marginal likelihood rewards a hypothesis with the correct (in)equality constraints. This is because the average likelihood value is higher when many small likelihood values are not taken into account. The smaller the parameter space, the less complex a model becomes. Therefore, the methodology combines model fit and model size of a hypothesis.

4.2.5 Bayes Factors

As was shown by Klugkist et al. (2005) informative hypotheses can be compared using the ratio of two marginal likelihood values, resulting in Bayes factors (denoted by BF), see Kass and Raftery (1995), for a statistical discussion of the Bayes factor (see also Hoijtink, Klugkist & Boelen, 2008). The outcome represents the amount of evidence in favour of one hypothesis compared with another hypothesis. The results may be interpreted as follows: $BF = 1$ states that the two hypotheses are equally supported by the data; $BF > 1$ states that the support for one hypothesis is higher than for another hypothesis.

In our simple example, the BF for H_C compared to H_A can be obtained from the marginal likelihoods of both hypotheses:

$$BF_{CA} = \frac{M_C}{M_A} = \frac{5.71e^{-67}}{2.83e^{-67}} \approx 2 \ . \tag{4.2}$$

This Bayes factor, BF_{CA}, implies that after observing the data, H_C receives two times more support from the data than H_A. For BF_{CB} the result implies that H_C receives

$$BF_{BC} = \frac{M_B}{M_C} = \frac{1.81e^{-68}}{5.71e^{-67}} \approx 31 \ , \tag{4.3}$$

as much support from the data than H_B.

Recall that Bayes factors provide a direct quantification of the support in the data for the constraints imposed on the means. With support we mean: the trade-off between model size and model fit. Every researcher will agree that 31 times more support seems considerable while, for example, 1.04 times as much support does not. However, clear guidelines are not provided in the literature, nor do we provide them here. We refrain from doing so because we want to avoid creating arbitrary decision rules. Remember the famous quote about p-values: "[. . .] surely, God loves the .06 nearly as much as the .05" (Rosnow & Rosenthal, 1989, p. 1277).

4.3 Guidelines

When evaluating a set of predetermined informative hypotheses using Bayesian model selection, we recommend the following three-step procedure.

4.3.1 Step 1

In the first step the informative hypotheses have to be formulated. That is, the expected ordering of the parameters needs to be specified. If there are conflicting expectations, multiple informative hypotheses may be specified. The informative hypotheses need to be formulated in terms of (in)equality constraints between parameters in the input file of the software (e.g. $\mu_O < \mu_U$). In addition, in Step 1 the strategy of analysis can be determined, if you want to evaluate all hypotheses at once, or if the best hypothesis is combined with other constraints, and so on. We provide an example of such a strategy in the next section.

4.3.2 Step 2

After running the software, each informative hypothesis under investigation is provided with a BF against the unconstrained hypothesis. That is, no constraints are imposed on any of the parameters of interest, and any ordering is equally likely, If this BF is larger than 1, it can be concluded that there is support from the data in favour of that particular informative hypothesis. If the BF of a certain informative hypothesis versus the unconstrained hypothesis is smaller than 1, it can be concluded that there no support in the data for the informative hypothesis. This procedure should be repeated for all informative hypotheses under investigation. The reason for calculating these BFs, is to enable inspection of the overall model fit of the hypotheses under investigation. With other words, you do not want to perform model selection among only poor hypotheses. Subsequently, the informative hypotheses can be divided into a set of 'supported' hypotheses and a set of 'unsupported' hypotheses.

In our simple example, H_A is the unconstrained hypothesis. The Bayes factor, BF_{CA}, for H_C compared with H_A is 2, indicating that H_C receives support from the data. The BF_{BA} for H_B compared with H_A is .06, which suggests that H_B receives less support from the data than the unconstrained hypothesis H_A. Hence, H_B can be considered as 'unsupported' while the H_C can be considered as 'supported'. After observing that a certain hypothesis is 'unsupported', it may be omitted from the set of hypotheses under investigation. However, if you still want to know how much better an 'unsupported' hypothesis is against other hypotheses, you can maintain the 'unsupported' hypothesis for Step 3.

4.3.3 Step 3

In the third step, all the informative hypotheses of interest are compared with each other (these might include 'unsupported' hypotheses). From these results, it can be concluded how much support there is for each of the informative hypothesis under investigation. There are three options for doing so.

First, when there are two or three informative hypotheses all mutual BFs can be computed. In our simple example, the two informative hypotheses are H_B and H_C. These hypotheses are directly compared with each other by calculating the BF_{CB}. The methodology allows for doing so if we use the BFs against the unconstrained hypothesis by calculating

$$BF_{CB} = \frac{BF_{CA}}{BF_{BA}} = \frac{2}{.06} \approx 31. \tag{4.4}$$

This result implies that H_B receives 31 times as much support from the data as H_C. So, if we were to choose between the informative hypotheses under investigation, hypothesis H_C would win the model selection competition. From this analysis it can be concluded that over-controlled adolescents score lower on externalizing problem behaviour than under-controlled adolescents.

Second, if more informative hypotheses are considered, it is not practical to present the BFs for all possible comparisons. Instead you can provide

the BFs comparing the hypothesis with the largest support of Step 1 against each of the others. H_A receives the highest BF of Step 1 and if there were more hypotheses under investigation the BFs would all be computed using this BF.

Third, an easy way to interpret a set large set of BFs is to convert them into a relative support measure, sometimes referred to as posterior model probabilities (PMPs). Note that the relative support measure is not a real probability, but it can be loosely interpreted as the probability on a 0-1 scale that the hypothesis at hand is the best of a set of finite hypotheses after observing the data. A PMP is computed for each model under consideration and this way an easy comparison of many models can be made. The relative fit of H_C is computed by dividing its BF compared with the unconstrained hypothesis by the sum of all BFs:

$$\frac{2}{(1 + 0.06 + 2)} = .65 \ . \tag{4.5}$$

The relative fit of H_A and H_B are .33 and .02 respectively.

4.4 Psychological Functioning, Personality and Support from Family

4.4.1 Introduction

Van Aken and Dubas (2004) investigated differences between three personality types among adolescents: resilient (R), over-controlled (O), and under-controlled adolescents (U). The main question was whether psychosocial functioning is the result of the interplay between personality and support from family.

The problem behaviour list (De Bruyn, Vermulst & Scholte, 2003) was used to obtain parent reports on adolescent's behavioural problems. Three subscales were used, namely externalizing (E), internalizing (I) and social

Table 4.1: Groups of Adolescents Based on Personality Type, Problem Behaviour and Support

		Problem behaviour		
		Externalizing	Internalizing	Social
Resilient	High family support	μ_{RHE}	μ_{RHI}	μ_{RHS}
	Low family support	μ_{RLE}	μ_{RLI}	μ_{RLS}
Over	High family support	μ_{OHE}	μ_{OHI}	μ_{OHS}
	Low family support	μ_{OLE}	μ_{OLI}	μ_{OLS}
Under	High family support	μ_{UHE}	μ_{UHI}	μ_{UHS}
	Low family support	μ_{ULE}	μ_{ULI}	μ_{ULS}

problem behaviour (S). Personality types (R, O, U) were denoted using big-five personality markers (Gerris et al., 1998). Finally, the relational support inventory (Scholte, Van Lieshout & Van Aken, 2001) was used to measure the support that children receive from their parents to obtain high (H) versus low (L) family support.

Based on personality type (R, O, U), high or low family support (H, L), 3 x 2 = 6 groups were constructed, see Table 4.1. Let μ denote the mean score on the dependent variable, then μ_{RHE} is the mean score for Resilient adolescents with High family support on the dependent variable Externalizing behaviour. To analyze this data we follow the three-step procedure described above.

4.4.2 Step 1

The first two expectations (H_A and H_B) are based on several studies showing that the three personality types have a distinct pattern of psychosocial and relational functioning (Van Aken, Van Lieshout, Scholte & Haselager, 2002). H_A states that under-controllers are expected to have the most externalizing problems and over-controllers are expected to have the most internalizing problems. Over-controllers and under-controllers are believed to score higher on social problems compared with resilient adolescents. Moreover, no constraints are specified with respect to high/low family support. The

informative hypothesis H_A can be formulated as

$$H_A : \begin{cases} (\mu_{RHE}, \mu_{RLE}, \mu_{OHE}, \mu_{OLE}) & < & (\mu_{UHE}, \mu_{ULE}) \\ (\mu_{RHI}, \mu_{RLI}, \mu_{UHI}, \mu_{ULI}) & < & (\mu_{OHI}, \mu_{OLI}) \\ (\mu_{RHS}, \mu_{RLS}) & < & (\mu_{OHS}, \mu_{OLS}, \mu_{UHS}, \mu_{ULS}) \ . \end{cases} \tag{4.6}$$

H_B states, additionally to H_A, that resilient adolescents function best in all psychosocial domains in comparison with the other two types of adolescents. Hence, the informative hypothesis H_B contains two additional constraints in comparison to H_A,

$$H_B : \begin{cases} (\mu_{RHE}, \mu_{RLE}) & < & (\mu_{OHE}, \mu_{OLE}) & < & (\mu_{UHE}, \mu_{ULE}) \\ (\mu_{RHI}, \mu_{RLI}) & < & (\mu_{UHI}, \mu_{ULI}) & < & (\mu_{OHI}, \mu_{OLI}) \\ (\mu_{RHS}, \mu_{RLS}) & < & (\mu_{OHS}, \mu_{OLS}) & , & (\mu_{UHS}, \mu_{ULS}) \ . \end{cases} \tag{4.7}$$

Previous research also indicates that it is the combination of personality type and the quality of social relationships that determines the risk level for experiencing more problem behaviour (Van Aken et al., 2002). Therefore, additional constraints are constructed for the third expectation (H_C). Over- and under-controllers with high perceived support from parents are expected to function better in psychosocial domains than those with low perceived support. For the resilient group, the level of support from parents is not related to problem behaviour. The constraints for informative hypothesis H_C are

$$H_C : \begin{cases} (\mu_{RHE} = \mu_{RLE}) & , & (\mu_{OHE} < \mu_{OLE}) & , & (\mu_{UHE} < \mu_{ULE}) \\ (\mu_{RHI} = \mu_{RLI}) & , & (\mu_{UHI} < \mu_{ULI}) & , & (\mu_{OHI} < \mu_{OLI}) \\ (\mu_{RHS} = \mu_{RLS}) & , & (\mu_{OHS} < \mu_{OLS}) & , & (\mu_{UHS} < \mu_{ULS}) \ . \end{cases} \tag{4.8}$$

The strategy of analysis is first to determine which hypothesis, H_A or H_B, receives the most support from the data. The best of these two hypotheses is then combined with the constraints of H_C to investigate whether these additional constraints are supported by the data. After having specified these hypotheses, we ran the software described in Mulder, Klugkist et al. (in press).

Table 4.2: Results of Bayesian model selection for the example of Van Aken and Dubas (2004)

Expectation	BF*	BF**	BF***
H_A	30.28	1	-
H_B	64.20	2.12	1
H_{BC}	1399.00	-	21.79

*BF compared with the unconstrained hypothesis

** BF between H_A and H_B

*** BF between H_B and H_{BC}

4.4.3 Step 2

The second step involves comparing H_A, H_B, and H_C with the unconstrained hypothesis, H_U. The results, see the second column of Table 4.2, show that all informative hypotheses have a BF larger than 1 versus H_U. For example, the BF between H_A and H_U is 30.28, indicating that H_A receives 30.28 times more support than H_U. From these BFs, it can be concluded that each of the hypotheses H_A, H_B, and H_C receives support from the data and have a good model fit.

4.4.4 Step 3

The second step involves comparing informative hypotheses with BFs. We first want to compare H_A with H_B to decide whether, additional to the constraints of H_A, resilient adolescents function best in all psychosocial domains. The BF of H_B against H_A is given by

$$BF_{BA} = \frac{BF_{BU}}{BF_{AU}} = \frac{64.20}{30.28} = 2.12 \ . \tag{4.9}$$

The support for H_B is about twice as strong as for H_A. From this analysis it can be concluded that additional to the constraints of H_A, there is also evidence that resilient adolescents score lower on externalizing behaviour than over-controlled adolescents and that resilient adolescents score lower on internalizing behaviour than under-controlled adolescents as was assumed by expectation H_B.

Secondly, we were interested to see whether the additional constraints of H_C presented in Equation (4.8) are supported by the data. Consequently, the additional constraints of H_C are combined with the constraints of H_B, leading to the informative hypothesis H_{BC},

$$H_{BC}: \begin{cases} (\mu_{RHE} = \mu_{RLE}) & < & (\mu_{OHE} < \mu_{OLE}) & < & (\mu_{UHE} < \mu_{ULE}) \\ (\mu_{RHI} = \mu_{RLI}) & < & (\mu_{UHI} < \mu_{ULI}) & < & (\mu_{OHI} < \mu_{OLI}) \\ (\mu_{RHS} = \mu_{RLS}) & < & (\mu_{OHS} < \mu_{OLS}) & , & (\mu_{UHS} < \mu_{ULS}) \end{cases} \tag{4.10}$$

We calculated the BF of H_{BC} versus H_B (see the fourth column in Table 4.2). These BFs show that there is much support in favour of HB_C compared with H_A or H_B. For example, the BF for H_{BC} against H_B is 21.79; in other words there is approximately 21 times as much support for H_{BC} as for H_B. From this analysis it can be concluded that the additional constraints of H_C are a meaningful addition to the constraints of H_B.

4.4.5 Conclusion

The results of Bayesian model selection for the example relating to personality types and problem behaviour provides strong support for the idea that it is the combination of personality type and the quality of social relationships that puts adolescents at risk of greater problem behaviour. Note that this is the same conclusion as in Van Aken and Dubas (2004), but now we learned even more about the data: how much support there is for each of the expectations. That is, there is approximately 21 times as much support for the additional expectations of H_C compared with the constraints of H_B alone.

Table 4.3: Means for the example of Van Aken and Dubas (2004)

		Problem behaviour		
		Externalizing	Internalizing	Social
Resilient	High family support ($n = 135$)	1.50	1.88	1.69
	Low family support ($n = 70$)	1.64	1.94	1.80
Over	High family support ($n = 76$)	1.43	2.05	1.77
	Low family support ($n = 81$)	1.58	2.18	1.94
Under	High family support ($n = 70$)	1.52	2.04	1.81
	Low family support ($n = 131$)	1.68	2.13	1.95

An examination of the means for all groups (see Table 4.3) shows that the constraints of H_{BC} are mostly supported by the data, but not perfectly. For example, in the over-controlled group with high family support social problem behaviour is lower than in the low family support group (1.77 vs. 1.94, respectively). In the resilient group high and low family support groups should have had the same levels of problem behaviour, but although the differences are small, this is not the case. The constraints imposed by H_{BC} on the means fit well enough for this hypothesis to win the model selection competition.

4.5 Discussion

In this paper we have shown that Bayesian model selection is a useful tool when evaluating informative hypotheses. The resulting Bayes factor quantifies the amount of support received from the data for each informative hypothesis. In the current paper we offer an introduction to the methodology for non-statisticians and we are the first to present a step-by-step approach to analyzing informative hypotheses with Bayesian model selection.

The major benefit of the method of evaluating informative hypotheses with Bayesian model selection is that: (i) there is a direct answer to the research question: how much better is one hypothesis versus another hypothesis; (ii) the amount of support for each hypothesis is quantified for

each hypothesis versus the unconstraint hypothesis to obtain overall model fit and for all informative hypotheses under investigation. The first step of the methodology described in this paper is to specify the constraints between the parameters of interest and hence to determine the admissible parameter space. Within the admissible parameter space a prior distribution needs to be specified. The methodology of evaluating informative hypotheses using Bayesian model selection employs a so called encompassing prior approach (Klugkist et al., 2005). The actual specification of this encompassing prior distribution is not considered to be the topic of this paper and the interested reader is referred to Mulder, Hoijtink and Klugkist (2009), and Mulder, Klugkist et al. (2009) for a detailed description of a default specification of the prior distribution used in the software we used in the current article.

Some research has been done on testing informative hypotheses in the null hypothesis framework, see the standard works of Barlow et al. (1972); Robertson et al. (1988); Silvapulle and Sen (2004), or see Van de Schoot, Hoijtink and Deković (2010) for testing informative hypotheses in structural equation models. However, these methods compare either a null hypothesis or an unconstrained hypothesis with one single informative hypothesis. These methods cannot deal with evaluating two or more informative hypotheses at the same time.

In conclusion, if researchers in psychology want to learn as much as possible from their data and if they want to judge the plausibility of expectations, Bayesian model selection, described in the current paper, is a promising and exciting tool.

5

Testing Inequality Constrained Hypotheses in SEM Models

Van de Schoot, R., Hoijtink, H., & Deković, M.

Published, 2010 in *Structural Equation Modeling, 17, 443-463*

Abstract

Researchers often have expectations that can be expressed in the form of inequality constraints among the parameters of a Structural Equation Model (SEM). It is currently not possible to test these so-called informative hypotheses in SEM software. We offer a solution to this problem using Mplus. The hypotheses are evaluated using plug-in p-values with a calibrated alpha level. The method is introduced and its utility is illustrated by means of an example.

Order restricted inference has been studied in the frequentist framework (see for example the books of: Barlow et al., 1972; Robertson et al., 1988; Silvapulle & Sen, 2004) as well as in the Bayesian framework (e.g. Hoijtink, Klugkist & Boelen, 2008; Klugkist et al., 2005). However, testing order constraints has received relatively little attention in the Structural Equation Modeling (SEM) literature (Gonzalez & Griffin, 2001; Stoel, Galindo-Garre, Dolan & Van den Wittenboer, 2006). SEM is often used and its attractiveness is largely due to its flexibility in specifying and testing hypotheses among both observed and latent variables in multiple groups.

SEM software can be used to impose inequality constraints among the parameters of interest. More specifically, in order to evaluate a research question, model parameters such as regression coefficients can be constrained to being greater or smaller than either a fixed value or other regression coefficients. We call hypotheses that contain inequality constraints *informative* hypotheses. Mplus (Muthén & Muthén, 2007) allows for such user-specified constraints and order constraint parameter estimation is available. The problem is that a null hypothesis test for the evaluation of an informative hypothesis is lacking in SEM software.

We offer a solution to this problem based on the parametric bootstrap method available in Mplus. Plug-in p-values are obtained using a likelihood ratio test. The performance of these p-values is evaluated and we will show that the alpha level should be calibrated. Using some examples, we demonstrate how this can be done.

5.1 Constraint Parameter Estimation and Hypothesis Testing

Ritov and Gilula (1993) proposed to obtain maximum likelihood estimates of order-restricted models by a pooling adjacent violators algorithm (see

also, Robertson et al., 1988, p.56). The procedure in Mplus (Muthén & Muthén, 2007), the slacking parameter method, is based on the solution of Ritov and Gilula. The core of the algorithm is estimating the parameters via the maximum likelihood method such that the likelihood is maximized using the sequential quadratic programming method (Han, 1977). In this method the parameters that contain inequality constraints are updated in an iterative process where inequality constraints are treated as equality constraints whenever the estimates do not fit the constraints imposed on the parameters. This is done by the introduction of 'slack' parameters into the model, see Schoenberg (1997) for more details.

How to test inequality constraint hypotheses has mainly been studied outside the SEM model, see the books of Barlow et al. (1972), Robertson et al. (1988), and Silvapulle and Sen (2004) for a comprehensive overview. Besides, in 2002 the *Journal of Statistical Planning and Inference* published a special issue on testing inequality constraint hypotheses (V. W. Berger & Ivanova, 2002; Chongcharoen, Singh & Wright, 2002; Khalil, Saikali & Berger, 2002; C. C. Lee & Yan, 2002; Perlman & Wu, 2002a, 2002b; Sampson & Singh, 2002; Silvapulle et al., 2002; Sen & Silvapulle, 2002), but none of these articles discussed constraints in SEM models. Testing informative hypotheses for SEM models has been described by Stoel et al. (2006). In this study, constraints were imposed on variance terms to obtain only positive values. Hypotheses tests were performed to test the benefit of these constraints (see also, Gonzalez & Griffin, 2001). Also, Tsonaka and Moustaki (2007) described testing parameter constraints in SEM models. In specific they described factor analysis where a parametric bootstrap was performed to obtain the results. However, they only considered a comparison between a constrained and an unconstrained model. In the present study we will also focus on constraint hypothesis testing within the SEM model, and although we will present examples of a path model, our solution is not limited to these kind of models. We will also show that the alpha values used

in constraint hypothesis testing need to be calibrated, which is not done in the studies described above.

In almost all of the books and papers described above, the likelihood ratio test (LRT) is used to test the inequality constraint hypothesis at hand. The null distribution of this test is a chi-square distribution with degrees of freedom equal to the difference between the number of parameters of the models under comparison (Bollen, 1989). An important result from the work of Barlow et al. (1972), Robertson et al. (1988), and Silvapulle and Sen (2004) is that one of the regularity conditions of the LRT does not hold when testing inequality constraint hypotheses (see also, Andrews, 1996, 2000; Ritov & Gilula, 1993; Stoel et al., 2006). Consequently, the asymptotic distribution of the LRT is no chi-square distribution and its p-value can not straightforwardly be computed.

Moreover, model selection criteria, such as the AIC or BIC, can not be used to distinguish between statistical models with inequality constraints between the parameters of interest. These criteria use the likelihood evaluated in its maximum as a measure of model fit, and the number of parameters of the model as a measure of complexity. The problem is that model selection criteria can not distinguish between hypotheses when these hypotheses do not differ in model fit but they only differ in the number of constraints imposed on the parameters of interest.

For example, consider the hypothesis $\{\theta_1 - \theta_2\} < \theta_3 > 0$ where $\theta_1 \dots \theta_3$ denote for example mean scores on some variable. Furthermore, suppose we want to compare this informative hypothesis to an unconstrained hypothesis where the parameters are allowed to have any value. Suppose the unconstraint parameter estimates fit the constraints, so that the estimated parameters agree with $\{\theta_1 - \theta_2\} < \theta_3 > 0$. In this case, both the constraint and unconstraint hypotheses do not differ in model fit, i.e. the maximum of the likelihood is the same for both models. The problem is how to account for model complexity. Because the parameters are restricted, the number of parameters used to determine model complexity is clearly not equal to 3. So

far, quantifying the number of parameters for constraint hypotheses received hardly any attention in the literature.

In conclusion, evaluating informative hypotheses in SEM models is neither possible with the likelihood ratio test, nor with traditional model selection criteria. Constraint parameter estimation and informative hypothesis testing has extensively been studied, but literature for SEM models is sparse. Also, as we will show in this paper, calibration of the alpha level is essential when testing inequality constraint hypotheses. This received hardly any attention in the literature described before. We will show how informative hypotheses can be tested in SEM models using Mplus, but we first introduce an example in which the hypothesis of interest are informative.

5.2 Ethnicity and Antisocial behaviour

The problem of testing inequality constrained hypotheses in SEM and its solution is illustrated using the following example. Deković et al. (2004) investigated whether the leading theories about antisocial behaviour in the dominant culture of adolescents can be generalized to members of different ethnic groups. For this example the dominant culture is the Dutch culture which is compared to the Moroccan, Turkish and Surinamese cultures in the Netherlands. Three aspects of the parent adolescent relationship were assessed: positive quality of the relationship (affection and intimacy), negative quality of the relationship (antagonism and conflict) and disclosure (how much adolescents tell the parents). The sample consists of 603 adolescents (mean age 14.4, range 14-16 years), 68% of the adolescents are Dutch ($n = 407$), 11% are Moroccan ($n = 68$), 13% are Turkish ($n = 79$) and 8% are Surinamese ($n = 49$). Adolescents were classified into these ethnic categories according to their responses on a single item in the questionnaire: "What ethnical group best describes you?" Using these data we present three examples where the hypothesis under investigation is informative.

The structural equation model used here is given by

$$\boldsymbol{y}_i^g = \boldsymbol{B}^g \boldsymbol{y}_i^g + \boldsymbol{\Gamma}^g \boldsymbol{x}_i^g + \boldsymbol{\zeta}_i^g \qquad \text{with } \boldsymbol{x}_i \sim N(\boldsymbol{\mu}_x^g, \boldsymbol{\Phi}^g) \tag{5.1}$$

where, if q is the number of dependent variables, r is the number of independent variables, $g = 1, \ldots, G$ denotes group membership and $i = 1, \ldots, I$ denote persons, then $\boldsymbol{y}_i^g$ is a $q \times 1$ vector of dependent variables for person i within group g, $\boldsymbol{B}^g$ is a $q \times q$ matrix of regression coefficients between y's where the diagonal must consist of zeros, $\boldsymbol{x}_i$ is a $r \times 1$ vector of independent variables, $\boldsymbol{\mu}_x^g$ is a $r \times 1$ vector of means for each independent variable with covariance matrix $\boldsymbol{\Phi}^g$, $\boldsymbol{\Gamma}^g$ is $q \times r$ matrix of regression coefficients between y's and x's, $\boldsymbol{\zeta}_i^g$ is $q \times 1$ vector with error terms which is assumed to have a multivariate-normal distribution, $\boldsymbol{\zeta}_i^g \sim N(0, \boldsymbol{\Psi}^g)$, which is independent of $\boldsymbol{y}$ and $\boldsymbol{x}$. Under these assumptions, the observed $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ have a multivariate-normal distribution with

$$\begin{bmatrix} \boldsymbol{y}_i^g \\ \boldsymbol{x}_i^g \end{bmatrix} \sim N_{q+r} \left(\begin{matrix} \boldsymbol{\mu}_y^g \\ \boldsymbol{\mu}_x^g \end{matrix} , \Sigma^g \right) \tag{5.2}$$

where $\boldsymbol{\Sigma}$ represents the implied covariance matrix which is given by

$$\Sigma^g = \begin{bmatrix} \underset{(q \times q)}{\Sigma_{yy}^g} & \underset{(r \times q)}{\Sigma_{xy}^g} \\ \underset{(q \times r)}{\Sigma_{yx}^g} & \underset{(r \times r)}{\Sigma_{xx}^g} \end{bmatrix} \tag{5.3}$$

with $\mathbf{I}$ being the identity matrix and

$$\begin{aligned} \Sigma_{xx}^g &= \boldsymbol{\Phi}^g \\ \Sigma_{yy'}^g &= (\mathbf{I} - \mathbf{B}^g)^{-1} (\boldsymbol{\Gamma}^g \boldsymbol{\Phi}^g \boldsymbol{\Gamma}'^g + \boldsymbol{\Psi}^g) (\mathbf{I} - \mathbf{B}^g)^{-1'} \\ \Sigma_{xy}^g &= \boldsymbol{\Phi}^g \boldsymbol{\Gamma}^g (\mathbf{I} - \mathbf{B}^g)^{-1'} \ . \end{aligned} \tag{5.4}$$

Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2\}$ with $\boldsymbol{\theta}^1 = \{\boldsymbol{B}^1, \ldots, \boldsymbol{B}^G, \boldsymbol{\Gamma}^1, \ldots, \boldsymbol{\Gamma}^G\}$ and $\boldsymbol{\theta}^2 = \{\boldsymbol{\Phi}^1, \ldots, \boldsymbol{\Phi}^G, \boldsymbol{\Psi}^1, \ldots, \boldsymbol{\Psi}^G\}$. Then, the likelihood function can be given by

$$\log f(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta}) \ = \sum_{g=1}^{G} \left(\tfrac{\mathrm{N}^g}{\mathrm{N}}\right) \mathrm{F}_{ML}^g \left[\mathbf{S}^g, \Sigma^g\right] \tag{5.5}$$

where N^g is the sample size for group g, $\mathbf{S}^g$ is the sample covariance matrix among the observed variables in group g, $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$, $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ and where F^g_{ML} is given by

$$\mathrm{F}^g_{ML} = \log \mid \Sigma^g \mid + \mathrm{tr}\Big[\mathbf{S}^g \Sigma^{g^{-1}} - \log \mid \mathbf{S}^g \mid -(\mathrm{q} + \mathrm{r})\Big] \tag{5.6}$$

We consider two types of hypothesis tests (Silvapulle & Sen, 2004), type A is of the form

$$\begin{aligned} \mathrm{H}_0 : &\quad \mathbf{A}\boldsymbol{\theta}^1 = \boldsymbol{c} \\ \mathrm{H}_1 : &\quad \mathbf{A}\boldsymbol{\theta}^1 > \boldsymbol{c}, \end{aligned} \tag{5.7}$$

and type B is of the form

$$\begin{aligned} \mathrm{H}_0 : &\quad \mathbf{A}\boldsymbol{\theta}^1 > \boldsymbol{c} \\ \mathrm{H}_1 : &\quad \text{unconstrained}, \end{aligned} \tag{5.8}$$

where the unconstrained model refers to a model without any constraints imposed on the parameters and where, if m is the number of inequality constraints imposed on the model and k the number of parameters involved, $\mathbf{A}$ is an $m \times k$ matrix of known constants, and $\mathbf{c}$ an $m \times 1$ vector of known constants. More specific examples will be given in the sequel.

5.2.1 Example 1: Simple Regression

The first example is a simple regression model where levels of antisocial behaviour are regressed on either a negative or positive relation with the parent and adolescent disclosure, see Figure 5.1, where $\boldsymbol{B} = \varnothing$ and

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 \end{bmatrix}, \tag{5.9}$$

$$\Psi = \begin{bmatrix} \psi_1 \end{bmatrix}, \tag{5.10}$$

$$\Phi = \begin{bmatrix} \phi_{11} & & \\ \phi_{12} & \phi_{22} & \\ \phi_{13} & \phi_{23} & \phi_{33} \end{bmatrix}. \tag{5.11}$$

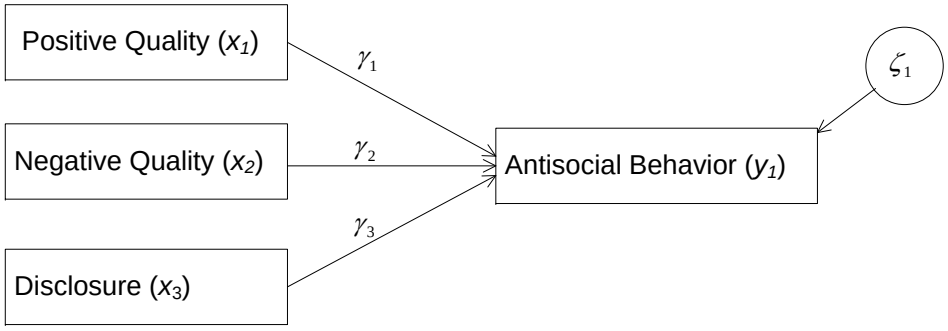


Figure 5.1: Path model between relationship characteristics, disclosure, and prevalence of antisocial behaviour

Note that this model is based on the total sample, therefore the superscript g is not needed.

Deković et al. (2004) state that adolescent disclosure is the strongest predictor of antisocial behaviour, followed by either a negative or positive relation with the parent (see also: Dishion & McMahon, 1998). We therefore hypothesize that the regression coefficients γ_1 and γ_2 are smaller than γ_3. Using

$$\boldsymbol{A} = \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix} , \tag{5.12}$$

$$\boldsymbol{\theta}^1 = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} , \tag{5.13}$$

$$\boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} , \tag{5.14}$$

the hypotheses tested for this example with $m = 2$, are for type A

$$H_0 : \quad \begin{bmatrix} \gamma_3 - \gamma_1 = 0 \\ \gamma_3 - \gamma_2 = 0 \end{bmatrix}$$

versus (5.15)

$$H_1 : \quad \begin{bmatrix} \gamma_3 - \gamma_1 > 0 \\ \gamma_3 - \gamma_2 > 0 \end{bmatrix} ,$$

and for type B

$$H_0 : \quad \begin{bmatrix} \gamma_3 - \gamma_1 > 0 \\ \gamma_3 - \gamma_2 > 0 \end{bmatrix}$$

versus (5.16)

$$H_1 : \quad \begin{bmatrix} \gamma_3 \ , \ \gamma_1 \\ \gamma_3 \ , \ \gamma_2 \end{bmatrix} .$$

Note that H_1 in (16) can also be written as $\{\gamma_1 \ , \ \gamma_2\} < \gamma_3$. This type of notation will be used in the remainder of this paper. We used M-plus version 5 (Muthén & Muthén, 2007) to estimate the unconstrained and constrained regression coefficients, see Table 5.1. As can be seen in this table, the unconstrained estimate of γ_2 is not smaller than γ_3. Consequently, the constrained estimates of γ_2 and γ_3 are set equal by the introduction of a 'slack' parameter, see the lower panel of Table 5.1.

5.2.2 Example 2: Multi Group Analysis

Research about the nature and impact of antisocial behaviour is dominated by studies conducted with white, western, middle class adolescents (Deković et al., 2004). It could be questioned whether the model in Figure 5.1 is the same for different ethnic groups living in The Netherlands: Dutch (indicated by $g = 1$), Turkish ($g = 2$), Moroccan ($g = 3$), and Surinamese adolescents

Table 5.1: Regression Coefficients for Example 1

Coefficient	B	SE
Unconstrained		
γ_1	.06	.02
γ_2	.24	.03
γ_3	.23	.03
Constrained		
γ_1	.06	.02
γ_2	.235	.03
γ_3	.235	.03

$(g = 4)$, where $\boldsymbol{B} = \varnothing$ and

$$\begin{aligned} \boldsymbol{\Gamma}^1 &= \begin{bmatrix} \gamma_1^1 & \gamma_2^1 & \gamma_3^1 \end{bmatrix} \\ &\vdots \\ \boldsymbol{\Gamma}^4 &= \begin{bmatrix} \gamma_1^4 & \gamma_2^4 & \gamma_3^4 \end{bmatrix}, \end{aligned} \tag{5.17}$$

$$\begin{aligned} \Psi^1 &= \begin{bmatrix} \psi_1^1 \end{bmatrix} \\ &\vdots \\ \Psi^4 &= \begin{bmatrix} \psi_1^4 \end{bmatrix}, \end{aligned} \tag{5.18}$$

$$\begin{aligned} \Phi^1 &= \begin{bmatrix} \phi_{11}^1 & & \\ \phi_{12}^1 & \phi_{22}^1 & \\ \phi_{13}^1 & \phi_{23}^1 & \phi_{33}^1 \end{bmatrix} \\ &\vdots \\ \Phi^4 &= \begin{bmatrix} \phi_{11}^4 & & \\ \phi_{12}^4 & \phi_{22}^4 & \\ \phi_{13}^4 & \phi_{23}^4 & \phi_{33}^4 \end{bmatrix}. \end{aligned} \tag{5.19}$$

The null hypothesis is that the regression coefficients for the predictors of antisocial behaviour are the same for all ethnic groups (see for example Greenberger & Chen, 1996):

$$H_0 : \begin{bmatrix} \gamma_1^1 = \gamma_1^2 = \gamma_1^3 = \gamma_1^4 \\ \gamma_2^1 = \gamma_2^2 = \gamma_2^3 = \gamma_2^4 \\ \gamma_3^1 = \gamma_3^2 = \gamma_3^3 = \gamma_3^4 \, . \end{bmatrix} \tag{5.20}$$

According to Deković et al. (2004) there are also indications in the literature that the same risk factors have different effects in different ethnic groups, a so-called process times context interaction phenomenon. The authors expected that cross-ethnic variations result in weaker relations between parent-child relations and adolescent behaviour compared to Dutch families. This was mainly expected because of differences in family expectations (Phalet & Schönpflung, 2001) and differences in intergenerational conflicts due to migration (Deković, Noom & Meeus, 1997). The informative hypothesis H_1 is:

$$H_1 : \quad \begin{bmatrix} \gamma_1^1 > \{\gamma_1^2, \gamma_1^3, \gamma_1^4\} \\ \gamma_2^1 > \{\gamma_2^2, \gamma_2^3, \gamma_2^4\} \\ \gamma_3^1 > \{\gamma_3^2, \gamma_3^3, \gamma_3^4\} \end{bmatrix} \quad . \tag{5.21}$$

The hypotheses that are tested for this example are for type A: (5.20) versus (5.21); and for type B: (5.21) versus the unconstrained model. In Table 5.2 the unconstrained and constrained regression coefficients are shown. As can be seen, not all unconstraint regression coefficients are in agreement with the constraints of H_1 in (21). For example γ_2 for the Dutch adolescents is smaller instead of higher than γ_2 for Moroccan adolescents. The bottom of Table 5.2 renders parameter estimates, obtained with the parameter slack method, that are in agreement with the constraints.

5.2.3 Example 3: Path Model

The third example includes the variable hanging around with deviant peers. The hypothesis states that problem behaviour is not only directly predicted by disclosure and a negative or positive relation with the parent, but is also indirectly predicted via hanging around with deviant peers, see Figure 5.2, where

$$\boldsymbol{B} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} , \tag{5.22}$$

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \end{bmatrix} , \tag{5.23}$$

Table 5.2: Regression Coefficients for Dutch, Moroccan, Turkish and Surinamese adolescents for Example 2 (Standard error between brackets)

	Ethnicity			
Coefficient	Dutch	Moroccan	Turkish	Surinamese
		Unconstrained		
γ_1	.05 (.03)	.12 (.09)	.04 (.08)	.15 (.09)
γ_2	.28 (.03)	.23 (.11)	.16 (.08)	.08 (.11)
γ_3	.20 (.04)	.33 (.13)	.25 (.10)	.24 (.14)
		Constrained		
γ_1	.06 (.02)	.06 (.02)	.03 (.07)	.06 (.02)
γ_2	.28 (.03)	.28 (.03)	.16 (.08)	.08 (.11)
γ_3	.22 (.03)	.22 (.03)	.22 (.03)	.16 (.12)

$$\Psi = \begin{bmatrix} \psi_{11} & \\ 0 & \psi_{22} \end{bmatrix}, \tag{5.24}$$

$$\Phi = \begin{bmatrix} \phi_{11} & & \\ \phi_{12} & \phi_{22} & \\ \phi_{13} & \phi_{23} & \phi_{33} \end{bmatrix}. \tag{5.25}$$

Note that this model is based on the total sample, therefore superscript g is not needed.

As is argued by Deković et al. (2004) children spend, especially in adolescence, more and more time with their peers without adult supervision (see also Mounts & Steinberg, 1995). During this period peers become the most important reference group for adolescents. Deković et al. (2004) state that especially in this period the association with deviant peers has emerged as the most prominent predictor of problem behaviour. The hypotheses tested for Example 3 are hypothesis type A:

$$\begin{aligned} &H_0: \quad \beta_{21} = \gamma_{21} = \gamma_{22} = \gamma_{23} \\ &\text{versus} \\ &H_1: \quad \beta_{21} > \{\gamma_{21}\ ,\ \gamma_{22}\ ,\ \gamma_{23}\} \end{aligned} \tag{5.26}$$

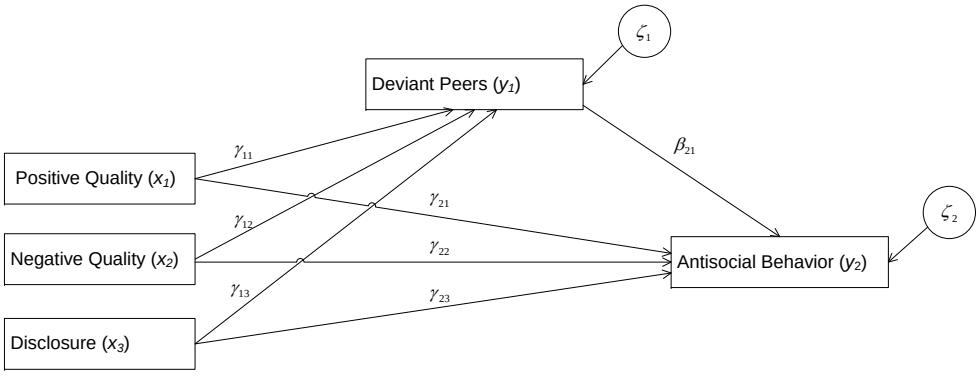


Figure 5.2: Path model between relationship characteristics, hanging around with deviant peers, and prevalence of antisocial behaviour

and hypothesis type B

$$\begin{aligned} H_0: &\quad \beta_{21} > \{\gamma_{21}\ ,\ \gamma_{22}\ ,\ \gamma_{23}\} \\ \text{versus} & \\ H_1: &\quad \beta_{21}\ ,\ \gamma_{21}\ ,\ \gamma_{22}\ ,\ \gamma_{23}\ . \end{aligned} \tag{5.27}$$

The unconstrained and constrained regression coefficients are shown in Table 5.3. As can be seen in this table, the constrained coefficients do not differ from their unconstrained counterparts. Hence, the constraints imposed by the informative hypothesis are not contradicted by the data.

5.3 Parametric Bootstrap

To evaluate informative hypotheses like presented in the previous section we make use of the parametric bootstrap. Bootstrapping is an approach for statistical inference falling within a broader class of resampling methods (Efron & Tibshirani, 1993). Various authors have suggested using the parametric bootstrap when the parameter space is restricted (Galindo-Garre & Vermunt, 2004, 2005; Ritov & Gilula, 1993; Stoel et al., 2006; Tsonaka & Moustaki, 2007).

Table 5.3: Regression Coefficients for Example 3

Coefficient	B	SE
	Unconstrained	
γ_{11}	.05	.03
γ_{12}	.31	.03
γ_{13}	.22	.04
β_{21}	.55	.02
	constrained	
γ_{11}	.05	.03
γ_{12}	.31	.03
γ_{13}	.22	.04
β_{21}	.55	.02

5.3.1 Bootstrap method

The method we advocate starts with the observed data $z = \{\boldsymbol{y}, \boldsymbol{x}\}$ and the likelihood in Equation (5.5) (see Start in Figure 5.3). Step 1 is a parametric bootstrap from a population in which the null hypothesis is true. First, θ is estimated under H_0 using the data z resulting in

$$f(z|\hat{\theta}_{H_0|z}) \ . \tag{5.28}$$

Using (5.28), T bootstrap samples of size n are generated, resulting in data sets z_t^{rep}, for $t = 1, \ldots, T$, see Figure 5.3.
Then, θ is estimated for each replicated data set under H_0, rendering

$$f(z_1^{rep}|\hat{\theta}_{H_0|z_1^{rep}}) \ldots f(z_T^{rep}|\hat{\theta}_{H_0|z_T^{rep}}) \ . \tag{5.29}$$

Further, θ is estimated under H_1, rendering

$$f(z_1^{rep}|\hat{\theta}_{H_1|z_1^{rep}}) \ldots f(z_T^{rep}|\hat{\theta}_{H_1|z_T^{rep}}) \ . \tag{5.30}$$

The second step, denoted by Step 2 in Figure 5.3, is to repeat these computations conditional on the observed data set and to compute $f(z|\hat{\theta}_{H_0|z})$ and $f(z|\hat{\theta}_{H_1|z})$.

The final step, see the lower part of the top panel of Figure 5.3, is to choose a test statistic, denoted by Λ, to investigate the compatibility of the null hypothesis with the observed data. Like many previous studies (e.g. Barlow et al., 1972; Robertson et al., 1988; Silvapulle & Sen, 2004), we will also use the LRT for evaluating the hypotheses at hand, but, as illustrated before, we do not use a p-value based on a chi-square distribution.

Since $f(z|\theta)$ is proportional to the likelihood, an LRT is performed for each replicated data set rendering

$$\Lambda_t = -2\log\left\{\frac{f(z_t^{rep}|\hat{\theta}_{H_0|z_t^{rep}})}{f(z_t^{rep}|\hat{\theta}_{H_1|z_t^{rep}})}\right\} \tag{5.31}$$

and for the observed data set it renders

$$\Lambda = -2\log\left\{\frac{f(z|\hat{\theta}_{H_0|z})}{f(z|\hat{\theta}_{H_1|z})}\right\} \ . \tag{5.32}$$

Now, a p-value can be computed using

$$p = \ \mathrm{P}(\Lambda_t > \Lambda \ |H_0, z) \ . \tag{5.33}$$

It can be approximated by the proportion of LRT-values from the simulated data sets that are equal or larger than the LRT-value of the observed data set, resulting in the definition of the plug-in p-value

$$p \approx \frac{\sum_{t=1}^{\mathrm{T}} I_t}{\mathrm{T}} \ , \tag{5.34}$$

where I_t is an indicator function taking the value 1 if the inequality holds and 0 otherwise:

$$I_t = \begin{cases} 1 & \text{if } \Lambda_t > \Lambda \\ 0 & \text{otherwise .} \end{cases} \tag{5.35}$$

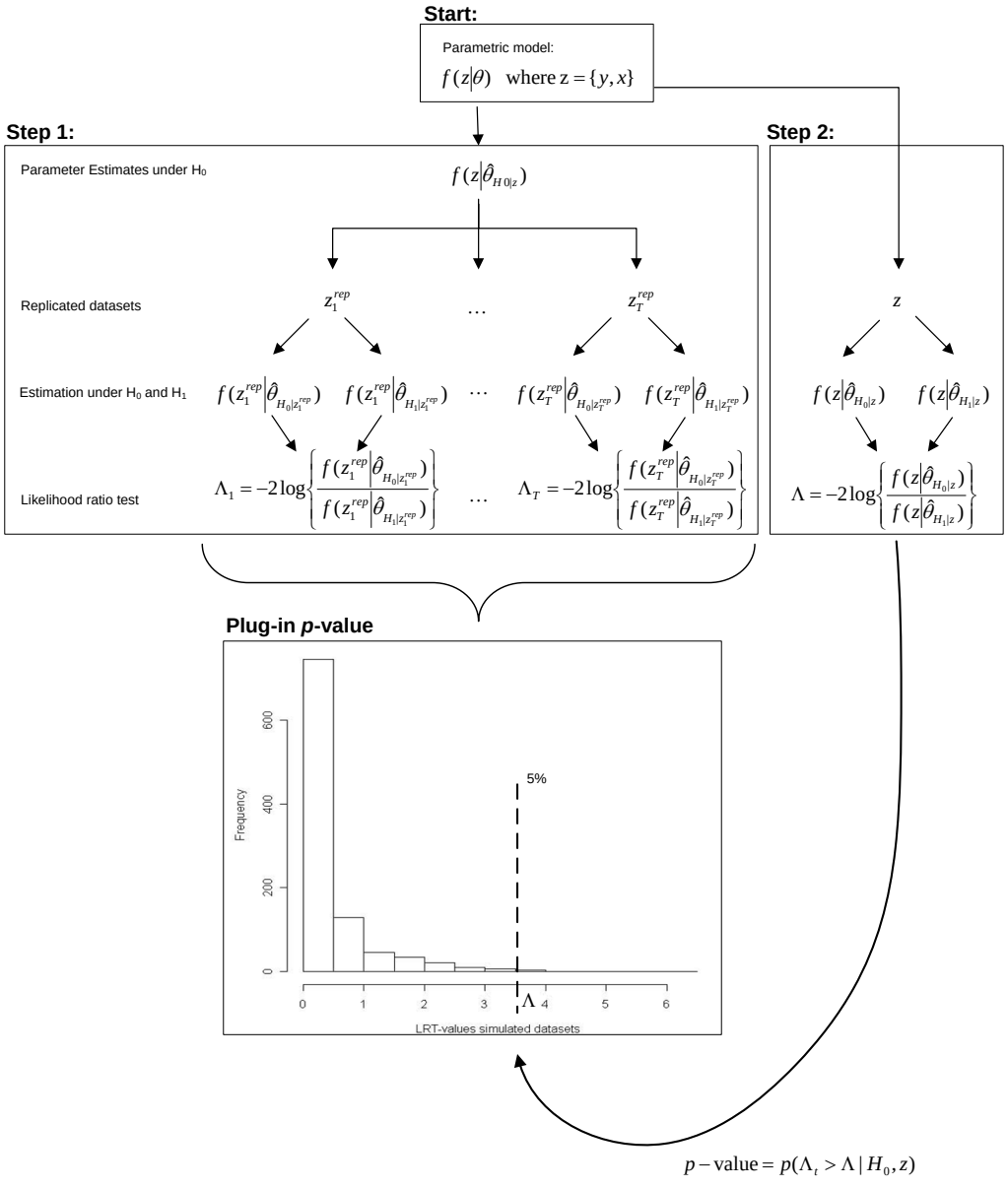


Figure 5.3: Graphical representation of the parametric bootstrap method

A hypothetical distribution is illustrated in the graph in the lower part of Figure 5.3. To determine whether the Λ-value from the observed data set stems from a population where the null hypothesis is true, it has to be smaller than a chosen alpha value. Traditional an alpha value of .05 is used and as such the observed Λ-value has to lie on the left hand side of the dotted line in Figure 5.3. However, as we will show in the next section in many situations the alpha value needs to be calibrated.

The procedure described above can be conducted for type A and B hypothesis testing. The parameter estimates and likelihood values can be obtained using Mplus and we developed R code that automatically computes the LRT values and the plug-in p-value from the output files of Mplus. Input files for all examples and R code can be downloaded from $http://www.fss.uu.nl/ms/schoot$.

5.4 Frequency Properties of the Asymptotic P-values

In the previous section we showed how to obtain plug-in p-values for the evaluation of informative hypotheses. An appealing property for any p-value, and consequently for our plug-in p-value, is, considered as a random variable, to be asymptotically uniform [0,1] under the null hypothesis: $\mathrm{P}(p < \alpha|H_0) = \alpha$. However, in some situations exact uniformity of p-values cannot be attained (Andrews, 2000; Bayarri & Berger, 2000; Galindo-Garre & Vermunt, 2004, 2005; Stoel et al., 2006). Andrews (2000) for example, showed that the results of the bootstrap procedure are not coherent when inequality constraints are imposed on the model parameters. Furthermore, Galindo-Garre and Vermunt (2004) showed in a simulation study that the parametric bootstrap may produce p-values that are higher than expected.

The bootstrap procedure described in the previous section may also lead to p-values that are biased. To determine whether the actual alpha level differs from its nominal level a double bootstrap procedure is used

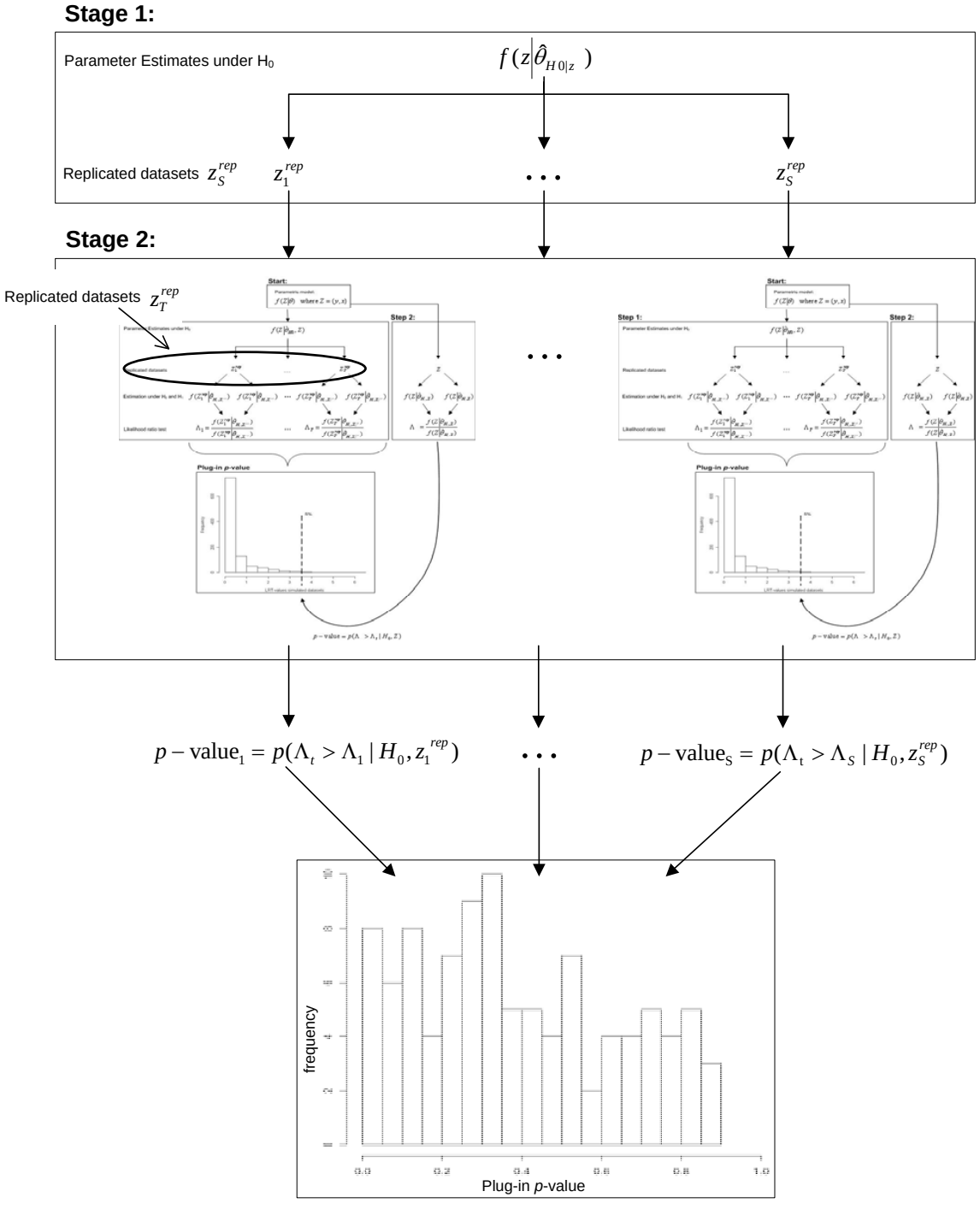


Figure 5.4: Graphical representation of the double bootstrap method

(Efron & Tibshirani, 1993). This procedure renders a calibrated alpha level: $\mathrm{P}(p < \alpha^* | H_0) = \alpha$, where α^* denotes the calibrated alpha level.

As can be seen in Figure 5.4, in the double bootstrap procedure there are two stages of bootstrapping. In Stage 1, data sets are generated using $f(z|\hat{\theta}_{H_0|z})$. Note that we make an important assumption here. We implicitly assume that in our procedure $\hat{\theta}_{H_0|z}$ is a good approximation of the true population values θ_{H_0} which are unknown. We assume that with n sufficiently large, the true population values will be close to the estimated values $f(z|\hat{\theta}_{H_0|z})$.

The result of Stage 1, are S data sets (for $s = 1, \ldots, S$) and the double bootstrap algorithm amounts to treating each bootstrap sample z_s^{rep} like an original data set in the second stage of the double bootstrap procedure, see Stage 2 in Figure 5.4. For each first stage data set, z_s^{rep}, a plug-in p-value can be computed based on the procedure described in the previous section.

In total, S plug-in p-values are computed and a hypothetical distribution of these values is shown in the lower part of Figure 5.4. As can be seen, it does not have an uniform distribution, but is skewed. In Figure 5.4, the 5^{th} percentile of generated plug-in p-values has a plug-in p-value of .02. That is, 5% of the p-values is smaller than .02, and for example 11% of the p-values is smaller than .05. In such a case, the alpha level for evaluation of the p-value computed for the observed data set needs to be calibrated. That is, the p-value should be compared to $\alpha^* = .02$ instead of $\alpha = .05$, because $\mathrm{P}(p < .02|H_0) = .05$.

5.5 Results for Examples

5.5.1 Example 1

To evaluate the performance of the plug-in p-value for Example 1, in total four double bootstraps are performed with $S = 1000$ and $T = 1000$: (A) hypothesis test type A with $n = 50$; (B) hypothesis test type B with $n = 50$; (C) hypothesis test type A with $n = 640$; and (D) hypothesis test type

Table 5.4: Results for the Double Bootstrap Procedure and the Parametric Bootstrap Procedure for Examples 1 to 3

	Hypothesis test	α^*	Λ	plug-in p-value
Example 1	type A ($n = 50$)	.048	-	-
	type A ($n = 640$)	.046	42.01	<.001
	type B ($n = 50$)	.024	-	-
	type B ($n = 640$)	.048	17.41	<.001
Example 2	type A	.038	2.55	.49
	type B	.058	1.07	.46
Example 3	type A	.056	132.42	<.001
	type B	-	0.0	>.999

B $n = 640$. In Figure 5.5 the four corresponding distributions of plug-in p-values are displayed.

As can be seen in Figure 5.5A and 5.5C the distribution for hypothesis test type A is almost uniform for both $n = 50$ and $n = 640$ with $\mathrm{P}(p < .048|H_0) = .05$ and $\mathrm{P}(p < .046|H_0) = .05$, respectively. As was shown by Silvapulle and Sen (2004, p. 32-33), the p-values of this statistical model and for hypothesis test type A are uniformly distributed. Our results for small and large sample sizes are pretty close to being uniform and the small deviations are sampling errors. Hence, the traditional alpha level of $\alpha = .05$ is used to evaluate the results for the observed data set.

For hypothesis test type B, however, the distribution is clearly not uniform, see the distributions in Figure 5.5B and 5.5D. High values of the plug-in p-value do not exist and low values appear too often. For small n, $\mathrm{P}(p < .02|H_0) = .05$ (see Table 5.4) and the alpha level for the analysis with the observed data set needs to be calibrated, $\alpha^* = .02$. Accidentally, for large n, $\mathrm{P}(p < .048|H_0) = .05$ and α^* does not need to be calibrated much, $\alpha^* = .048$. However, as can be seen in Figure 5.5D the distribution is clearly not uniform, for example $\mathrm{P}(p < .30|H_0) = .50$ and $\mathrm{P}(p < .38|H_0) = .70$, indicating that calibration is necessary for different alpha levels.

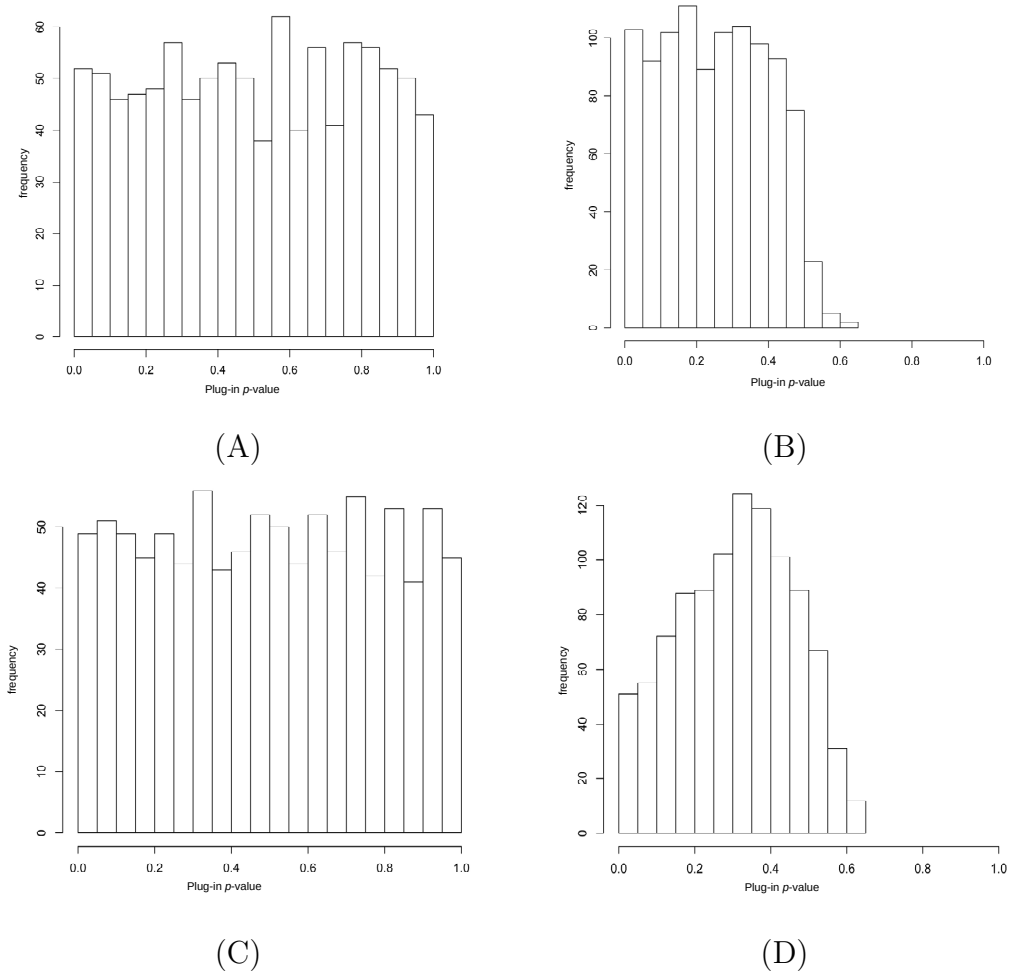
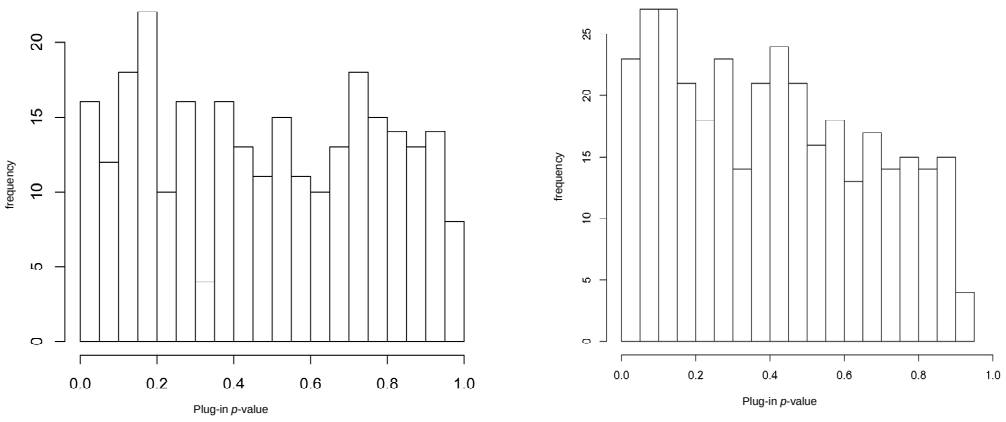


Figure 5.5: Distribution of plug-in p-values for example 1: In (A) and (C) hypothesis test type A is evaluated, in (B) and (D) hypothesis test B is evaluated; in (A) and (B) $n = 50$, in (C) and (D) $n = 640$

To evaluate the hypotheses in (5.15) and (5.16) for the observed data set, a parametric bootstrap is performed where 1000 data sets were generated. Each of these samples was fitted under H_0 for hypothesis test type A and B. Based on the results shown in Table 5.4, it can be concluded that for hypothesis test type A, H_0 can be rejected ($p < .001, \alpha = .05$). This implies



(A) (B)

Figure 5.6: Distribution of plug-in p-values for Example 2: in (A) test type A is evaluated, and in (B) test type B is evaluated

that the hypothesis $H_0 : \gamma_1 = \gamma_2 = \gamma_3$ is rejected in favor of the informative hypothesis, $H_1 : \{\gamma_1, \gamma_2\} < \gamma_3$. Moreover, the result of hypothesis test type B, indicates that the informative hypothesis is rejected ($p < .001, \alpha^* = .048$) in favor of the unconstrained model $H_1 : \gamma_1, \gamma_2, \gamma_3$. Inspection of Table 5.1 reveals that $\gamma_2 > \gamma_3$ and as such the unconstrained parameter estimates do not fit either $H_0 : \gamma_1 = \gamma_2 = \gamma_3$ or $H_1 : \{\gamma_1, \gamma_2\} < \gamma_3$.

In conclusion, levels of antisocial behaviour are not evenly predicted by how much adolescents tell the parents and by either a positive or a negative quality of the relationship with the parents (rejection of H_0). However, the expectation that disclosure is the best predictor does not hold (rejection of H_1).

5.5.2 Example 2

To evaluate the hypotheses for Example 2, two double bootstraps are performed to determine the correct alpha level for hypothesis test type A

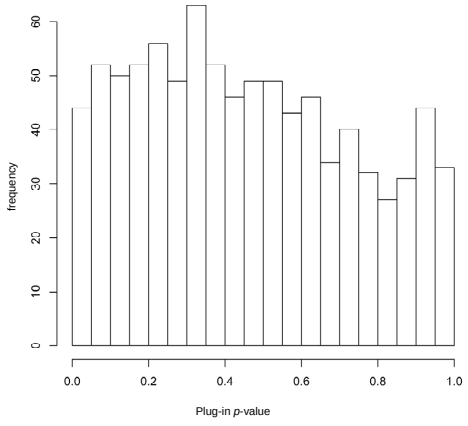


Figure 5.7: Distribution of plug-in p-values for Example 3, test type A.

and type B, with $S = 500$, $T = 500$ and group sizes for group $g = 1, \ldots, 4$ equal to the sample sizes.

The results are shown in Table 5.4 and the distribution of p-values is shown in Figure 5.6A and B for hypothesis test type A and B, respectively. The hypothesis H_0, shown in (5.20) can not be rejected in favor of hypothesis H_1 ($p = .49, \alpha^* = .04$), shown in (5.21). For hypothesis test type B the informative hypothesis can not be rejected in favor of the unconstrained hypothesis ($p = .46, \alpha^* = .056$). So, after testing the informative hypothesis it appears that the observed differences shown in Table 5.2 are too small to reject H_0 shown in (5.20). This makes sense since confidence intervals, if computed row-wise using 1.96 times SE, overlap and as such provide a lot of support for H_0 in (5.20).

In conclusion, compared to Dutch adolescents, adolescents from different ethnic groups are satisfied to a similar degree with their relationships with parents. Besides, Dutch adolescents disclose as much information as adolescents from different ethnic groups.

5.5.3 Example 3

For Example 3, a double bootstrap was performed with $S = 1000$, $T = 1000$ and $n = 640$ for hypothesis test type A and B. The results are shown in Table 5.4 and the distribution of p-values for hypothesis test type A is shown in Figure 5.7. For this hypothesis test, the null hypothesis is rejected in favor of the informative hypothesis ($p < .001, \alpha^* = .04$).

For hypothesis test type B, it appears that for all S bootstraps $p = 1$. This result implies that in none of the S bootstraps, the constraint parameter estimates violated the inequality constraints imposed on the regression coefficients. Also, for the observed data set $p = 1$. We wanted α to have the property $\mathrm{P}(p < .05|H_0) = .05$, but for this example we observed $\mathrm{P}(p < .05|H_0) = 0$. This simulation study shows that that we can not make an incorrect conclusion with respect to hypothesis test type B.

Thus, the association with deviant peers is the most prominent predictor of problem behaviour. A visual inspection of Table 5.3 confirms this conclusion since the unconstraint estimate β_{21} is larger than the unconstraint estimates of γ_{21}, γ_{22} and γ_{23}.

5.6 Concluding Remarks

Traditional hypothesis tests and model selection criteria are not equipped to deal with informative hypotheses formulated in terms of inequality constraints among the parameters of a structural equation model. In this paper we presented a solution for this problem using Mplus. Some issues that need further elaboration are now discussed.

First, p-values are often used in SEM and are evaluated using the traditional alpha level of .05. Using the double bootstrap procedure we evaluated the frequency properties of the plug-in p-values resulting from our method. These results show clearly that the distribution of the p-values is not always uniform and calibration is needed. This is especially the case when evaluating hypotheses of type B.

Second, in this paper we used rather simple SEM models. However, there are no technical limitations to use inequality constraints for more complicated models in Mplus, for example including latent effects, second order effects or categorical variables. In Figure 5.8 a hypothetical SEM model is shown with one latent variables, and seven observed variables. For this model constraints could be imposed on for example $\gamma_1 \dots \gamma_3$. In this paper we only discussed informative hypotheses with inequality constraints of the type $\gamma_1 > \gamma_2 > \gamma_3$. There are however no technical limitations to evaluate an informative hypothesis consisting of combinations of equality and inequality constraints of the form $\gamma_1 > \gamma_2 = \gamma_3$, $\{\gamma_1 - \gamma_2\} > 2$, or $\{\gamma_1 - \gamma_2\} < \gamma_3 > 1$.

A limitation of the procedure is that computational time can be substantial. To compute the examples in this paper we used pentium computers (3.20MHz) containing 2 duo-processors (Mplus can deal with multiple processors) with 1 GB memory. The models for Example 1 and 3 took approximately two days to compute, but Example 2 took more than two weeks.

So, although inequality constraints can be tested using the approach proposed, further research should focus on decreasing computation time. We therefore recommend to implement our procedure in Mplus. If this can be achieved, the method will be attractive for researchers like Deković et al. (2004), because they will be able to evaluate informative hypotheses easily and quickly.

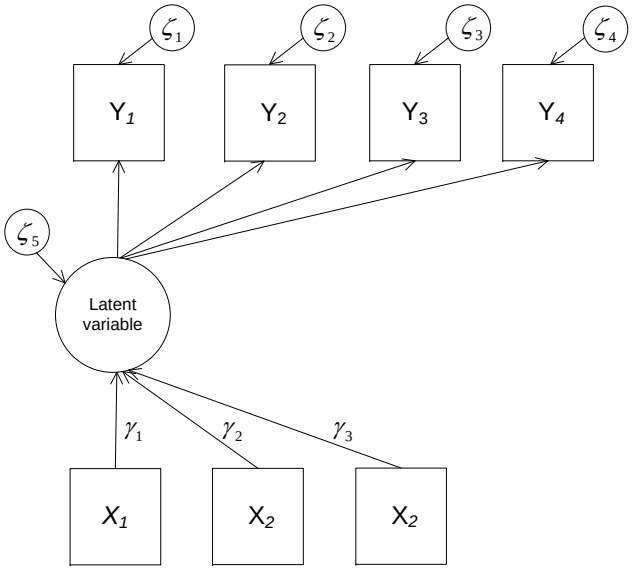


Figure 5.8: Hypothetical SEM model

6

A Prior Predictive Loss Function for the Evaluation of Inequality Constrained Hypotheses

Van de Schoot, R., Hoijtink, H., Brugman, D.,
& Romeijn, J.-W.

Manuscript under review

Abstract

The posterior predictive Deviance Information Criterium (*posterior* DIC) was proposed as a model selection tool by Spiegelhalter et al. (2002). In many types of statistical modeling inequality constraints are imposed between the parameters of interest. As we will show in this paper, the *posterior* DIC fails when comparing inequality constrained hypotheses. In this paper we will derive the prior DIC and show that it also fails when comparing inequality constrained hypotheses. However, it will be shown that a modification of the prior predictive loss function that is minimized by the DIC, and consequently a modification of the prior DIC does have the properties needed in order to be able to compare inequality constrained hypotheses. This new criterion will be called the Prior Information Criterium (PIC) and will be illustrated and evaluated using simulated data and examples.

Within the Bayesian framework, there are two perspectives on model selection: a prior predictive approach (e.g. Box, 1980; Kass & Raftery, 1995) and a posterior predictive approach (e.g. Gelman, Carlin, Stern & Rubin, 2004; Gelman et al., 1996). Spiegelhalter, Best, Carlin and Van Der Linde (2002) derived the posterior predictive Deviance Information Criterium (*posterior* DIC) to choose between a set of competing hypotheses. In this paper we will derive the prior predictive Deviance Information Criterium (*prior* DIC).

In many types of statistical modeling inequality constraints are imposed between the parameters of interest (Barlow et al., 1972; Hoijtink, Klugkist & Boelen, 2008; Robertson et al., 1988; Silvapulle & Sen, 2004; Van de Schoot, Hoijtink & Deković, 2010). More specifically, model parameters such as means or regression coefficients can be constrained to being greater or smaller than either a fixed value or other means or regression coefficients. Phrases like "the mean outcome in both experimental groups is expected to be larger than in the control group" and "women score higher than men in each condition" can be found in many papers. These specific expectations may be derived from theories, empirical evidence, or both.

As we will show in this paper, the *posterior* DIC can not be used to choose between a set of inequality constrained hypotheses. We also show that the *prior* DIC can only be used to choose between a set of constrained hypotheses if the population is fully in agreement with the inequality constrained hypothesis, but fails when the population is not in agreement with the constraints. To accommodate for this, the predictive loss function is modified. This new loss function can be approximated by the Prior Information Criterion (PIC) which can be used to evaluate a set of inequality constrained hypotheses. Simulated data and examples will be used to illustrate the performance of the PIC.

6.1 behaviour of The Posterior Predictive DIC in Constrained Model Selection

The *posterior* DIC is proposed in Spiegelhalter et al. (2002) as a Bayesian criterion for minimizing the posterior predictive loss. The *posterior* DIC has an important role in statistical model comparison. In this section we briefly show how the *posterior* DIC is obtained (based on the derivation of Spiegelhalter et al.), thereafter we show with two simple examples that the *posterior* DIC fails when comparing inequality constrained hypotheses.

6.1.1 Posterior Predictive DIC

The *posterior* DIC minimizes the posterior expectation of the expected loss (Gelman et al., 2004). It can be seen as the error that is expected when a statistical model based on the observed data set $\mathbf{y}$ is applied to a future data set $\mathbf{x}$. Let $f(\cdot)$ denote the likelihood, then the expected loss is given by

$$\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta}^*)}[-2\log f(\mathbf{x} \mid \bar{\boldsymbol{\theta}}_y)] \;, \tag{6.1}$$

where $-2\log f(\cdot)$ is the loss function of a future data set $\mathbf{x}$ in which $\bar{\boldsymbol{\theta}}_y$ is the expected a-posteriori estimate of the model parameters $\boldsymbol{\theta}$ based on the observed data set $\mathbf{y}$. If we would know the true parameter value $\boldsymbol{\theta}^*$, the expectation in (6.1) could be computed. However, since these are unknown, the *posterior* DIC takes the posterior expectation of (6.1) resulting in the posterior expectation of the expected loss given by

$$\begin{aligned}
&\mathrm{E}_{g(\boldsymbol{\theta}|\mathbf{y})}\Big\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[-2\log f(\mathbf{x} \mid \bar{\boldsymbol{\theta}}_y)\big]\Big\} = \\
&\quad -2\log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y) + \mathrm{E}_{g(\boldsymbol{\theta}|\mathbf{y})}\Big\{c(\mathbf{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}_y)\Big\} \approx \\
&\quad -2\log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y) + 2\big[-2\overline{\log f(\mathbf{y} \mid \boldsymbol{\theta})} + 2\log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)\big] = \\
&\quad \textit{posterior}\ \text{DIC} \;,
\end{aligned} \tag{6.2}$$

where $\mathrm{E}_{g(\theta|\mathbf{y})}$ denotes the expectation with respect to the posterior distribution $g(\boldsymbol{\theta} \mid \mathbf{y})$ and the term between square brackets is the penalty term,

often interpreted as the effective number of parameters. Let $\boldsymbol{\theta}^1 \dots \boldsymbol{\theta}^L$ be L draws from the posterior distribution $g(\boldsymbol{\theta} \mid \mathbf{y})$, then $-2\overline{\log f(\mathbf{y} \mid \boldsymbol{\theta})}$ can be estimated by

$$\sum_{l=1}^{L} \frac{-2 \log f(\mathbf{y} \mid \boldsymbol{\theta}^l)}{L} \ , \tag{6.3}$$

and $-2 \log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)$ can be estimated by

$$-2 \log f(\mathbf{y} \mid \sum_{l=1}^{L} \frac{\boldsymbol{\theta}_1^l}{L}, \dots, \sum_{l=1}^{L} \frac{\boldsymbol{\theta}_k^l}{L}) \ , \tag{6.4}$$

where k $(k = 1, \dots, K)$ index the parameters in $\boldsymbol{\theta}$.

6.1.2 Inequality Constrained Hypotheses: Example 1

To show that the *posterior* DIC fails when evaluating a set of inequality constrained hypotheses, consider an example where persons from two groups receive a score on one dependent variable, y_i $(i = 1, \dots, N)$:

$$y_i = \mu_1 d_{i1} + \mu_2 d_{i2} + \epsilon_i \ , \tag{6.5}$$

where μ_1 and μ_2 denote the mean score on y for group 1 and 2 respectively and where the residuals ϵ_i are assumed to be normally distributed $N(0, \sigma^2)$. The group membership of a person is denoted by $d_{ig} \in 0, 1$, where 1 and 0 denote that a person is either a member or not a member of group g. Suppose we want to evaluate two hypotheses: $H_0 : \mu_1, \mu_1$ and $H_1 : \mu_1 < \mu_2$. There are situations where the *posterior* DIC is unable to distinguish between H_0 and H_1.

Let $g_0(\mu_1, \mu_2, \sigma^2|\mathbf{y}) = g_1(\mu_1, \mu_2, \sigma^2|\mathbf{y}) \times c$, where c denotes the constant needed to normalize $g_1(\mu_1, \mu_2, \sigma^2|\mathbf{y})$ because it has density zero for all combinations of μ_1 and μ_2 not in agreement with H_1. The subscript in $g_0(\cdot)$ and $g_1(\cdot)$ refers to the posterior distribution of H_0 and H_1 respectively. Then, for $\mu_2 - \mu_1 \to \infty$, $g_0(\mu_1, \mu_2, \sigma^2|\mathbf{y})$ - $g_1(\mu_1, \mu_2, \sigma^2|\mathbf{y}) \to 0$. That is, if the population from which the data is generated is strongly in agreement with

H_1, the difference between the posterior distributions for H_0 and H_1 becomes zero. Since, the *posterior* DIC is computed using samples of μ_1, μ_2 and σ^2 obtained from the posterior distribution, for $\mu_2 - \mu_1 \to \infty$, samples obtained under H_0 and H_1 are exchangeable. Consequently, DIC_{H_0} and DIC_{H_1} have the same values. This result is counterintuitive and unwanted because H_1 is more parsimonious than H_0 and hence it contains more information (e.g. Sober, 2006), so it should be preferred by the DIC.

6.1.3 Inequality Constrained Hypotheses: Example 2

Consider a second example with two dependent variables (denoted by y_{1i} and y_{2i} for $i = 1, \ldots, N$),

$$\begin{aligned} y_{1i} &= \mu_1 + \epsilon_{i1} \\ y_{2i} &= \mu_2 + \epsilon_{i2} \ , \end{aligned} \tag{6.6}$$

where the residuals are assumed to be normally distributed

$$\begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{bmatrix} \sim N\Big(0, \Sigma\Big) \ , \Sigma = \begin{bmatrix} \sigma^2_{y_1} & \rho\sigma_{y_1}\sigma_{y_2} \\ \rho\sigma_{y_1}\sigma_{y_2} & \sigma^2_{y_2} \end{bmatrix} . \tag{6.7}$$

Suppose we want to evaluate two hypotheses: $H_0 : \mu_1, \mu_2$ and $H_1 : \mu_1 > 0;\ \mu_2 > 0$. Analogously to the previous example, also in this situation the *posterior* DIC is unable to distinguish between these hypotheses. Again we assume $g_0(\mu_1, \mu_2, \Sigma|\mathbf{y_1}, \mathbf{y_2}) = g_1(\mu_1, \mu_2, \Sigma|\mathbf{y_1}, \mathbf{y_2}) \times c$. Then, for $\mu_1 \to \infty$ and $\mu_2 \to \infty$, $g_0(\mu_1, \mu_2, \Sigma|\mathbf{y_1}, \mathbf{y_2}) - g_1(\mu_1, \mu_2, \Sigma|\mathbf{y_1}, \mathbf{y_2}) \to 0$. Since, the *posterior* DIC is computed using samples of μ_1, μ_2 and Σ obtained from the posterior distribution, for $\mu_1 \to \infty$ and $\mu_2 \to \infty$, samples obtained under H_0 and H_1 are exchangeable. Analogously to the previous example, DIC_{H_0} and DIC_{H_1} have the same values.

6.2 Derivation of Prior Predictive DIC

In this section we show how to obtain the *prior* DIC based on the derivation of the *posterior* DIC presented in Spiegelhalter et al. (2002). The point of

departure for the *prior* DIC is the same as for the *posterior* DIC, namely the expected loss given in (6.1). However, to deal with the unknown parameters $\boldsymbol{\theta}^*$, we take the expectation with respect to the *prior* distribution, $h(\boldsymbol{\theta})$, instead of the *posterior* expectation of the expected loss:

$$\begin{aligned} &\mathrm{E}_{h(\boldsymbol{\theta})}\Big\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[-2\log f(\mathbf{x} \mid \bar{\boldsymbol{\theta}}_y)\big]\Big\} = \\ &\quad -2\log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y) + \mathrm{E}_{h(\boldsymbol{\theta})}\big[c(\mathbf{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}_y)\big] \ . \end{aligned} \tag{6.8}$$

The main problem now, is to find an expression for the second term on the right hand side in (6.8). Using $D(\mathbf{a}, \mathbf{b}) = -2\log f(\mathbf{a} \mid \mathbf{b})$, $c(\mathbf{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}_y)$ in (6.8) can be rewritten to

$$\begin{aligned} c(\mathbf{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}_y) &= \mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x}, \bar{\boldsymbol{\theta}}_y) - D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y)\big] \\ &= \mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x}, \bar{\boldsymbol{\theta}}_y) - D(\mathbf{x}, \boldsymbol{\theta})\big] \\ &+ \mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x}, \boldsymbol{\theta}) - D(\mathbf{y}, \boldsymbol{\theta})\big] \\ &+ D(\mathbf{y}, \boldsymbol{\theta}) - D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y) \ . \end{aligned} \tag{6.9}$$

Now, $D(\mathbf{x}, \bar{\boldsymbol{\theta}}_y)$ in (6.9) can be approximated by taking a second order Taylor expansion about $\boldsymbol{\theta}$,

$$\begin{aligned} D(\mathbf{x}, \bar{\boldsymbol{\theta}}_y) \approx\ & -2\log f(\mathbf{x} \mid \boldsymbol{\theta}) - 2\left\{\frac{\partial \log f(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\}^T \big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big) - \\ & -\big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big)^T \left\{\frac{\partial^2 \log f(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\}\big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big) \ . \end{aligned} \tag{6.10}$$

Since $-2\log f(\mathbf{x} \mid \boldsymbol{\theta})$ is equal to $D(\mathbf{x}, \boldsymbol{\theta})$ and the expectation of the second term on the right hand side of (6.10) with respect to $f(\mathbf{x} \mid \boldsymbol{\theta})$ is zero (p. 604 Spiegelhalter et al., 2002),

$$\begin{aligned} &\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x}, \bar{\boldsymbol{\theta}}_y) - D(\mathbf{x}, \boldsymbol{\theta})\big] \approx \\ &\quad \mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\left[-\big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big)^T \left\{\frac{\partial^2 \log f(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\}\big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big)\right] \ . \end{aligned} \tag{6.11}$$

The expression on the right hand side of (6.11) can be rewritten as $\mathrm{tr}\big\{\mathbf{I}(\boldsymbol{\theta})\big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big)\big(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta}\big)^T\big\}$ and since $\mathbf{x}$ and $\mathbf{y}$ stem from the same data generating

mechanism, the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta})$ can be approximated by the observed Fisher information matrix, $\mathbf{I}(\bar{\boldsymbol{\theta}}_y)$ (p. 604 Spiegelhalter et al., 2002), where $\mathbf{I}(\bar{\boldsymbol{\theta}}_y) = -\partial^2 \log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$. Using $\mathrm{E}\{\mathrm{tr}(\cdot)\} = \mathrm{tr}\{\mathrm{E}(\cdot)\}$, the prior expectation of $c(\mathbf{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}_y)$ can now be approximated by:

$$\begin{aligned} \mathrm{E}_{h(\boldsymbol{\theta})}\big[c(\mathbf{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}_y)\big] &\approx \mathrm{tr}\big\{\mathbf{I}(\bar{\boldsymbol{\theta}}_y)\boldsymbol{\Lambda}\big\} + \\ &\quad + E_{h(\boldsymbol{\theta})}\Big\{E_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x}, \boldsymbol{\theta}) - D(\mathbf{y}, \boldsymbol{\theta})\big]\Big\} + d \ , \end{aligned} \tag{6.12}$$

where $\boldsymbol{\Lambda} = E_{h(\boldsymbol{\theta})}[(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta})(\bar{\boldsymbol{\theta}}_y - \boldsymbol{\theta})^T]$ denotes the variation in the prior distribution around $\bar{\boldsymbol{\theta}}_y$. The last term on the right hand side of (6.12) is defined as

$$\begin{aligned} d &= \mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big] - \mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y)\big] \\ &= \mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big] - D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y) \ . \end{aligned} \tag{6.13}$$

To show that $\mathrm{tr}\big\{\mathbf{I}(\bar{\boldsymbol{\theta}}_y)\boldsymbol{\Lambda}\big\}$ is approximately equal to d, we use a second order Taylor expansion about $\bar{\boldsymbol{\theta}}_y$:

$$\begin{aligned} \mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big] \approx \mathrm{E}_{h(\boldsymbol{\theta})}\Bigg[& D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y) - 2\Big\{\frac{\partial \log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)}{\partial \boldsymbol{\theta}}\Big\}^T \big(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_y\big) - \\ & - \big(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_y\big)^T \Big\{\frac{\partial^2 \log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\Big\}\big(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_y\big)\Bigg] \ . \end{aligned} \tag{6.14}$$

Since, $\bar{\boldsymbol{\theta}}_y \to \bar{\boldsymbol{\theta}}_{ML}$ for $n \to \infty$, $-2\Big\{\frac{\partial \log f(\mathbf{y}|\bar{\boldsymbol{\theta}}_y)}{\partial \boldsymbol{\theta}}\Big\}^T$ is asymptotically zero (Gelman et al., 2004). This way, $\mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big]$ can now be approximated by

$$\begin{aligned} \mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big] &\approx D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y) + \mathrm{E}_{h(\boldsymbol{\theta})}\Big[\mathrm{tr}\Big\{-\frac{\partial^2 \log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\big(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_y\big)\big(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_y\big)^T\Big\}\Big] \\ &\approx D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y) + \mathrm{tr}\big\{\mathbf{I}(\bar{\boldsymbol{\theta}}_y)\boldsymbol{\Lambda}\big\} \ . \end{aligned} \tag{6.15}$$

To show that $\mathrm{tr}\big\{\mathbf{I}(\bar{\boldsymbol{\theta}}_y)\boldsymbol{\Lambda}\big\}$ is approximately equal to d, $D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y)$ is subtracted from both sides of (6.15)

$$\mathrm{tr}\big\{\mathbf{I}(\bar{\boldsymbol{\theta}}_y)\boldsymbol{\Lambda}\big\} \approx \mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big] - D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y) = d \ . \tag{6.16}$$

Equation (6.12) then becomes

$$\begin{aligned} \mathrm{E}_{h(\boldsymbol{\theta})}\big[c(\mathbf{y},\boldsymbol{\theta},\bar{\boldsymbol{\theta}}_y)\big] \;\approx\; & \mathrm{E}_{h(\boldsymbol{\theta})}\Big\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x},\boldsymbol{\theta}) - D(\mathbf{y},\boldsymbol{\theta})\big]\Big\} + \\ & +2\Big\{\mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y},\boldsymbol{\theta})\big] - D(\mathbf{y},\bar{\boldsymbol{\theta}}_y)\Big\} \;. \end{aligned} \tag{6.17}$$

The *prior* DIC can now be written as

$$\mathrm{E}_{h(\boldsymbol{\theta})}\Big\{E_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x},\boldsymbol{\theta})\big]\Big\} - D(\mathbf{y},\bar{\boldsymbol{\theta}}_y) + \mathrm{E}_{h(\boldsymbol{\theta})}\big[D(\mathbf{y},\boldsymbol{\theta})\big] \;, \tag{6.18}$$

whereas, using the same notation, the *posterior* DIC can be written as

$$D(\mathbf{y},\bar{\boldsymbol{\theta}}_y) + 2\Big\{\mathrm{E}_{g(\boldsymbol{\theta}|\mathbf{y})}\big[D(\mathbf{y},\boldsymbol{\theta})\big] - D(\mathbf{y},\bar{\boldsymbol{\theta}}_y)\Big\} \;. \tag{6.19}$$

Note the two major differences between the *prior* and *posterior* DIC: (i) the first term on the right hand side of (6.18) does not have a corresponding part in the *posterior* DIC and (ii) the third term on the right hand side of (6.18) is the expectation with respect to the prior distribution whereas the corresponding term in (6.19) is the expectation with respect to the posterior distribution.

In what follows we will first show in Section 6.3 how to choose $h(\boldsymbol{\theta})$. Thereafter, we prove in Section 6.4 that $\mathrm{E}_{h(\boldsymbol{\theta})}\Big\{E_{f(\mathbf{x}|\boldsymbol{\theta})}\big[D(\mathbf{x},\boldsymbol{\theta})\big]\Big\}$ is constant when comparing inequality constrained hypotheses. In Section 6.5 we inspect the behaviour of the *prior* DIC with respect to the evaluation of inequality constrained hypotheses for the examples presented in Section 6.1.2 and 6.1.3. In Section 6.6 we will introduce a new loss function and a new model selection tool for the evaluation of inequality constrained hypotheses, the Prior Information Criterium (PIC). Finally in Section 6.7 the PIC is used in an application.

6.3 Prior Distributions for Constrained Hypotheses

An important issue when computing the *prior* DIC is the specification of the prior distribution $h(\boldsymbol{\theta})$. A key characteristic is that constraints can be

incorporated in $h(\boldsymbol{\theta})$ and we show how to do so in this section. Finally, we provide the prior distributions we will use to evaluate Example 1 and 2 introduced in Section 6.1.2 and 6.1.3, respectively.

6.3.1 Truncated Prior Distribution

We want to use the *prior* DIC to choose between a set of hypotheses that differ in the specification of the constraints between the parameters of interest. In order to incorporate the constraints in the prior distribution, we use the encompassing prior approach as was proposed by Klugkist et al. (2005); Mulder, Hoijtink and Klugkist (2009); Mulder, Klugkist et al. (2009).

Let $h(\boldsymbol{\theta})$ be the prior distribution for the model parameters $\boldsymbol{\theta}$. Furthermore, let $H_t(t = 0, \ldots, T)$ denote a hypothesis specified using constraints and let H_0 denote an unconstrained hypothesis. All hypotheses H_t are nested in H_0, therefore $h_t(\boldsymbol{\theta})$ is proportional to $h_0(\boldsymbol{\theta})$, with

$$h_t(\boldsymbol{\theta}) : \begin{cases} c_t^{-1} h_0(\boldsymbol{\theta}) & \text{if} \boldsymbol{\theta} \in H_t \\ 0 & \text{otherwise} \end{cases}, \tag{6.20}$$

where c_t is a normalization constant given by

$$c_t = \int_{\boldsymbol{\theta} \in H_t} h_0(\boldsymbol{\theta}) \partial \boldsymbol{\theta} \ . \tag{6.21}$$

Using this encompassing prior approach only the prior distribution for H_0 needs to be specified. This is in agreement with the principle of compatibility which is best illustrated using a quote from Leucari and Consonni (2003) "If nothing was elicited to indicate that the two priors should be different, then it is sensible to specify [the prior of the constrained hypothesis] to be, ..., as close as possible to [the prior of the unconstrained hypothesis]. In this way the resulting Bayes factor should be least influenced by dissimilarities between the two priors due to differences in the construction processes, and could thus more faithfully represent the strength of the support that the data lend to each [hypothesis]" (see also, Roverato & Consonni, 2004).

Now, let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_c, \boldsymbol{\theta}_n\}$ where $\boldsymbol{\theta}_c$ denotes model parameters subjected to constraints and $\boldsymbol{\theta}_n$ denotes model parameters that are unconstrained, then $h_0(\boldsymbol{\theta}) = h_0(\boldsymbol{\theta}_c)h_0(\boldsymbol{\theta}_n)$. Let k $(k = 1, \ldots, K)$ index the parameters in $\boldsymbol{\theta}_c$, then, assuming prior independence between both groups of parameters,

$$h_0(\boldsymbol{\theta}_c) = \prod_{k=1}^{K} h_0(\boldsymbol{\theta}_{c_k}) \ . \tag{6.22}$$

The parametrization of $h_0(\boldsymbol{\theta}_{c_k})$ and $h_0(\boldsymbol{\theta}_n)$ depend on the statistical model used. In the current paper we will use a scaled inverse chi-square distribution or an inverse Wishart distribution for $h_0(\boldsymbol{\theta}_n)$, and for $h_0(\boldsymbol{\theta}_{c_k})$ normal prior distributions will be used.

Choosing the parameters of the prior distribution in constrained hypotheses selection is for example explained in Mulder, Hoijtink and Klugkist (2009) or in Mulder, Klugkist et al. (2009). We will not repeat their elaboration, but we will limit ourselves to the specification of $h_0(\boldsymbol{\theta}_{c_k})$ and $h_0(\boldsymbol{\theta}_n)$ for the examples used in the current paper.

An important characteristic of the prior distribution for $h_0(\boldsymbol{\theta}_{c_k})$ is that the means of the prior distribution need to be located at the boundary of the inequality constrained parameter space. Also, the prior variance need to be restricted in the same way as the prior means. As was shown by Mulder, Hoijtink and Klugkist (2009) only in this situation the prior operationalizes the complexity of a model adequately (for more details see: Mulder, Hoijtink & Klugkist, 2009; Mulder, Klugkist et al., 2009). In the sequel two examples and an application will be used to illustrate this is also the case for our own model selection tool.

To account for this important feature of the prior distribution we will use a simple data-based procedure to obtain values for the parameters of the prior distribution. Our procedure is in line with the training data approach and the fractional Bayes factors as is presented in J. Berger and Pericchi (1996, 2004); O'Hagan (1995); Perez and Berger (2002). There are also other methods available, such as described in Klugkist et al. (2005); Mulder, Hoijtink and Klugkist (2009); Mulder, Klugkist et al. (2009). In the current

paper we will not compare different specifications of $h_0(\boldsymbol{\theta}_{c_k})$ but limit ourself to our simple method that will be illustrated in the sequel.

6.3.2 Prior Specification for Example 1 and 2

According to Mulder, Hoijtink and Klugkist (2009), the prior distribution for Example 1 should be given by

$$h_0(\mu_1, \mu_2, \sigma^2) = N(\mu_1|\mu_0, \tau_0^2) \times N(\mu_2|\mu_0, \tau_0^2) \times Inv\chi^2(\sigma^2|\upsilon_0, \sigma_0^2), \quad (6.23)$$

where μ_0 is the prior mean and τ_0^2 is the prior variance. Note that if μ_1 and μ_2 have the same prior distribution, then μ_0 will be located on the boundary of the inequality constrained parameter space. In our rather simple solution, μ_0 and τ_0^2 will be based on a fraction of the information in the data with respect to the overall mean that corresponds to a minimal training sample of size 2:

$$\mu_0 = \sum_{i=1}^{N} \frac{y_i}{N}, \quad (6.24)$$

and

$$\tau_0^2 = \sum_{i=1}^{N} \frac{(y_i - \mu_0)^2}{N} \times \frac{1}{2}, \quad (6.25)$$

where the correction term of $\frac{1}{2}$ renders the expected variance of the mean in a minimal training sample of size 2. The scaled inverse chi squared prior distribution for σ^2 in (6.23), has υ_0 degrees of freedom and has scale parameter σ_0^2, with $\upsilon_0 = 2$ and $\sigma_0^2 = \tau_0^2 \times 2$. Our straightforward solution to obtain reasonable values for υ_0 and σ_0^2 such that this prior is vague.

According to Mulder, Hoijtink and Klugkist (2009), for Example 2,

$$h_0(\mu_1, \mu_2, \Sigma) = MVN(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\tau}_0^2) \times W^{-1}(\Sigma|\upsilon_0, \boldsymbol{\Sigma}_0), \quad (6.26)$$

where $\boldsymbol{\mu}_0 = \{0, 0\}$, and $\boldsymbol{\tau}_0^2$ is a diagonal matrix with variances τ_{01}^2 and τ_{02}^2. Note that we assume independence on the off-diagonal element in (6.27) and (6.29) in line with Mulder, Hoijtink and Klugkist (2009).

The specification of μ_0 is somewhat different because of the use of the constants in hypothesis $H_1 : \mu_1 > 0;\ \mu_2 > 0$. In this situation, the prior should be centered at the boundary of the admissible parameter space. In our case this is zero. This is because zero is the boundary of the inequality constrained parameter space (see, Mulder, Hoijtink & Klugkist, 2009; Mulder, Klugkist et al., 2009).

For $\boldsymbol{\tau}_0^2$ and $\boldsymbol{\Sigma}_0$ will be determined using the expected information in a training sample of size 2. $\boldsymbol{\tau}_0^2$ is given by

$$\begin{bmatrix} \tau_{01}^2 & 0 \\ 0 & \tau_{02}^2 \end{bmatrix} , \tag{6.27}$$

where

$$\tau_{0\cdot}^2 = \sum_{i=1}^{N} \frac{(y_{\cdot i})^2}{N} \times \frac{1}{2} . \tag{6.28}$$

For the Inverse Wishart, υ_0 are degrees of freedom ($\upsilon_0 = 2$) and $\boldsymbol{\Sigma}_0$ is the scale matrix with

$$\begin{bmatrix} \sigma_{01}^2 & 0 \\ 0 & \sigma_{02}^2 \end{bmatrix} , \tag{6.29}$$

where

$$\sigma_{0\cdot}^2 = \frac{2}{N} \sum_{i=1}^{N} (y_{\cdot i})^2 . \tag{6.30}$$

6.4 Simplifying the Prior DIC for Constrained Hypotheses

As we will prove in this section, $\mathrm{E}_{h_t(\boldsymbol{\theta})} \Big\{ E_{f(\mathbf{x}|\boldsymbol{\theta})} \big[D(\mathbf{x}, \boldsymbol{\theta}) \big] \Big\}$ in (6.18) is constant between constrained hypotheses. In this context the *prior* DIC reduces to

$$\textit{prior}\ \mathrm{DIC} = C + 2 \log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y) + \mathrm{E}_{h_t(\boldsymbol{\theta})} \big[-2 \log f(\mathbf{y} \mid \boldsymbol{\theta}) \big] , \tag{6.31}$$

where $C = \mathrm{E}_{h_t(\boldsymbol{\theta})} \Big\{ E_{f(\mathbf{x}|\boldsymbol{\theta})} \big[-2 \log f(\mathbf{x} \mid \boldsymbol{\theta}) \big] \Big\}$ and can be ignored for all H_t.

6.4.1 Example 1 Continued

For Example 1, $h_t(\boldsymbol{\theta}_c)h_t(\boldsymbol{\theta}_u) = h_t(\mu_1, \mu_2)h_t(\sigma^2)$ where $h_t(\sigma^2)$ is the same, but $h_t(\mu_1, \mu_2)$ differs across hypotheses because of the normalization of the prior distribution in Equation (6.20). In the remainder of this subsection we drop the subscript t to simplify the notation. We will prove that

$$E_{h(\sigma^2)h(\mu_1,\mu_2)}\Big\{E_{f(\mathbf{x}|\mu_1,\mu_2,\sigma^2)}\big[-2\log f(\mathbf{x} \mid \mu_1, \mu_2, \sigma^2)\big]\Big\}$$

is constant over all hypotheses under consideration. When comparing constrained hypotheses we have to prove that the term within accolades is independent of μ_1, μ_2, and σ^2. First using

$$f(\mathbf{x} \mid \mu_1, \mu_2, \sigma^2) = \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^N \exp\Big[-\frac{1}{2}\frac{\sum_{i=1}^N (x_i - \mu_1 d_1 - \mu_2 d_2)^2}{\sigma^2}\Big] \ , \quad (6.32)$$

the term being constant can be written as

$$\int_{\sigma^2}\int_{\mu_1,\mu_2}\int_{\mathbf{x}} 2N\log\sqrt{2\pi\sigma^2}\ \partial f(\mathbf{x} \mid \mu_1, \mu_2, \sigma^2)\partial h(\mu_1, \mu_2)\partial h(\sigma^2) + \quad (6.33)$$
$$+\int_{\sigma^2}\int_{\mu_1,\mu_2}\int_{\mathbf{x}}\sum_{i=1}^N \frac{(x_i - \mu_1 d_1 - \mu_2 d_2)^2}{\sigma^2}\ \partial f(\mathbf{x} \mid \mu_1, \mu_2, \sigma^2)\partial h(\mu_1, \mu_2)\partial h(\sigma^2) \ .$$

The first term of (6.33) is independent of μ_1, μ_2, and since $h(\sigma^2)$ is the same for each hypothesis, the second term integrated over σ^2 in (6.33) should be constant for every value for σ^2 to render (6.33) constant. Let $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2\}$ denote subgroups with sample sizes N_1 and N_2 for $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. Omitting the integral over σ^2, we can now rewrite the second term in (6.33) to

$$\int_{\mu_1}\int_{\mathbf{x}_1}\sum_{i=1}^{N_1}\frac{(x_i - \mu_1)^2}{\sigma^2}\ \partial f(\mathbf{x}_1 \mid \mu_1, \sigma^2)\partial h(\mu_1) + \quad (6.34)$$
$$+\int_{\mu_2}\int_{\mathbf{x}_2}\sum_{i=1}^{N_2}\frac{(x_i - \mu_2)^2}{\sigma^2}\ \partial f(\mathbf{x}_2 \mid \mu_2, \sigma^2)\partial h(\mu_2) \ .$$

Note, that for the first group in (6.34) $x_i \sim N(\mu_1, \sigma^2)$ and for the second group $x_i \sim N(\mu_2, \sigma^2)$. Using $x_i^* = \frac{x_i - \mu_1}{\sigma^2}$ with $x_i^* \sim N(0,1)$ in the first group, and $x_i^* = \frac{x_i - \mu_2}{\sigma^2}$ with $x_i^* \sim N(0,1)$ in the second group, the integral over μ_1 and μ_2 drop out of (6.34):

$$\int_{\mathbf{x}_1^*} \sum_{i=1}^{N} (x_i^*)^2 \; \partial f(\mathbf{x}_i^* \mid 0,1) + \int_{\mathbf{x}_2^*} \sum_{i=1}^{N} (x_i^*)^2 \; \partial f(\mathbf{x}_i^* \mid 0,1) \; . \tag{6.35}$$

Consequently, for every value of σ^2, (6.34) is independent of μ_1, μ_2. That is, for this example, $E_{h(\sigma^2)h(\mu_1,\mu_2)}\Big\{E_{f(\cdot)}\big[-2\log f(\cdot)\big]\Big\}$ is constant over constrained hypotheses.

6.4.2 Example 2 Continued

For Example 2, $h_t(\boldsymbol{\theta}_c)h_t(\boldsymbol{\theta}_u) = h_t(\mu_1, \mu_2)h_t(\Sigma)$ where $h_t(\Sigma)$ is the same, but $h_t(\mu_1, \mu_2)$ differs across hypotheses because of the normalization of the prior distribution in Equation (6.20). In the remainder of this subsection we drop the subscript t to simplify the notation. We now have to prove that the term between accolades in

$$E_{h(\mu_1,\mu_2)h(\sigma_{x1},\sigma_{x2},\rho)}\Big\{E_{f(\cdot)}\big[-2\log f(\mathbf{x}_1, \mathbf{x}_2 \mid \mu_1, \mu_2, \sigma_{x1}, \sigma_{x2}, \rho)\big]\Big\} \tag{6.36}$$

is constant over hypotheses for μ_1, μ_2, and Σ. Using

$$\begin{aligned} f(\mathbf{x}_1, \mathbf{x}_2 \mid \mu_1, \mu_2, \sigma_{x1}, \sigma_{x2}, \rho) \;\; = \;\; & \Big(\frac{1}{2\pi\sigma_{x_1}\sigma_{x_2}\sqrt{1-\rho^2}}\Big)^N \exp\Big[-\frac{1}{2(1-\rho^2)} \\ & \Big\{\frac{\sum_{i=1}^N (x_{1i}-\mu_1)^2}{\sigma_{x_1}^2} + \frac{\sum_{i=1}^N (x_{2i}-\mu_2)^2}{\sigma_{x_2}^2} - \\ & - \frac{2\rho\sum_{i=1}^N (x_{1i}-\mu_1)(x_{2i}-\mu_2)}{\sigma_{x_1}\sigma_{x_2}}\Big\}\Big], \end{aligned} \tag{6.37}$$

(6.36) can be written as the sum of

$$\begin{aligned} \int_{\sigma_{x1},\sigma_{x2},\rho} \int_{\mu_1,\mu_2} \int_{\mathbf{x}_1,\mathbf{x}_2} & 2N \log 2\pi\sigma_{x_1}\sigma_{x_2}\sqrt{1-\rho^2} \\ & \partial f(\mathbf{x}_1, \mathbf{x}_2 \mid \mu_1, \mu_2, \sigma_{x1}, \sigma_{x2}, \rho)\partial h(\mu_1, \mu_2)\partial h(\sigma_{x1}, \sigma_{x2}, \rho) \; , \end{aligned} \tag{6.38}$$

and

$$\int_{\sigma_{x1},\sigma_{x2},\rho}\int_{\mu_1,\mu_2}\int_{\mathbf{x}_1,\mathbf{x}_2}\frac{1}{(1-\rho^2)} \left\{\frac{\sum_{i=1}^N(x_{1i}-\mu_1)^2}{\sigma_{x_1}^2}+\frac{\sum_{i=1}^N(x_{2i}-\mu_2)^2}{\sigma_{x_2}^2}-\frac{2\rho\sum_{i=1}^N(x_{1i}-\mu_1)(x_{2i}-\mu_2)}{\sigma_{x_1}\sigma_{x_2}}\right\} \partial f(\mathbf{x}_1,\mathbf{x}_2 \mid \mu_1,\mu_2,\sigma_{x1},\sigma_{x2},\rho)\partial h(\mu_1,\mu_2)\partial h(\sigma_{x1},\sigma_{x2},\rho)\ . \tag{6.39}$$

Since $h(\Sigma)$ is the same for each hypothesis, the integrals in (6.39) integrated over $\sigma_{x1},\sigma_{x2},\rho$ should be constant for every value of $h(\Sigma)$ to render (6.39) constant. Also, in this situation (6.38) is constant over constrained hypotheses. Using $x_{1i}^*=\frac{x_{1i}-\mu_1}{\sigma}$ and $x_{2i}^*=\frac{x_{2i}-\mu_2}{\sigma}$, (6.39) can be rewritten into

$$\int_\rho\int_{\mathbf{x}_1^*,\mathbf{x}_2^*}\sum_{i=1}^N\frac{1}{(1-\rho^2)}\Big\{(x_{1i}^*)^2+(x_{2i}^*)^2-2\rho^2x_{1i}^*x_{2i}^*\Big\} \partial f(\mathbf{x}_1^*,\mathbf{x}_2^*\mid 0,0,1,1,\rho)\partial(\rho)\ . \tag{6.40}$$

Consequently, for every Σ, (6.39) is independent of μ_1 and μ_2. That is, for this example, $E_{h(\mu_1,\mu_2)h(\sigma_{x1},\sigma_{x2},\rho)}\Big\{E_{f(\cdot)}\big[-2\log f(\cdot)\big]\Big\}$ is constant over constrained hypotheses.

6.4.3 Multivariate Models

Finally, consider a multivariate example with two groups with mean scores on two dependent variables:

$$\begin{aligned} y_{1i}&=\mu_{11}d_{ig1}+\mu_{12}d_{ig2}+\epsilon_{1i}\\ y_{2i}&=\mu_{21}d_{ig1}+\mu_{22}d_{ig2}+\epsilon_{2i}\ , \end{aligned} \tag{6.41}$$

where $\mu_{1\cdot}$ and $\mu_{2\cdot}$ denote the mean score on y_1 and y_2 respectively and where $\mu_{\cdot 1}$ and $\mu_{\cdot 2}$ denote the mean for group 1 and 2 respectively. Again, group membership of a person is denoted by $d_{ig}\in 0,1$ and the residuals are assumed to be normally distributed with

$$\begin{bmatrix}\epsilon_{i1}\\ \epsilon_{i2}\end{bmatrix}\sim N\Big(0,\Sigma\Big)\ ,\Sigma=\begin{bmatrix}\sigma_{y_1}^2 & \rho\sigma_{y_1}\sigma_{y_2}\\ \rho\sigma_{y_1}\sigma_{y_2} & \sigma_{y_2}^2\end{bmatrix}\ . \tag{6.42}$$

Note that this example is a combination of (6.5) and (6.6). Also for constrained hypotheses in this multivariate example it can be proved that

$$E_{h_t(\mu_{11},\mu_{12},\mu_{21},\mu_{22})h_t(\Sigma)}\Big\{E_{f(\cdot)}\big[-2\log f(\cdot)\big]\Big\}$$

is constant over constrained hypotheses. Even so, using the same steps as presented in Section 6.4.1 and 6.4.2, it can be proved for the general multivariate normal linear model Press (2005, pp. 252-257), that

$$\mathrm{E}_{h_t(\boldsymbol{\theta})}\Big\{E_{f(\mathbf{x}|\boldsymbol{\theta})}\big[-2\log f(\mathbf{x}\mid\boldsymbol{\theta})\big]\Big\}$$

is constant over constrained hypotheses.

6.5 Evaluating Inequality Constrained Hypotheses

In this section we show how to compute the *prior* DIC. We also show that the *prior* DIC can be used to choose between a set of constrained hypotheses if the population from which the data is generated is fully in agreement with the most constrained hypothesis, where the *posterior* DIC fails to do so. Furthermore, we show that the *prior* DIC also fails to choose between a set of inequality constrained hypotheses if the population is *not* in agreement with the constrained hypothesis. To accommodate for this inconvenience, we will show in the next section that the prior predictive loss that is estimated by the prior DIC needs to be adjusted in order to be able to evaluate inequality constrained hypotheses.

6.5.1 Estimation of the Prior DIC

Let $\boldsymbol{\theta}^1 \ldots \boldsymbol{\theta}^L$ be L draws from the posterior distribution, then $D(\mathbf{y}, \bar{\boldsymbol{\theta}}_y)$, in Equation (6.31) can be estimated by

$$2\,\log\, f(\mathbf{y}\mid \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\theta}_1^l, \ldots, \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\theta}_k^l)\;. \tag{6.43}$$

Furthermore, let $\boldsymbol{\theta}^1 \ldots \boldsymbol{\theta}^K$ be K draws from the prior distribution, then $\mathrm{E}_{h_t(\boldsymbol{\theta})}\big[D(\mathbf{y}, \boldsymbol{\theta})\big]$ in Equation (6.31) can be estimated by

$$\frac{1}{K}\sum_{k=1}^{K} \text{-2 } \log f(\mathbf{y} \mid \boldsymbol{\theta}^k) \ . \tag{6.44}$$

We will show for each example how both can be obtained, but we start with a close examination of the behaviour of the *prior* DIC for Example 1 for inequality constrained hypotheses.

6.5.2 Example 1 continued

To show that the *prior* DIC can be used to choose between a set of constrained hypotheses if the population from which the data is generated is fully in agreement with the most constrained hypothesis, whereas the *posterior* DIC fails to do so, we reconsider Example 1, see Section 6.1.2 with hypotheses $H_0 : \mu_1, \mu_2$ and $H_1 : \mu_1 < \mu_2$.

If we would compare H_0 and H_1 with the *prior* DIC, the first term of the *prior* DIC given in Equation (6.31) is constant, as was shown in Section 6.4.1. Now, consider the same situation as in Section 6.1.2 where the population from which the data was generated is strongly in agreement with H_1. In this case, the second term in Equation (6.31) does also not differ between H_0 and H_1, because for $\mu_1 - \mu_2 \to \infty$, $\bar{\mu}_1|H_0 \to \bar{\mu}_1|H_1$ and $\bar{\mu}_2|H_0 \to \bar{\mu}_2|H_1$. So, the third term, $\mathrm{E}_{h_t(\mu_1,\mu_2,\sigma^2)}\big[D(\mathbf{y}, \mu_1, \mu_2, \sigma^2)\big]$, should make the difference between H_0 and H_1.

Since samples of μ_1 and μ_2 are taken from the prior distribution $h_0(\mu_1, \mu_2, \sigma^2)$, see Equation (6.44), and since $h_0(\mu_1, \mu_2, \sigma^2) \neq h_1(\mu_1, \mu_2, \sigma^2)$ because of the normalization of the prior distribution according to Equation (6.20), samples from the prior distribution are different for H_0 and H_1. For $\mu_1 - \mu_2 \to \infty$, the third term of (6.31), when computed for H_0 is based on more large values of $-2\log\ f(\mathbf{y} \mid \mu_1, \mu_2, \sigma^2)$ then when it is computed for H_1. Consequently, the third term of (6.31) for H_1 is smaller then the third term of (6.31) for H_0.

The actual computation of the second term of (6.31) for H_0 and H_1 can be done using samples from $g_0(\mu_1, \mu_2, \sigma^2|\mathbf{y})$ and $g_1(\mu_1, \mu_2, \sigma^2|\mathbf{y})$, respectively, that can be used in (6.43). These samples can be obtained using the Gibbs sampler for $g_0(\cdot)$ (see, Gelman et al., 2004) and the constrained Gibbs sampler for $g_1(\cdot)$ (see, Klugkist et al., 2005). The third term of (6.31) can be computed using a sample from the prior distribution of the hypotheses under investigation which subsequently can be used in (6.44).

We also consider an equality constrained hypothesis to investigate the performance of the *posterior* and *prior* DIC, $H_2 : \mu_1 = \mu_2$. For this hypothesis, μ_1 and μ_2 in (6.32) can be replaced by μ, $h(\mu)$ has mean μ_0 and variance $\frac{1}{1/\tau_0^2+1/\tau_0^2}$ and $g_2(\mu, \sigma^2|\mathbf{y})$ is proportional to this likelihood and this prior distribution. For H_2 both (6.43) and (6.44) are computed from the posterior and prior distribution, respectively.

A simulation study was performed where data sets from seven different populations were considered and the three hypotheses were evaluated with the *prior* and *posterior* DIC, see Figure 6.1. Note that the first four data sets are in agreement with the constraints of H_1, whereas the last three data sets are constructed in such a way that they violate the constraints of H_1. The difference between the seven data sets is that the size of the difference between the two group means varies from small (difference of .02) to large (difference of 2). Data were constructed in such a way that the sample means and variance are exactly equal to the population parameters (with $\sigma^2 = 1$ and $n = 20$ for each group). The specification of μ_0, τ_0^2, υ_0 and σ_0^2 is described in Section 6.3.2. For population 1 with $\mu_1 = -1$ and $\mu_2 = 1$, the priors are $\mu_0 = 0$, $\tau_0^2 = 0.97$, $\upsilon_0 = 2$ and $\sigma_0^2 = 1.95$.

Inspection of Figure 6.1 leads to two important observations: (1) there are situations that the *posterior* DIC fails to correctly distinguish between H_0, H_1 and H_2; (2) there are also situations where the *prior* DIC fails to correctly distinguish between H_0, H_1 and H_2.

First, consider the performance the *posterior* DIC. When looking at populations 1-5 in Figure 6.1, it can be seen that the the values for the

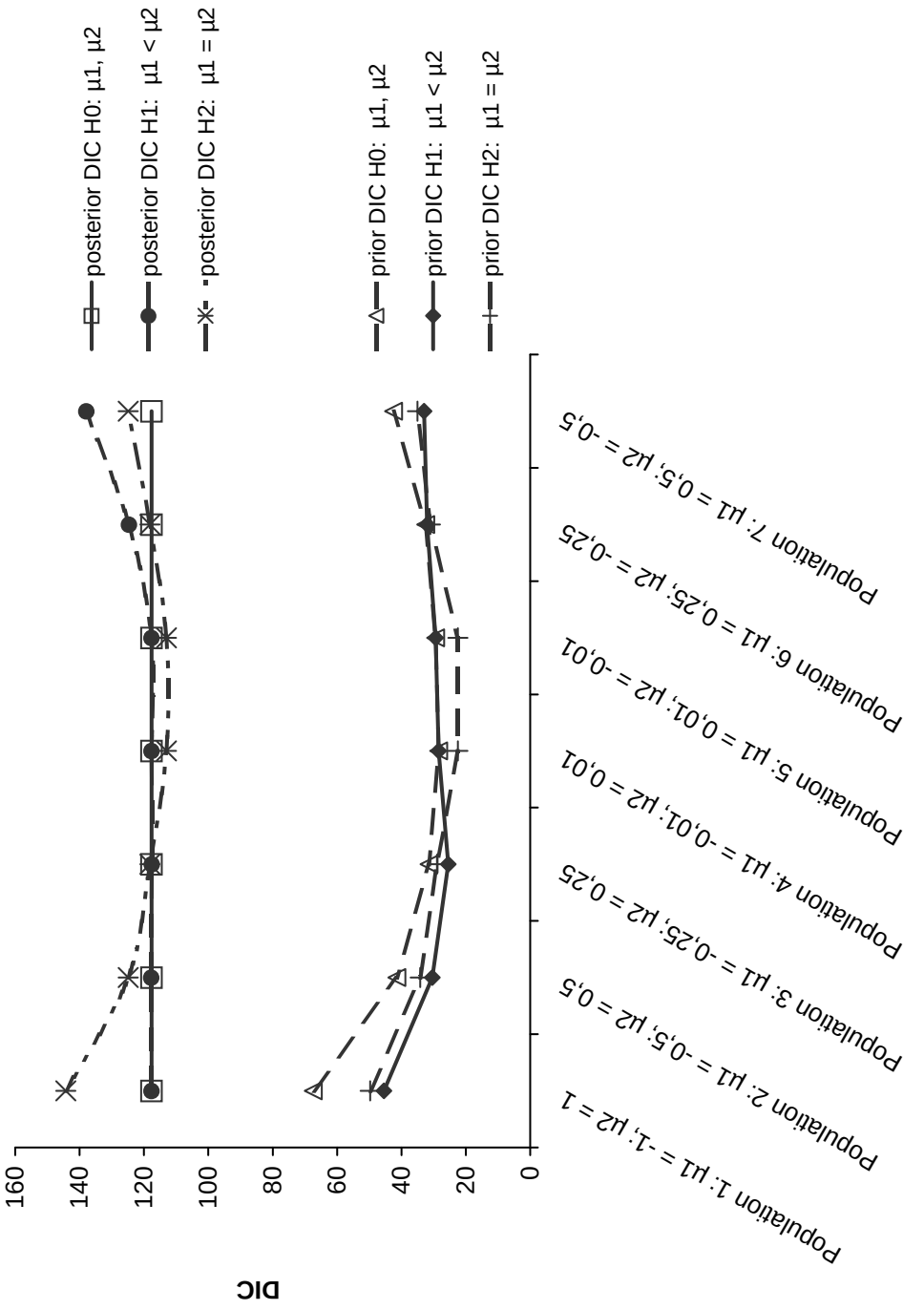


Figure 6.1: Posterior and prior DIC for populations 1-7 in Example 1.

posterior DIC for both H_0 and H_1 are equal. Hence, the *posterior* DIC can not distinguish H_0 and H_1, which is counterintuitive because the population values satisfy the constraints of H_1. For population 4 and 5, the two data sets with the smallest difference in sample means, the value of the *posterior* DIC for H_2 is lowest, which makes sense because the means are approximately equal. When the means do not fit the constraints of H_1, populations 6 and 7, the values for the *posterior* DIC for H_0, H_1 and H_2 are in line with what would be expected: the lowest DIC value for H_0 followed by H_2 and H_1, respectively. In sum, the *posterior* DIC fails to distinguish between hypotheses H_0 and H_1 when the data is strongly in agreement with the most constrained hypothesis, H_1.

Second, consider the performance the *prior* DIC. In contrast to the *posterior* DIC, the *prior* DIC is able to correctly distinguish between H_0 and H_1 when the data are in agreement of the constraints of H_1, see populations 1-3 in Figure 6.1, where the the *prior* DIC is lowest for H_1. For the data with the smallest differences in sample means (population 4 and 5), the *prior* DIC is lowest for H_2. When the constraints are not supported by the data, populations 6-7, the value for H_0 should be the lowest value, but as can be seen in Figure 1, this is not the case! So, when the data is fully in agreement with H_1 the the *prior* DIC outperforms the the *posterior* DIC, but when the data do not support H_1, the *prior* DIC fails to correctly distinguish H_0 from H_1 and H_2.

In conclusion, neither the *prior* DIC, nor the *posterior* DIC are proper model selection tools for the evaluation of inequality constrained hypotheses. In the next section the prior predictive loss function will be adjusted such that its estimate, the PIC, can be used to select the best of a set of equality and inequality constrained hypotheses.

6.6 A New Loss Function for the Evaluation of Inequality Constrained Hypotheses

In this section we take a closer look at why the *prior* DIC fails and how this problem can be solved by introducing a new loss function..

6.6.1 The Problem

Consider Figure 6.2, where estimates of (6.43) and (6.44) are displayed for all populations described in the previous section. First, as described in the previous section, a problem arises for population 6 and 7 where H_1 instead of H_0 has smaller values of the *prior* DIC see Figure 6.1. Consider the prior expectation of the expected loss given in (6.1), which is approximated by the *prior* DIC as was shown in Section 6.2:

$$\begin{aligned} &\mathrm{E}_{h_t(\boldsymbol{\theta})}\Big\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[-2\log f(\mathbf{x} \mid \bar{\boldsymbol{\theta}}_y)\big]\Big\} \\ &\quad\approx 2\log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y) + \mathrm{E}_{h_t(\boldsymbol{\theta})}\big[-2\log f(\mathbf{y} \mid \boldsymbol{\theta})\big]\ . \end{aligned} \tag{6.45}$$

Note that this loss function determines how well replicated data fit a certain hypothesis, that is, how good $\bar{\boldsymbol{\theta}}_y$ is a summary of $\mathbf{x}$. However, this loss function does not accommodate 'bad' fitting hypotheses, that is, if for a hypothesis $\bar{\boldsymbol{\theta}}_y$ is not a good summary of $\mathbf{y}$, this will not be detected by the loss function in (6.45).

Consider the situation of Example 1 and suppose that a population is not in agreement with the inequality constrained hypothesis, $H_1 : \mu_1 < \mu_2$, for example Population 7 with population means $\mu_1 = 0.5; \mu_2 = -0.5$. In this situation the *prior* DIC chooses H_1 as the best hypothesis, see Figure 6.1. This result is unwanted because in the data $\mu_1 > \mu_2$. This result is due to the fact that the prior $\overline{\mu}_1 \approx 0; \overline{\mu}_2 \approx 0$ because of the truncation of the prior distribution. Subsequently for the computation of (6.45), data is replicated based on $\boldsymbol{\theta}$ from a prior distribution with $\mu_0 = 0$. These replicated data are adequately summarized by $\overline{\mu}_1$ and $\overline{\mu}_2$. However, what is not accounted for in

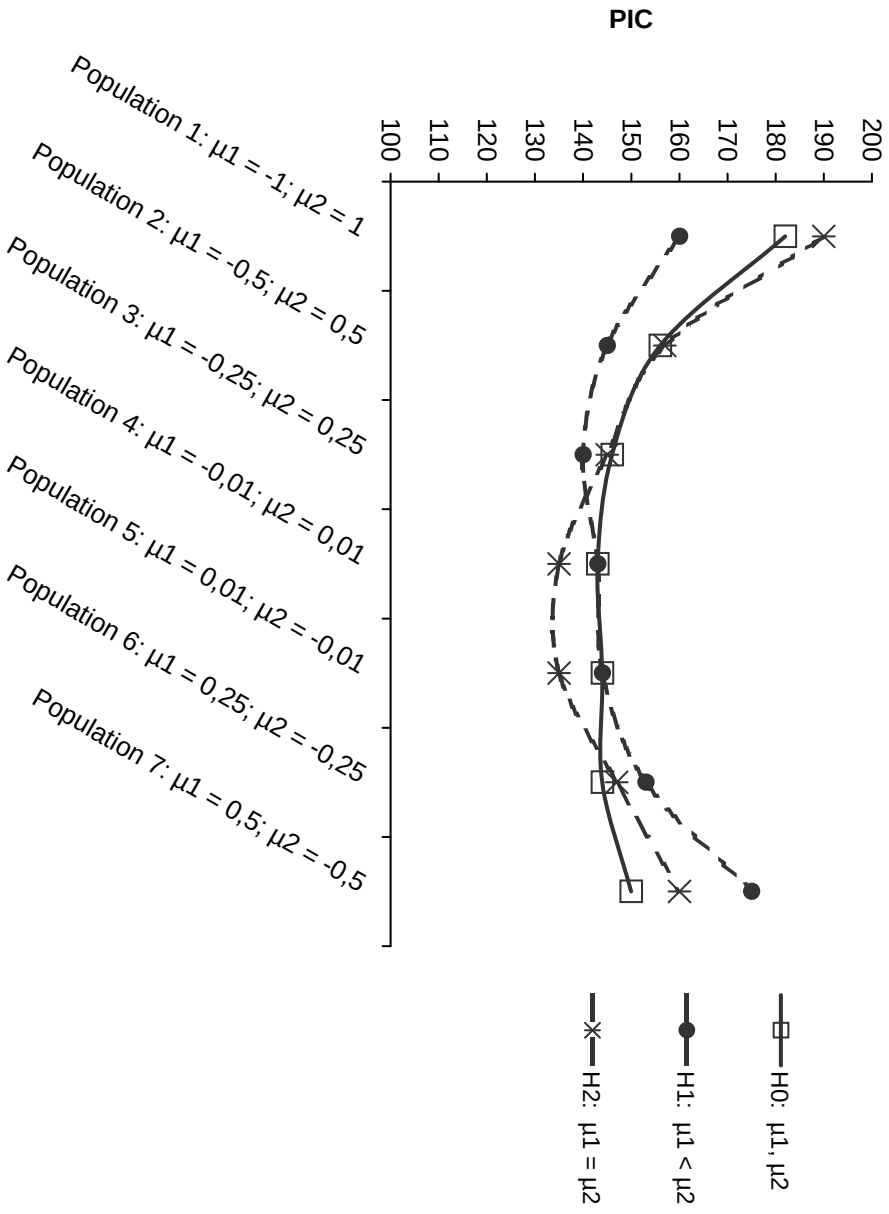


Figure 6.2: The PIC for populations 1-7 of Example 1.

(6.45) is that the observed data $\mathbf{y}$ are not adequately summarized by $\overline{\mu}_1$ and $\overline{\mu}_2$. This leads to situations where the loss function in (6.45) has a preference for 'bad' fitting inequality constrained hypotheses.

6.6.2 The Solution

The solution of the aforementioned problem is to adjust the loss function that is used to select the best hypothesis such that it also accounts for the agreement between $\bar{\boldsymbol{\theta}}_y$ and $\mathbf{y}$. The loss function in (6.45) can be rewritten as

$$\begin{aligned} -2\ \mathrm{E}_{h_t(\boldsymbol{\theta})}\Big\{\mathrm{E}_{f(\mathbf{x}|\boldsymbol{\theta})}\big[\log f(\mathbf{x} \mid \bar{\boldsymbol{\theta}}_y) + \log f(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_y)\big]\Big\} \\ \approx \mathrm{E}_{h_t(\boldsymbol{\theta})}\big[-2\log f(\mathbf{y} \mid \boldsymbol{\theta})\big]\ . \end{aligned} \tag{6.46}$$

The new loss function determines not only how well replicated data fit with a certain hypothesis (the first term between accolades in 6.46), but it also determines how well a hypothesis fits the data (the second term between accolades in 6.46). It is approximated by the third term of the *prior* DIC and is our final model selection tool, to be called Prior Information Criterium (PIC). In Figure 6.2 the PIC values for populations 1-7 of Example 1 are shown. As can be seen, the PIC chooses for H_1 as the best hypothesis in situations where this hypothesis is true in the population, see populations 1-3. The PIC chooses for H_2 as the best hypothesis where this hypothesis is strongly supported by the population values, see populations 4 and 5. Finally, the PIC chooses for the unconstrained hypothesis, H_0, where the (in)equality constraints for both H_1 and H_2 are not supported by the data, see populations 6 and 7. These results makes the PIC outperform both the *posterior* and *prior* DIC in all situations.

6.6.3 Prior Sensitivity

In this section we evaluate the influence of the prior specification on the PIC. To do so, we performed a simulation study where μ_0, τ_0^2, υ_0 and σ_0 are varied. We evaluated H_0, H_1 and H_2 for populations 1, 4, and 7 with: (1) $\mu_0 - 1$,

$\mu_0 + 0$ and $\mu_0 + 1$; (2) $\tau_0 \times .5$, $\tau_0 \times 1$, and $\tau_0 \times 5$; (3) $\upsilon_0 = 2$ and $\upsilon_0 = 5$; (4) $\sigma_0 \times .5$, $\sigma_0 \times 1$, and $\sigma_0 \times 5$.

The results are presented in Table 6.1 with in bold the correct conclusions. As can be seen, the specification of the prior influences the results and it is sensitive for different values for μ_0, τ_0, υ_0 and σ_0. However, as can be seen for different values of the priors the influences mainly the height of PIC and not the relative ordering of H_0, H_1, and H_2. It is therefore to be expected that any reasonable data based method can be used to specify the parameters of the prior distribution if the goal is to select the best of a set of (in)equality constrained hypotheses using the PIC.

6.6.4 Example 2 Continued

Let us return to Example 2 with $H_0 : \mu_1, \mu_2$ and $H_1 : \mu_2 > 0, \mu_1 > 0$. The situation for this example is analogous to Example 1, that is, also for Example 2, the first and the third term of Equation (6.31) are similar for H_0 and H_1 if the population from which the data was generated is fully in agreement with H_1. Again, since, $h_0(\mu_1, \mu_2, \Sigma) \neq h_1(\mu_1, \mu_2, \Sigma)$, because of the normalization of the prior distribution, samples from the prior distribution are different for H_0 and H_1. For $\mu_1 \to \infty, \mu_2 \to \infty$, the integral used to compute $\mathrm{E}_{h_0(\cdot)}[\cdot]$ takes all values of $-2 \log f(\mathbf{y_1}, \mathbf{y_2} \mid \mu_1, \mu_2, \Sigma)$ into account, whereas $\mathrm{E}_{h_1(\cdot)}[\cdot]$ only takes values of -2 log $f(\mathbf{y_1}, \mathbf{y_2} \mid \mu_1, \mu_2, \Sigma)$ into account where $\mu_1 > 0, \mu_2 > 0$. Since, in the latter case less small values for $\mathrm{E}_{h_1(\cdot)}[\cdot]$ are sampled, *prior* $\mathrm{DIC}_{H_0} >$ *prior* DIC_{H_1}.

Analogously to Example 1, the *prior* DIC does not correctly distinguish H_0 and H_1 because the loss function does not take 'bad' fitting hypotheses into account. Therefore we use the the PIC, see (6.46) to select the best hypothesis. Note that we also evaluated $H_2 : \mu_1 = 0; \mu_2 = 0$.

To evaluate H_0, H_1 and H_2 we performed a simulation study where data sets from six different populations were considered. The deviance from zero for the two means varies from small to large, see Figure 6.3. Populations 1-4 satisfy the constraints of H_1 and populations 5-6 are not in agreement with

Table 6.1: Sensitivity of the PIC. The bold numbers represent the hypothesis that should be preferred by the PIC and is also preferred by the PIC.

		Population 1	Population 4	Population 7
H_0 with	$\mu_0 - 1$	203	186	**189**
	μ_0	182	144	**155**
	$\mu_0 + 1$	202	187	**188**
H_1 with	$\mu_0 - 1$	**180**	185	200
	μ_0	**164**	143	167
	$\mu_0 + 1$	**181**	187	199
H_2 with	$\mu_0 - 1$	203	**178**	228
	μ_0	183	**136**	162
	$\mu_0 + 1$	202	**178**	228
H_0 with	$\tau_0 \times .5$	185	150	**162**
	$\tau_0 \times 1$	182	144	**155**
	$\tau_0 \times 5$	292	215	**236**
H_1 with	$\tau_0 \times .5$	**160**	151	174
	$\tau_0 \times 1$	**164**	143	167
	$\tau_0 \times 5$	**270**	214	257
H_2 with	$\tau_0 \times .5$	186	**143**	175
	$\tau_0 \times 1$	183	**136**	162
	$\tau_0 \times 5$	195	**207**	238
H_0 with	$\upsilon_0 = 2$	182	144	**155**
	$\upsilon_0 = 5$	168	130	**141**
H_1 with	$\upsilon_0 = 2$	**160**	143	167
	$\upsilon_0 = 5$	**145**	129	152
H_2 with	$\upsilon_0 = 2$	186	**136**	162
	$\upsilon_0 = 5$	171	**127**	146
H_0 with	$\sigma_0^2 \times .5$	213	167	**180**
	$\sigma_0^2 \times 1$	183	144	**155**
	$\sigma_0^2 \times 5$	200	169	**177**
H_1 with	$\sigma_0^2 \times .5$	**167**	165	203
	$\sigma_0^2 \times 1$	**161**	143	167
	$\sigma_0^2 \times 5$	**196**	169	181
H_2 with	$\sigma_0^2 \times .5$	246	**156**	218
	$\sigma_0^2 \times 1$	186	**136**	162
	$\sigma_0^2 \times 5$	177	**159**	179

H_1. Data sets were constructed in such a way that the sample means and variance-covariance matrix are exactly equal to the population parameters ($\rho = .4; \sigma_1^2 = 1; \sigma_2^2 = 1$). For each of these data sets, we computed the *posterior* DIC and the PIC for H_0, H_1 and H_2. The specification of μ_0, τ_0, υ_0 and σ_0^2 is described in Section 6.3.2. Like for Example 1, samples for the prior distribution can be used to compute the PIC. For population 1 with $\mu_1 = 1$ and $\mu_2 = 1$, the priors are $\mu_0 = 0$, $\tau_0 = 0.98$, $\upsilon_0 = 3$ and $\sigma_0^2 = 3.95$.

The results are shown in Figure 6.3 and it can be seen that the PIC outperforms the *posterior* DIC when the data is in agreement with the constraints of H_1. Compare, for example, the results for population 2, where the PIC has the lowest value for H_1 but the *posterior* DIC is indifferent between H_0 and H_1.

6.7 Moral Judgment Competence

In this section we use the PIC as a model selection tool in an application. Leenders and Brugman (2005) investigated whether moral judgment competence and attitude towards delinquent behaviour create a domain shift in young adolescents. That is, a certain behavior which in society as a whole is considered to be not moral (e.g. aggression, violence), might be a group convention in certain adolescent groups. In total 135 pupils of intermediate secondary schools in the Netherlands were asked to report about self conducted aggressive acts. They were also asked to judge aggressive acts and vandalistic acts in hypothetical situations on how moral they thought the behaviour was. The researchers had specific ideas about differences in the level of morality in these hypothetical situations between pupils that did or did not report to conduct aggressive acts themselves.

The statistical model is analogous to Equation (6.41) where $\mu_{1\cdot}$ and $\mu_{2\cdot}$ denote the mean score on the hypothetical construct vandalism (denoted by y_1) and the hypothetical construct aggression (denoted by y_2) and where $\mu_{\cdot 1}$ and $\mu_{\cdot 2}$ denote the mean for the group reported not to conduct aggressive acts

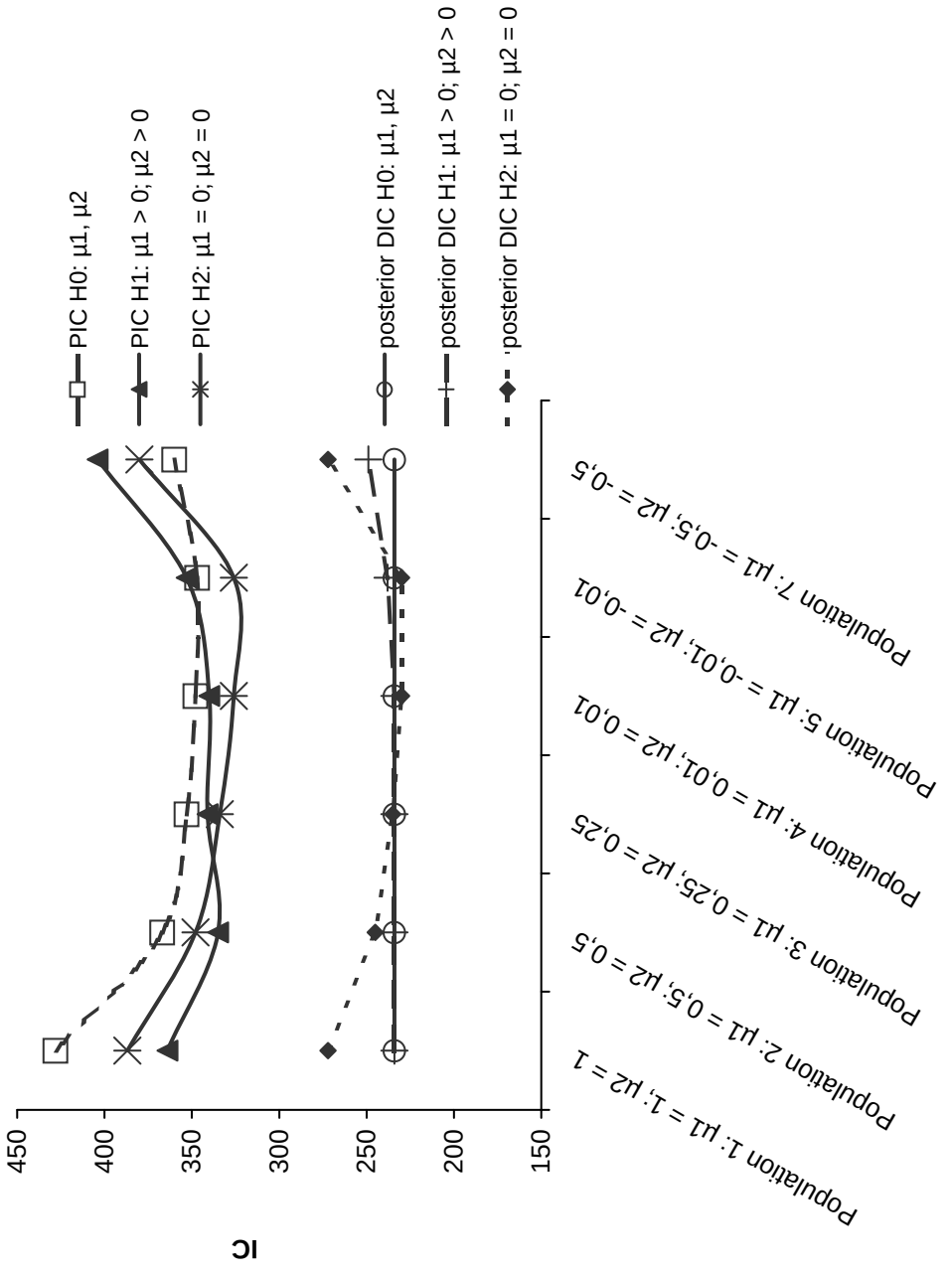


Figure 6.3: The results of Example 2 for populations 1-6 for the *posterior* DIC and the PIC

and the group that did report to conduct aggressive acts, respectively. Again, group membership of a person is denoted by $d_{ig} \in 0, 1$ and the residuals are assumed to be normally distributed, see Equation (6.42).

There are three hypotheses of interest

$$\begin{aligned} H_0 &: \mu_{12}, \mu_{11} \text{ and } \mu_{22}, \mu_{21} \\ H_1 &: \mu_{12} > \mu_{11} \text{ and } \mu_{22} > \mu_{21} \\ H_2 &: \mu_{12} = \mu_{11} \text{ and } \mu_{22} > \mu_{21} \ . \end{aligned} \tag{6.47}$$

The first hypothesis, is an unconstrained hypothesis (H_0). A second hypothesis, H_1, postulates that the aggressive group ($\mu_{\cdot 2}$) also judge the same behaviour in all hypothetical situations to be more conventional and as such morally more appropriate than their peers who do not report such behaviour ($\mu_{\cdot 1}$). The third hypothesis, H_2, is that there is a domain shift in the judgement about hypothetical situations. That is, for pupils that reported to have conducted some delinquent behavior (i.e. aggression), in the same hypothetical situation, they will judge it to be more morally accepted compared to adolescents that did not report to conduct the same behavior. However, in hypothetical situations concerning other delinquent behaviour that was not reported by these same adolescents (i.e. vandalism), they will judge the hypothetical situation to be equally morally condemnable as adolescents that did not report any antisocial behaviour.

The prior distribution, $h_0(\boldsymbol{\theta}_c, \boldsymbol{\theta}_n) = h_0(\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})\ h_0(\Sigma)$, is chosen again such that the prior mean is on the border of the admissible parameter space. We used a multivariate normal distribution for the means and an inverse Wishart distribution for the variance-covariance matrix (see, Mulder, Hoijtink & Klugkist, 2009):

$$h_0(\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, \Sigma) = MVN(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\tau}_0^2) \times W^{-1}(\Sigma|\upsilon_0, \boldsymbol{\Sigma}_0), \tag{6.48}$$

where $\boldsymbol{\mu} = \{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\}$ and $\boldsymbol{\mu}_0 = \{\mu_0, \mu_0, \mu_0, \mu_0\}$ where

$$\mu_0 = \sum_{i=1}^{N} \frac{y_{1i} + y_{2i}}{2N}, \tag{6.49}$$

Table 6.2: Descriptive Statistics ($n_1 = 38; n_2 = 97; \rho = .52$)

	Mean	SD
μ_{11}	5.37	1.23
μ_{12}	5.68	1.62
μ_{21}	5.27	1.27
μ_{22}	6.71	2.14

$\boldsymbol{\tau}_0^2$ is the prior variance-covariance matrix with

$$\begin{bmatrix} \tau_0^2 & 0 \\ 0 & \tau_0^2 \end{bmatrix} , \tag{6.50}$$

where

$$\tau_0^2 = \sum_{i=1}^{N} \frac{(y_{1i})^2 + (y_{2i})^2}{2N} \times \frac{1}{2} . \tag{6.51}$$

For the Inverse Wishart, $\upsilon_0 = 3$ and $\boldsymbol{\Sigma}_0$ is the scale matrix

$$\begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_0^2 \end{bmatrix} , \tag{6.52}$$

where

$$\sigma_0^2 = \frac{2}{2N} \sum_{i=1}^{N} (y_{1i})^2 + (y_{2i})^2 . \tag{6.53}$$

In Table 6.2 group means and standard deviations (SD) are provided. The results of the model selection procedure are presented in Table 6.3. As can be seen in this table the *posterior* DIC is indifferent for all hypotheses, whereas the PIC chooses for H_2. This result can be confirmed when looking at the group means in Table 6.2 where μ_{22} is larger then μ_{21} and μ_{11} is close to μ_{12}.

6.8 Conclusion

In this paper we showed how to obtain the *prior* DIC based on the derivation of the *posterior* DIC presented in Spiegelhalter et al. (2002). The point of

Table 6.3: Model Selection Results for the application

Hypothesis	PIC	*posterior* DIC
H_0	1044	935
H_1	1023	935
H_2	**872**	935

departure for the *prior* DIC is the same as for the *posterior* DIC, namely the expected loss. The derivation of the *prior* DIC is provided and the choice for the prior distribution, which is based on training data is motivated (see also Mulder, Hoijtink & Klugkist, 2009). Its performance is illustrated using examples and we showed that the *prior* DIC can be used to choose between a set of constrained hypotheses if the population from which the data is generated is fully in agreement with the most constrained hypothesis, where the *posterior* DIC failed to do so. However, the *prior* DIC fails to choose between a set of inequality constrained hypotheses if the population is *not* in agreement with the constrained hypothesis.

In conclusion, neither the *prior* DIC, nor the *posterior* DIC are proper model selection tools for the evaluation of inequality constrained hypotheses. To accommodate for this, the loss function that is minimized by the *prior* DIC was adjusted. The proposed loss function determines not only how well replicated data fit with a certain hypothesis, but it also determines how well a hypothesis fits the data. It is approximated by a new model selection tool, the Prior Information Criterium (PIC). We demonstrated with three examples that the PIC is able to select the best of a set of (in)equality constrained hypotheses.

PART III

Applications

7

Do Delinquent Young Adults Have a High or a Low Level of Self-concept?

Van de Schoot, R. & Wong, T.

In press for *Self & Identity*

Abstract

This study explored the levels of self-concept of delinquent young adults (n = 873). This question is of theoretical and practical importance, as therapeutic programmes addressing the self-concept must be based on clear evidence. The present study demonstrated that self-concept is related to delinquent behaviour and that men and women differ both in the strength and direction of the association. Furthermore, Bayesian latent class analysis revealed that both high-delinquent and non-delinquent men and women fall into two groups: those with high levels of self-concept and those with low levels of self-concept. This pattern emerged across the 12 different domains of self-concept assessed. These results may help to explain inconsistent results of previous studies on the link between self-concept and delinquency.

7.1 Introduction

In recent years the association between delinquent behaviour and self-concept in children, adolescents and young adults has received much attention (see, e.g., the review of Baumeister, Boden & Smart, 1996). A common finding in this type of research is that a low self-concept is a significant and powerful risk factor for many types of negative life outcomes, including delinquency (e.g. Donnellan, Trzesniewski, Robins, Moffitt & Caspi, 2005; Fergusson & Horwood, 2002; Webster, Kirkpatrick, Nezlek, Smith & Paddock, 2007). This has led to many intervention programmes aiming to reduce delinquent behaviour by improving offenders' self-concepts (Mason, 2003). However, other researchers argue that delinquency in young adults stems from a high self-concept that is threatened or disputed by others (e.g., Bushman & Baumeister, 1998). If this is correct, then intervention programmes aiming at increasing the level of self-concept may make things worse.

Given the theoretical attention surrounding this issue as well as its clinical implications, there is a need to gather clear evidence regarding the link between self-concept and delinquent behaviour: Do young adult offenders have a high or a low self-concept?

7.1.1 Self-concept and Delinquent Behaviour

Throughout the life span, self-esteem and self-perceived competencies (i.e., self-concept) are considered essential determinants of well-being and functioning (Baumeister, Campbell, Krueger & Vohs, 2003; Harter, 1990). In this study we focus on self-concept. Since there is often confusion about the difference between self-concept and self-esteem, it is useful to define these two terms and explain how they differ. Self-concept can be defined as the knowledge, appreciation and understanding a person has of him/herself (e.g. Cole, Chan & Lytton, 1989). As a person grows older, perceived competencies are characterized by increasing differentiation of competence domains (Harter, 1990, 1999). Self-esteem, however, is the evaluative component of self-

concept (Harter, 1990, 1999). It addresses "how one feels about the self when it is viewed as an object of evaluation" (Campbell, 1990, p. 539). We focus on the self-concepts of young adults and their life tasks, which include domains such as work-related achievements, finding a spouse, and leaving the parental home (see also Visser-Van Balen, Laak, Treffers, Sinnema & Geenen, 2007).

Previous research into the association between self-concept and delinquent behaviour yielded contradictory conclusions that resulted in four different notions about this association. In the current study, we translated these notions into four predictions through which we examined whether young adult offenders would show low or high self-concepts.

Prior literature suggests competing predictions about the link between self-concept and delinquency. The first prediction would be that young adults who commit offences have a *lower* self-concept compared to young adults who do not commit offences. This negative association has been found among different nationalities, age groups, and different assessments of the self-concept and delinquent behaviour (e.g. Donnellan et al., 2005; Murphy, Stosny & Morrel, 2005; Trzesniewski et al., 2006). Trzesniewski et al. (2006), for example, tested the hypothesis that low self-concept predicts negative real-world consequences such as delinquency. Using prospective data, they found that adolescents with a low self-concept were more likely to be convicted of a crime during adulthood than adolescents with a higher self-concept. An explanation might be that negative self-views sabotage the ability to cope successfully with events. That is, as Swann, Chang-Schneider and McClarty (2007) state, in the wake of failure experiences, people with negative self-views are more likely to suffer emotional trauma than people with positive self-views.

The second prediction would be that young adults who commit offences have similar levels of self-concept when compared to young adults who do not commit offences. 'Similar' here does not refer to equal levels of self-concept; instead it refers to unrelatedness between level of self-concept and

delinquency. Hence, the prediction would be for no association between delinquent behaviour and self-concept. Many studies have failed to find an association (e.g. Bushman & Baumeister, 1998; Jang & Thornberry, 1998; Neumark-Sztainer, Story, French & Resnick, 1997; Salmivalli, Kaukiainen, Kaistaniemi & Lagerspetz, 1999). Bushman and Baumeister (1998), for example, did experiments to find an explanation of the role of self-concept in aggression. However, there appeared to be no significant correlation between levels of self-concept and aggression in any of the three situations they studied.

The third prediction would be that young adults who commit offences have a *higher* self-concept than young adults who do not commit offences (e.g. Piko, Fitzpatrick & Wright, 2005; Rigby & Slee, 1993; Spencer, Josephs & Steele, 1993). Baumeister et al. (1996), who reviewed various bodies of findings, found that perpetrators of aggression might have positive and perhaps even inflated views of themselves. A possible explanation Baumeister et al. provide is that delinquents seem to believe they are superior to others and therefore might feel entitled to help themselves to the resources of other people.

These contradictory findings might be reconciled by a fourth prediction: There are two groups of young adults who commit offences, a group that has *low* levels of self-concept and a group that has *high* levels of self-concept (Baumeister et al., 2003; Diamantopoulou, Rydell & Henricsson, 2008; Boden, Fergusson & Horwood, 2007; Vermeiren, Bogaerts, Ruchkin, Deboutte & Schwab-Stone, 2004). When the relation of self-concept to delinquency is examined without disentangling low and high levels of self-concept, the possible roles of low and high self-concept might cancel each other out (Salmivalli, 2001).

7.1.2 Domain Differences

Over and above these four predictions, it may be argued that the association between delinquent behaviour and self-concept is dependent on the domain

of self-concept. Swann et al. (2007) proposed that both the self-concept (as a general concept) as well as its metacognitive aspects (that is, different relevant domains) are important (see also Marsh & Craven, 2006; Rosenberg, Schooler & Schoenbach, 1989). Hence, it may be that the association between self-concept and delinquent behaviour in young adults differs according to the domain of self-concept. Carroll, Houghton, Wood, Perkins and Bower (2007) used three dimensions of self-concept to study the relationship with level of involvement in delinquent activities. They found that students highly involved in delinquent activities reported significantly lower classroom, peer, and confidence self-concepts. Vermeiren et al. (2004) used different domains and found that a low self-concept regarding family climate and school competence and a high self-concept regarding friendships were significantly related to juvenile delinquency.

An important caveat is that self-concept might not be a one-dimensional construct; furthermore, different domains of self-concept may be differently related to delinquent behaviour. In the current study we assessed self-concept as a set of domains domains judged to be relevant for the age group of the respondents including a global self-worth scale, as developed by Harter (1983, 1987, 1990, 1999). See Appendix A for a description of all the relevant domains used in the current study. Since this study was the first to use all 13 domains for young adults in examining the association with delinquent behaviour, we approached the domain differences in an exploratory way.

7.1.3 Gender Differences

Besides domain-specific differences, gender differences may also exist in the relationship between self-concept and delinquency. Piko et al. (2005), for example, found that female delinquency was associated with a high self-concept, whereas male delinquency was not associated with level of self-concept. Vermeiren et al. (2004) concluded that for both sexes, low self-concept regarding academic competence showed the strongest association with delinquent behaviour, but the variance explained was consistently higher

for men than for women. Gender differences in this association were also found by Diamantopoulou et al. (2008) and Webster et al. (2007). We thus explored gender differences and have reported findings for men and women separately where possible.

7.1.4 Main Aims of this Study

It is far from clear how the association between self-concept and delinquent behaviour can best be described. Studies have suggested that both low self-concept and high self-concept may be related to delinquent behaviour, while other studies have failed to find any relationship at all. One way to resolve these discrepancies might be to propose that either a low or a high self-concept could be related to delinquency. Furthermore, different predictions might be valid for different domains of self-concept and for men and women separately. In this study we had two aims: (1) to explore the association between self-concept and delinquent behaviour and possible gender differences using a regression approach and (2) to adopt a group-based approach in order to find out whether subgroups exist in accordance with the predictions previously introduced. To do this we used confirmatory latent class analyses (Hoijtink & Boom, 2008; Laudy, Boom & Hoijtink, 2005) to test the competing predictions and also to determine whether the findings would differ based on self-concept domain or participant gender.

7.2 Method

7.2.1 Participants

Participants were 899 young Dutch adults (male = 34%) with ages ranging from 18 to 24 years ($M = 20.51$; $SD = 1.81$). Of the participants 87.5% were currently following some level of education. Among these, 29.7% were in intermediate vocational education, 25.1% in higher vocational education and 44.1% at university (For a detailed explanation of the Dutch educational

system, see De Graaf, De Graaf & Kraaykamp, 2000). Forty-five percent of the participants worked more than 10 hours per week and 27% worked more than 30 hours per week. The participants were living in different parts of the Netherlands with divergent degrees of urbanization and were from different ethnic groups. Five percent of the sample was born outside the Netherlands.

7.2.2 Procedure

The total sample was recruited by Bachelor students in Developmental Psychology and Clinical & Health Psychology at Utrecht University, the Netherlands. E-mail (50%) and paper- and-pencil versions of the questionnaires were used, most of them distributed at schools, factories, and stores in towns and cities in different parts of the Netherlands.

7.2.3 Measures

Delinquent behaviour. To assess self-reported delinquent behaviour we used a questionnaire that consisted of five categories of offences and violations: property offences, crimes of violence, traffic violations, drug-related offences, and vandalism (see Table 7.1). The questionnaire was based on a study of delinquent behaviour commissioned by the Dutch government (Jennissen & Blom, 2005). It was pretested in a pilot study in which 72 female and 53 male students participated (mean age = 21; $SD = 0.19$). Participants were Dutch students in pre-vocational education (40%) and higher education. The first version of the questionnaire consisted of 25 items, 5 questions per category.

Each item (e.g. stealing money in the home) consisted of the question: "Did you engage in this behaviour?" scored on a three-point scale (never, just once, often). Participants in the pilot study were asked to provide feedback on all items. On the basis of this feedback we deleted some of the items (e.g., committing tax fraud, joy-riding, discriminating against a person), and excluded one item with no variance (robbing someone). To select the final set of items we performed an exploratory factor analysis in Mplus 4.1 (Muthén & Muthén, 2007). Items per category with factor loadings lower

Table 7.1: Frequency Statistics for the Delinquency Items (n =895)

		Answering categories		
Categories of delinquent behaviour	Items	Never	Just once	Often
Crimes of violence				
	Using physical aggression	96.5%	3%	0.4%
	Threatening someone	99.4%	0.2%	0.3%
Traffic violations				
	Driving under the influence of alcohol	88.3%	9.3%	2.5%
	Driving through red traffic lights	97.5%	2.5%	0%
Drug-related offences				
	Using soft drugs	93.3%	4.8%	1.9%
	Using hard drugs	97.2%	1.7%	1.1%
Vandalism				
	Destroying public property	90.8%	8.6%	0.6%
	Setting fire to public property	92.3%	5.9%	1.8%
Property offences				
	Stealing from supermarket	93.5%	5%	1.5%
	Stealing a bicycle	93.4%	6.0%	0.6%
	Selling stolen goods	88.8%	7.3%	3.9%

Note Percentages add up to 100% in rows.

than .30 were omitted. For the final questionnaire, items that decreased the Cronbach's alpha were also deleted. In total, 11 items were selected with a final Cronbach's alpha of .85.

For the final data set, the 11 items described above were used (see Table 7.1 for the frequency statistics). To evaluate model fit, items per category were included as categorical variables in a confirmatory factor analysis (CFA) using Mplus. Model comparison indices indicated a better fit for a one-factor model (BIC= 6583.376) compared to a two-factor model (BIC = 6584.664). The Cronbach's alpha was .74.

Self-concept. To assess the multiple domains of self-concept we used the Self-Perception Profile for Young Adults (SPP-YA, Dutch version; Visser-Van Balen et al., 2007). This questionnaire is based on the Dutch version of the Self-Perception Profile for Adolescents (Treffers et al., 2002). The SPP-YA consists of 60 items to assess perceived competencies and global self-worth in young adults aged 18 to 24. The items are divided into 13 scales: a global self-worth scale (five items) to assess general self-concept and 12 perceived competence scales for each domain that is important for young adults (Harter, 1990, 1999). See Appendix A for a description of the scales . The items have a 4-point answering format and in every item two alternatives are presented: "Some are X" but "Others are not X". An example is "Some young people are not productive in their work BUT other young people are very productive in their work". Another example is "Some young people are better then me at sport BUT other young people are not better then me at sport". The respondent is asked to judge to which group he/she belongs, and to mark whether this is 'sort of true' or 'really true' for him/her. Items per scale were included as categorical variables in a confirmatory factor analysis using Mplus. Using a confirmatory factor analysis, we replicated the factor structure as was found by Visser-Van Balen et al. (2007) who used exploratory factor analysis. Fit indices for our data indicated a good fit of a 13-factor model (CFI = .94; TLI = .93; RMSEA = .05; Cronbach's alphas between .66 and .86).

7.2.4 Strategy of analysis

To examine our research question, we adopted two different approaches. First, we explored the data using a MANOVA to test for gender differences on all variables, together with a multiple group regression analysis where delinquent behaviour was predicted by the global self-worth scale and the 12 domain-specific scales (See Appendix A). We used the mean scores for each variable. The delinquency variables were Poisson distributed and we used the count option in Mplus to run the regression analysis. We also ran a multi-group model to investigate gender differences in the relationship between level of self-concept and delinquency. The Akaike's Information Criterium (AIC) (Akaike, 1973) and the Bayesian Information Criterium (BIC) (Schwarz, 1978) were used to select the best model in terms of model fit and model complexity. Model 1 constrained regression coefficients to be equal for men and women, whereas Model 2 did not have this constraint.

To examine whether there were subgroups that differed in level of self-concept and delinquency, we used confirmatory latent class analysis (C-LCA: Hoijtink, 1998, 2001; Hoijtink & Boom, 2008; Hoijtink & Molenaar, 1997; Laudy, Boom & Hoijtink, 2005). Because it is a fairly new technique, we devote some space here to introducing the methodology. The main goal of traditional LCA is to determine groups of persons with similar item responses. However, we had specific expectations not only about the number of groups, but also about the answering patterns in each subgroup. In other words, we had specific expectations about the ordering of the latent class probabilities. Our analytic needs led us to use the software developed by Laudy et al.. The main elements of this methodology are introduced below and are described in more technical detail in Appendix B; however for more detailed information about this method, see also Hoijtink and Boom (2008). Comparisons between more traditional ways of analyzing data and confirmatory Bayesian approaches are described in (Van de Schoot, Hoijtink, Mulder et al., 2010) and Kuiper and Hoijtink (2010), for example.

The first step in Bayesian model selection is to specify the inequality constraints among the latent class probabilities of interest. For example, suppose we expect there to be two groups of young adults and that these groups can be distinguished in terms of their levels of delinquency and self-concept. Assume that the probability of giving the answer 'yes' to questions regarding any of the offences is higher in group 1 than in group 2. We would then expect that the probability of choosing the 'disagree' alternative on the self-concept items, indicating a low self-concept, would be lower in group 1 than in group 2. This would result in two groups of young adults: one group with a high level of delinquent behaviour in combination with a low self-concept, and a group with a low level of delinquent behaviour in combination with a high self-concept.

In this way, a set of constraints could be constructed for each prediction. We evaluated four different models based on the four predictions outlined in the introduction. However we also explored a number of additional models to ascertain whether our results were trustworthy. In Appendix B, we describe these models in statistical terms. To summarize the four main models were as follows: (1) a first group of young adults with *high* levels of delinquency in combination with *low* levels of self-concept and a second group with *low* levels of delinquency in combination with *high* levels of self-concept; (2) a first group with *high* levels of delinquency and a second group with *low* levels of delinquency where no constraints are imposed on the latent class probabilities for the items of the self-concept scale; (3) a first group with *high* levels of delinquency in combination with *high* levels of self-concept and a second group with *low* levels of delinquency in combination with *low* levels of self-concept; (4) two groups with *high* levels of delinquency, one with *low* levels of self-concept and one with *high* levels of self-concept. For the global self-worth scale several alternatives were explored and the best solutions were validated on the 12 other domains.

After confronting the set of inequality constraints for each model with the data, the software provides each model with the marginal likelihood value.

This is a model selection tool which quantifies the degree of support for the constraints imposed on the latent class probabilities provided by the data. It has a close link with traditional model criteria such as the AIC and BIC that also can be used to evaluate models in latent class analyses. However, in contrast to Bayesian model selection, these traditional criteria are as yet unable to deal with inequality constraints specified between latent class probabilities. The easiest way to interpret these marginal likelihood values is to translate them into Posterior Model Probabilities (PMP). These PMPs reflect the probability that the prediction at hand is the best of the set of predictions under consideration. The hypothesis with the highest PMP receives most support from the data.

The software used in this paper, which is described in Laudy, Boom and Hoijtink (2005) and in Hoijtink and Boom (2008), is based on two assumptions: (1) no missing data is allowed and (2) only dichotomous variables are allowed.

With respect to the first assumption, most of the data was fully observed (84.7%), 8.2% of the participants had only 1 missing value, 2.8% had two missing values and 4.3% had more missing values. Of these latter cases, four questionnaires were completely blank and were omitted from the data set. In total, a percentage of less than 1% of all the data points was missing. Missing items were imputed using the MICE package (Van Buuren & Oudshoorn, 1999, 2005). MICE can be used to impute categorical data. Further analyses were executed on the sample of imputed cases ($n = 895$).

To deal with the second assumption (only dichotomous variables are allowed) we chose to dichotomies the items instead of the scale scores. The items for the delinquency scale were measured on a three-point scale whereas the items for the self concept were measured on a four-point scale. Dichotomizing these items is more straightforward than dichotomizing a scale score because much information would be lost if we performed the latter operation. In addition, using several items instead of one single scale provided us with more detailed information on the subgroups.

The items were dichotomized by combining answering categories. For the delinquency items we recoded the answering categories into: 0 = not engaging any delinquent acts; 1 = once or often engaging in delinquent acts. It appeared that, on average, 48% of the sample had committed offences during the previous year. Of those, 32% committed three or more offences. The items for the self-concept scale were dichotomized as follows: If respondents chose the 'agree' alternative (either 'sort of true' or 'really true'), this was coded 1 (i.e., high); it was coded 0 (i.e., low) if they chose the 'disagree' alternative.

7.3 Results

7.3.1 Regression and MANOVA Results

To evaluate whether mean delinquency level and the 13 self-concept scales differed by gender, a multivariate analysis of variance (MANOVA) was performed. Results of evaluation of assumptions were satisfactory. Significant multivariate main effects were found ($F(14, 880) = 19.44$; $p < .002$, partial $\eta 2 = .24$) . Univariate analyses showed significant effects for some of the scales. See Table 7.2 for means and standard deviations for the total group and for men and women separately.

A noteworthy finding is that men scored higher on delinquency than women, as well as on global self-worth, and physical appearances. The opposite pattern held for athletic competence, behaviour and consequence, close friendships, nurturance, household management, and sense of humour, on which women scored higher.

To evaluate which domains were related to delinquent behaviour and whether these differed for men and for women, we performed a multiple group Poisson regression in Mplus. The results are presented in Table 7.3, and we will limit the discussion of the results to the most relevant findings.

First, we compared a model which did not allow for different estimates for men than for women to a model which did allow for different estimates

based on gender. It appeared the latter model had a better trade off between model fit and model complexity (ΔAIC = 342; ΔBIC = 275), suggesting that the association between self-concept and delinquent behaviour differed between men and women. This can be illustrated by, for example, the regression coefficient for close friendships. This coefficient was almost zero (B = .05) for the constrained model, but in the gender-based model the coefficient was .42 for women and -.26 for men. In addition, results revealed a different set of significant predictors for men (close friendships, job competence, romantic relations) than for women (scholastic competence, social competence, physical appearance, nurturance, household management). Additionally, there were some predictors common to both sexes (athletic competence, behaviour and conscience, sense of humour). Relationships with parents appeared to be the only non-significant predictor for both men and women. In addition, the direction of the effect for athletic

Table 7.2: Mean scores for Delinquent and Self-Concept Scales with Standard Deviation Between Brackets (n=895)

	Male	Female
Delinquency scale*	1.73 (2.39)	0.36 (0.84)
Self-worth*	2.44 (0.52)	2.36 (0.52)
Scholastic Competence	2.11 (0.48)	2.04 (0.48)
Athletic Competence*	2.20 (0.51)	2.30 (0.48)
Social Acceptance	2.20 (0.51)	2.34 (0.42)
Physical Appearance*	2.21 (0.48)	2.12 (0.48)
Behaviour and Conscience*	2.11 (0.46)	2.29 (0.46)
Close Friendships*	2.65 (0.51)	2.79 (0.44)
Job Competence	2.33 (0.56)	2.34 (0.49)
Nurturance*	2.01 (0.44)	2.15 (0.39)
Household Management*	2.31 (0.68)	2.56 (0.69)
Romantic Relations	2.14 (0.50)	2.17 (0.50)
Sense of Humour*	2.38 (0.46)	2.43 (0.44)
Relationship with Parents	2.38 (0.46)	2.42 (0.44)

* denote significant gender differences ($p < .05$)

competence was different for men (negative) and women (positive), as shown in Table 7.3.

7.3.2 Bayesian Results: Higher or Lower Overall Self-concept?

We started the Bayesian analyses with an exploratory approach and investigated 4 models without any constraints imposed between the latent class probabilities. Each of these 4 models was an unconstrained model and the models differed only in the number of latent classes (M01, . . ., M04 with 1 to 4 latent classes respectively). The results are presented in the top panel of Table 7.4.

As can be seen, a model with 4 latent classes has the lowest marginal likelihood value and the highest PMP value (.99). The gender-specific analyses produced equivalent results. A model with 5 latent classes resulted

Table 7.3: Multiple group Poisson regression of the 13 subscales of self-concept on delinquent behaviour for men and women seperately ($n = 965$)

	Women		Men	
	B	SE	B	SE
Self-worth	.10	.15	-.08	.09
Scholastic Competence	.32*	.15	.15	.09
Athletic Competence	.36*	.14	-.24*	.09
Social Acceptance	-.48*	.17	.12	.12
Physical Appearance	-.14*	.15	-.15	.10
Behaviour and Conscience	-.45*	.15	-.29*	.09
Close Friendships	-.26	.15	.42*	.09
Job Competence	-.22	.15	.22*	.09
Nurturance	.35*	.19	-.003	.10
Household Management	-.27*	.10	-.14*	.07
Romantic Relations	.15	.14	.37*	.09
Sense of Humour	.50*	.15	.23*	.09
Relationship with Parents	.21	.17	.08	.10
Intercept	-1.24	.81	-1.32	.43

* denote significant gender differences ($p < .05$)

Table 7.4: Marginal Likelihood and Posterior Model Probabilities for Global Self-Worth Scale (n=895).

	Total sample		Men		Women	
Model	ML[1]	PMP[2]	ML[1]	PMP[2]	ML[1]	PMP[2]
M01 (1 latent class, no constraints)	-5182.41	.00	-2239.27	.00	-2721.77	.00
M02 (2 latent classes, no constraints)	-4785.25	.00	-2103.23	.00	-2484.68	.00
M03 (3 latent classes, no constraints)	-4610.96	.00	-2036.79	.00	-2441.10	.27
M04 (4 latent classes, no constraints)	**-4577.68**	**.99**	**-2014.66**	**.99**	**-2440.50**	**.73**
M1: (high delinquency & low self-concept and low delinquency & high self-concept)	-4973.59	.00	-2156.58	.00	-2456.58	.00
M2: (high levels of delinquency and low levels of delinquency; no constraints for self-concept)	-4956.17	.00	-2149.03	.00	-2484.58	.00
M3: (high delinquency & high self-concept and low delinquency & low self-concept)	-4955.92	.00	-2106.11	.00	-2484.58	.00
M4a: (one group that has no constraints)	-4612.49	.00	-2039.04	.00	-2443.99	.00
M4b: (one group that has levels of self-concept that are in between the high and low levels of the two offender groups)	-4650.18	.00	-2043.98	.00	-2452.66	.00
M4c: (two groups: one group with high levels and one group with low levels of self-concept)	-4576.93	**.99**	**-2014.66**	**.99**	**-2433.99**	**.99**

[1] ML= Marginal Likelihood

[2] PMP= Posterior Model Probability

Table 7.5: Posterior estimates for the latent class memberships for the 4-group solution without any constraints

	Class 1		Class 2		Class 3		Class 4	
Delinquency items	CSP[1]	95% C.I.[2]	CSP[1]	95% C.I.[2]	CSP[1]	95% C.I.[2]	CSP[1]	95% C.I.[2]
1. Using physical aggression	0.00	0.00 - 0.01	**0.43**	0.25 - 0.60	**0.22**	0.13 - 0.33	0.01	0.00 - 0.03
2. Threatening someone	0.04	0.02 - 0.06	**0.52**	0.35 - 0.72	**0.37**	0.25 - 0.50	0.04	0.02 - 0.06
3. Driving under the influence of alcohol	0.06	0.04 - 0.08	**0.48**	0.31 - 0.64	**0.42**	0.29 - 0.55	0.11	0.08 - 0.14
4. Driving through red traffic lights	0.07	0.05 - 0.10	**0.51**	0.36 - 0.69	**0.41**	0.29 - 0.55	0.08	0.05 - 0.10
5. Using soft drugs	0.00	0.00 - 0.01	**0.23**	0.12 - 0.38	**0.21**	0.12 - 0.33	0.02	0.01 - 0.03
6. Using hard drugs	0.03	0.01 - 0.05	**0.69**	0.51 - 0.86	**0.49**	0.35 - 0.64	0.05	0.03 - 0.08
7. Destroying public property	0.02	0.01 - 0.03	**0.27**	0.14 - 0.42	**0.07**	0.02 - 0.13	0.01	0.00 - 0.02
8. Setting fire to public property	0.03	0.01 - 0.04	**0.32**	0.18 - 0.48	**0.34**	0.22 - 0.47	0.06	0.03 - 0.08
9. Stealing from supermarket	0.04	0.02 - 0.07	**0.42**	0.26 - 0.58	**0.28**	0.16 - 0.40	0.03	0.01 - 0.04
10. Stealing a bicycle	0.00	0.00 - 0.01	**0.07**	0.01 - 0.15	**0.04**	0.01 - 0.09	0.01	0.00 - 0.01
11. Selling stolen goods	0.03	0.01 - 0.04	**0.53**	0.37 - 0.70	**0.29**	0.18 - 0.42	0.04	0.02 - 0.06
Global self-worth items								
12. Disappointment with oneself	0.17	0.14 - 0.20	**0.60**	0.43 - 0.76	0.24	0.13 - 0.36	**0.38**	0.33 - 0.42
13. Satisfied with life	0.32	0.28 - 0.36	**0.78**	0.64 - 0.94	0.40	0.27 - 0.53	**0.68**	0.63 - 0.73
14. Satisfied with oneself	0.15	0.12 - 0.19	**0.91**	0.76 - 1.00	0.16	0.05 - 0.28	**0.90**	0.86 - 0.94
15. Satisfied with who you are	0.19	0.15 - 0.22	**0.83**	0.68 - 0.95	0.21	0.09 - 0.32	**0.85**	0.81 - 0.89
16. Satisfied with how you are	0.08	0.05 - 0.12	**0.90**	0.75 - 0.99	0.20	0.08 - 0.32	**0.85**	0.81 - 0.90

Note: latent class probabilities in bold are interpreted as high and latent class probabilities in non-bold as low.

[1] CSP = Class-specific probability

[2] 95% credibility intervals

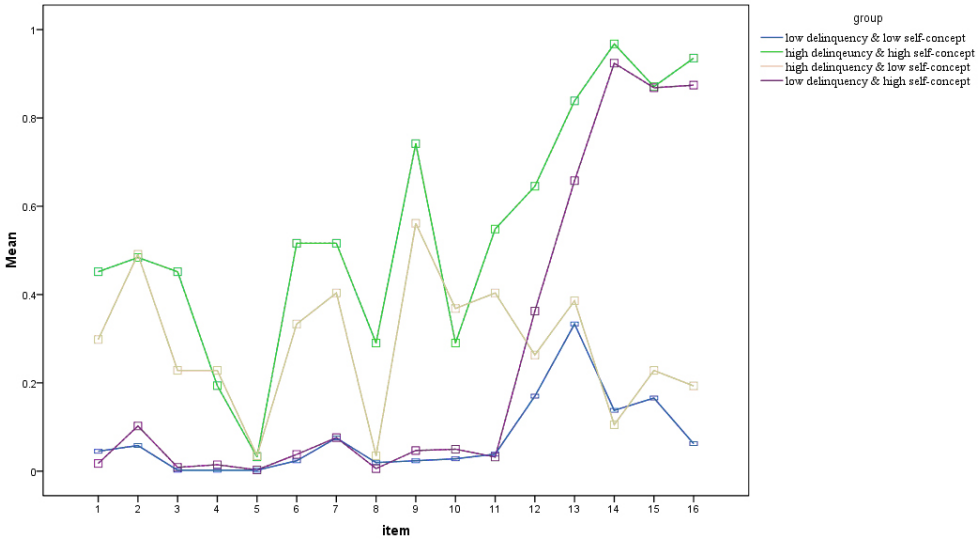


Figure 7.1: Means on the delinquency items (1-11, see Table 7.5) and the self-concept items (12-16, see Table 7.5) for the 4-group solution without any constraints.

in such small sample sizes for some of the classes that this model cannot be interpreted from a theoretical point of view. The unconstrained latent class probabilities for the total sample are shown in Table 7.5. Inspecting these unconstrained latent class probabilities provided an initial test of our competing predictions. Figure 7.1 displays (in graphical form) the mean score for each group on all items used in analyses. Combining the information on means presented in the figure and the latent class probabilities presented in Table 7.5, the classification of the four groups can be interpreted as follows: There was a large group (51%) with the lowest levels of delinquency and the lowest levels of self-concept and a moderate-sized group (38%) with low levels of delinquency and high levels of self-concept. Furthermore, there was a small group (7%) with high levels of delinquency and low levels of self-concept. Finally, there was a small group (4%) with the highest levels of delinquency and high levels of self-concept.

A more traditional paper using LCA would stop here, and the latent class probabilities would be interpreted. In this project, however, we went

one step further and formulated inequality constraints between the latent class probabilities according to our predictions, which enabled us to test our predictions directly. We describe this process in the remainder of this section.

The bottom panel in Table 7.4 shows the marginal likelihood values and PMPs for the models 1-3 as described in the strategy of analyses section. For prediction 4 we evaluated for the low-delinquency group what the best alternative was regarding the constraints in the level of self-concept. Several alternative models were explored: (4a) one group that had no constraints; (4b) one group with levels of self-concept that were in between the high and low levels of the two offender groups; (4c) two groups: one group with high levels of self-concept and another with low levels of self-concept.

Table 7.4 shows that the PMP for prediction 4c was .99. Remember that a PMP value is on a probability scale and runs between 0 and 1. This PMP value indicated very strong support for the set of inequality constraints imposed on the latent class probabilities. In other words, there was evidence for two groups of delinquents and non-delinquents: one group with high levels and one group with low levels of self-concept. The gender-specific analyses produced equivalent results, although the support for model 4c for the women was only moderate. This result thus provided strong support for the hypothesis that either a low or a high self-concept is associated with delinquent behaviour. As such, there seem to be two distinct groups of young adults who commit offences: one with a high self-concept and one with a low self-concept.

The four theoretically based models (including the model that appeared to be most adequate for prediction 4, i.e. Model 4c) were used for evaluating the domain-specific analyses. We ran the software for these four models for each domain and for men and women separately. Theses analyses resulted in 96 marginal likelihood values and are not presented in the current paper for reasons of space. The results, however, clearly pointed to Model 4c as the best model, since its PMP value exceeded .99 for each and every domain and for both men and women.

7.4 Discussion

The present research tested competing models of the relationship between self-concept and delinquent behaviour using Bayesian methodology. This association has been debated in the literature, but there is as yet no consensus regarding their precise relationship. New Bayesian analysis tools allowed a new form of differentiated relative evaluation.

One thing seemed to be evident in the present study: Self-concept is related to delinquent behaviour, either positively or negatively. Our preliminary results from the MANOVA and regression analyses showed that it clearly depends on the domain whether the relationship between self-concept and delinquency is positive or negative. These differences became even more pronounced when men and women were examined separately. Men and women appeared to differ both in the strength and the direction of the association. This indicates that, depending on the domain of self-concept and on gender, there is a negative or positive relationship between self-concept and delinquent behaviour in young adults.

The latent class results went one step further and revealed that for all the domains, both high-delinquent and non-delinquent men and women fall into two groups: those with high levels of self-concept and those with low levels of self-concept. This finding could help to explain some of the inconsistent results of previous studies on the link between self-concept and delinquency. It is possible that studies that found a positive association between the two were correct, but so were studies that found a negative association. Our study showed that there are two groups of delinquents, those with a high self-concept and those with a low self-concept.

These are important findings, since programmes developed to reduce delinquent behaviour often aim to improve self-concept (Mason, 2003). These results suggest that programmes focusing on self-concept should first consider the level of each individual's self-concept before making decisions about

whether to boost or temper it. Ideally, this should be done in each domain, as the level of an individual's self-concept may differ between domains.

The two groups of delinquents exist for both genders and in every domain of self-concept. We will highlight some of the domains for which this finding was quite remarkable. For scholastic and job competence, for example, a high self-concept seems unexpected. Delinquent young adults generally come from low-SES households and score low on school achievement (Thornberry & Krohn, 2003), two important factors related to low occupational prospects (J. W. Lynch, Kaplan & Salonen, 1997). Furthermore, it is known that offenders usually have poor jobs (Laub, Nagin & Sampson, 1998). According to the strain theory (Agnew, 1992), people who have fewer chances of attaining high socio-economic status, perhaps because of poorer job prospects, might engage in delinquent behaviour in order to reach their goals. It is surprising, then, that some delinquent men and women considered their scholastic and job prospects in such a positive light. Perhaps their views of their career competences give them unrealistically positive views of their own futures.

Furthermore, some of the delinquent young adults rated their competence regarding behaviour at a low level. This could imply, for example, that some delinquents knew quite well that they should not engage in such 'bad' behaviour, but it did not keep them from doing it, possibly because they did not realise that such behaviour might bring negative consequences for them in the future. Although this statement was not investigated in the current study, previous literature indeed shows that delinquents do not think about the negative effects that criminal behaviour may have on important aspects of life, like their futures (Modecki, 2009). This is related to our finding that delinquent persons viewed their job prospects positively. Although delinquent young adults seem to know that it is not a good thing to be delinquent, it is possible that they are unconcerned with their behaviour or do not see its likely consequences.

Previous literature suggests that a good relationship with parents can restrain juveniles from delinquent behaviour since they do not want to jeopardise this bond. If the parent-child relationship is of low quality, this incentive disappears. Such juveniles can become delinquent because they are not directed by their emotional attachment to their parents (Gottfredson & Hirschi, 1990). However, our study showed that juveniles who have a good relationship with their parents can also become delinquent. Combining this result with the aforementioned literature, we could speculate that it may be that parents are no longer so important for our respondents, since they are already in young adulthood. Close friends and romantic partners might have a more pronounced impact on them. However, the level of self-concept in these domains could also be high for delinquent young adults. Delinquents with a high self-concept in terms of relationships with close friends and romantic partners might have friends and romantic partners who are also delinquent. A good relationship then might then actually increase the odds of turning to delinquent behaviors.

Another remarkable finding was that young adults who perceive themselves as nurturant could also be delinquents. Nurturance is closely linked to empathy (Batson, Lishner, Cook & Sawyer, 2005), and people who have high levels of empathy are usually less likely to offend than those with low levels of empathy (Jolliffe & Farrington, 2004). This could be because people who share someone else's feelings caused by their own delinquent behaviour may be less inclined to engage in this behaviour (see also, Feshbach, 1975). The same goes for nurturance: People who like to take care of others will probably not indulge in (delinquent) behaviour that can cause harm to others. An explanation for our finding of delinquents with high levels of nurturance could, for example, be that these delinquents only committed victimless offences. It would therefore be interesting to examine the levels of self-concept in different types of offenders (i.e., violent, property, public order offences).

This study has certain limitations. Although we found that low and high self-concept are related to delinquency, we could not clarify using our data whether delinquency is related, more specifically, to a positive self-concept, an inflated self-concept or to both. Future research should focus on the question of whether it is sufficient to have a positive self-concept to become delinquent or whether it is necessary to have an inflated view of the self. Additionally, our results might not be valid for more serious delinquents but can only be generalised to a relatively 'normal' population. It would therefore be interesting to repeat this study with a population of young adults with a history of serious delinquent behaviour or with a population of imprisoned young adults. Furthermore, it could be argued that translating our variables into dichotomous variables leads to information loss. However, Bayesian software that can deal with inequality constraints in Latent Class Analysis for other than dichotomous variables is not yet available. Although this study provides evidence for the association of delinquency with both low and high self-concept, we did not examine the causal effects of self-concept on delinquent behaviour. Longitudinal studies are needed to verify whether the relationship is causal or not. Furthermore, the process and direction of influence should be analysed.

A strength of this study was that the presented predictions were pitted against each other using Bayesian model selection. This technique enabled a direct comparison of predictions about whether delinquent young adults have a high or a low self-concept, a distinction that to our knowledge has never been made before. Moreover, to date, little research has focused on how gender may account for differences across multidimensional self-concept constructs.

In sum, by applying the new Bayesian analysis tools, the relative plausibility of complex alternative hypotheses regarding the association of level of self-concept and delinquent behaviour could be clearly and convincingly evaluated and determined. Knowing how and why self-concept affects delinquent behaviour can be of importance in prevention and

intervention programmes for delinquents. The current study showed that delinquents, male or female, can either have a low or a high self-concept. Programmes should therefore not necessarily focus on improving delinquents' self-concept, which is often the purpose of current intervention programmes (Mason, 2003). By gaining knowledge about the process of influence, programmes designed to prevent and intervene in delinquent behaviour can be better geared to the development of delinquent behaviour.

APPENDICES

A Overview of the Self-concept scales

- *Global Self-worth.* Self evaluations that represent global characteristics of the individual (e.g. I am a worthwhile person), also referred to as self-esteem or general self-concept (5 items).

- *Scholastic Competence.* Self evaluations related to academic expectations, such as being good at school work, knowing the right answers at school, finishing homework quickly (5 items).

- *Athletic Competence.* Self evaluations related to sporting activities, such as being good at different sports, being good at outdoor sports (5 items).

- *Social Acceptance.* Self evaluations related to social networks, such as being popular, having friends and being liked by others (5 items)

- *Physical Appearance.* Self evaluations related to own physical appearance, such as liking your own body, being happy with body composition, being attractive (5 items).

- *Behaviour and Conscience.* Self evaluations related to committing offences, such as abiding by the rules, doing things that get you into trouble and behaving correctly (5 items).

- *Close Friendships.* Self evaluations related to having close friends to share things with, such as engaging in activities together, sharing secrets, and having close friends (5 items).

- *Job Competence.* Self evaluations related to job experiences, such as being good at/satisfied with/productive in your work (4 items).

- *Nurturance.* Self evaluations related to caring, such as being good at nurturing others, providing adequately for the needs of others, and liking to support others (6 items).
- *Household Management.* Self evaluations related to managing a household, such as running a household smoothly, being organised and efficient (4 items).
- *Romantic Relations.* Self evaluations related to intimate relations, such as being scared of unrequited love, having difficulty establishing romantic relations (4 items).
- *Sense of Humour.* Self evaluations related to humour, such as laughing at yourself, having a good sense of humour, and laughing at others' jokes (4 items).
- *Relationship with Parents.* Self evaluations related to interaction with parents, such as acting naturally around your parents and being yourself with parents (3 items).

B Technical Details of the Bayesian Methodology

In this paper confirmatory latent class analysis (C-LCA) is used to analyse our data (Hoijtink & Boom, 2008; Hoijtink, 2001; Hoijtink & Molenaar, 1997; Laudy, Boom & Hoijtink, 2005; Laudy, Zoccolillo et al., 2005). In this Appendix we present an introduction to the methodology behind the software of Laudy et al. We refer more interested readers to the book chapter of Hoijtink and Boom (2008) for a more technical description of the method. The second part of this appendix is a summary of the book chapter.

B.1 Bayesian Model Selection

For readers not familiar with Bayesian statistics in general we refer to S. Lynch (2007). For readers not familiar with Bayesian model selection

we refer to Van de Schoot, Hoijtink, Mulder et al. (2010) or Hoijtink, Klugkist and Boelen (2008). Here we present a short introduction to the main components of Bayesian statistics.

An evaluation of a hypothesis of interest according to Bayes' formula puts together three components:

1. Component 1: a set of K hypotheses of interest, H_k $(i = 1, \ldots, K)$, is specified. For each hypothesis an 'a priori model probability' $p(H_k)$ is determined, as a measure of the degree of belief that H_k is true, before inspection of the data.

2. Component 2: empirical observations, or data D, are considered regarding the degree of support they provide to the different hypotheses. For each hypothesis the 'likelihood' $p(D|H_k)$ is determined as the probability of that the empirical data, given H_k is true.

3. Component 3: The a priori probabilities $p(H_k)(k = 1, \ldots, K)$ and the likelihoods $p(D|H_k)$ are combined to determine the 'marginal likelihoods' $p(H_k) \times p(D|H_k)$ whose relative values give the relative support for each hypothesis after observing the data.

Then, Bayes' formula is given by

$$p(H_k|D) = \frac{p(H_k) \times p(D|H_k)}{\sum_i^n p(H_k) \times p(D|H_k)} \ .$$

B.2 Confirmatory Latent Class Analysis

For a more detailed introduction of C-LCA we refer to Hoijtink and Boom (2008). LCA, in general, is used to group responses of persons on items $(i = 1, \ldots, I)$ into latent classes $(j = 1, \ldots, J)$ such that persons with similar responses are assigned to the same class. Let π_{ij} indicate the unconditional probability that a person's latent class membership equals group j. Based on theoretical expectations, we can construct a model using

inequality constraints of the following types $\pi_{i,j} > \pi_{i',j'}$ and $\pi_{i,j} < \pi_{i',j'}$ for $i \neq i'$ and $j \neq j'$. The methodology has, un till now, been restricted to dichotomous data, so that $\pi_{i,j}$ indicates the probability of the response '1' on item i in class j. The restriction $\pi_{1,1} > \pi_{1,2}$ implies that the probability of the response '1' for item 1 is larger for the first latent class compared to the second latent class.

The models under investigation, as described in the strategy of analysis section, can be specified in terms of these item probabilities. For example, let the constraints for the delinquency item 'stealing money in the home' ($i = 1$) to be equal to $\pi_{1,1} > \pi_{1,2}$, indicating that for group 1 the probability of answering 'yes' to this question is higher than for group 2. The constraint for the self-concept item 'sense of humour' ($i = 16$) is $\pi_{16,2} > \pi_{16,1}$. Using this approach we can construct statistical hypotheses for each of the predications under investigation, as shown in Table 7.6.

For the computation of the marginal likelihood, a prior distribution needs to be specified. Note that this distribution is set to default in the software and cannot be changed. It is chosen to be non-informative for all combinations of parameter values allowed by a constrained model. Since prior information about the models is included in the prior distributions via inequality constraints, in that respect the priors are informative; however, the actual distribution used is not. The methodology employs a truncated prior distribution over the parameter space by: (a) assigning some non-informative probability distribution to the admissible parameter space where the inequality constraints imposed on the model hold and (b) assigning a zero prior probability to the parameter space for which these inequality constraints are violated. The actual specification of the prior for confirmatory LCA is a (truncated) Beta(1,1) distribution. The conjugate prior for the class weights is a Dirichlet distribution parameterised such that a priori all combinations of weight values are equally likely (see Hoijtink & Boom, 2007, for more details). The prior takes the value 1 if the model parameters are in accordance with the constraints imposed by the model and zero if not.

Since it is not trivial to obtain a sample from a multivariate posterior distribution, an MCMC procedure, the Gibbs sampler, is applied. The general algorithm is described by Gelfand et al. (1992) and the direct application to confirmatory LCA can be found in Hoijtink (1998). The algorithm renders samples from the joint posterior of the parameters by repeatedly sampling from the distribution of the parameter at hand, given all the other parameters. The naive way to do so is to sample from the correct (non-truncated) Beta distribution until a deviate is sampled that satisfies the constraints. However, this is fairly inefficient when only a small range of the distribution is admissible. Inverse probability sampling solves this problem.

The actual computation of the marginal likelihood is based on Kass and Raftery (1995) and is described in Hoijtink and Boom (2007, p.233). The marginal likelihood is computed using samples from the prior and posterior distribution. However, when a constrained model is compared with an inequality constrained model and this latter model is supported by the data (i.e., most or all inequality constraints imposed on $\pi_{i,j}$ are correct) then the sampling procedure will render approximately similar values for the constrained and unconstrained marginal likelihood values. This is due to the fact that the samples are obtained from the posterior distribution. Since the posterior distribution of the unconstrained model is fairly similar to that of a posterior distribution based on a perfect model, the values obtained from these distributions will be fairly similar. Stated otherwise, in this situation the marginal likelihood is biased against the constrained model. So, if the value for the marginal likelihood for the constrained model is close to the marginal likelihood value of the unconstrained model, the first model should be preferred. This can be confirmed by visual inspection of the unconstrained latent class membership probabilities to ascertain whether they fit the constraints imposed on the model.

Table 7.6: Inequality Constraints of the Models under Investigation

	Score on delinquent behaviour			Score on Self-concept		
	Item $i = 1$	. . .	Item $i = 11$	Item $i = 12$	. . .	Item $i = 16$
M0: $j = 1$	$\pi_{11,2}$	. . .	$\pi_{11,1}$	π_{12}	. . .	π_{16}
M1: $j = 1, 2$	$\pi_{1,1} > \pi_{1,2}$	. . .	$\pi_{11,1} > \pi_{11,2}$	$\pi_{12,1} < \pi_{12,2}$	. . .	$\pi_{16,1} < \pi_{16,,2}$
M2: $j = 1, 2$	$\pi_{1,1} > \pi_{1,2}$	. . .	$\pi_{11,1} > \pi_{11,2}$	$\pi_{12,1}$, $\pi_{12,2}$	. . .	$\pi_{16,1}$, π_{162}
M3: $j = 1, 2$	$\pi_{1,1} > \pi_{1,2}$	. . .	$\pi_{11,1} > \pi_{11,2}$	$\pi_{12,1} > \pi_{12,2}$	. . .	$\pi_{16,1} > \pi_{16,2}$
M4a: $j = 1, 2, 3$	$(\pi_{1,1}, \pi_{1,2}) > \pi_{1,3}$	. . .	$(\pi_{1,1}, \pi_{1,2}) > \pi_{1,11}$	$(\pi_{12,1}, \pi_{12,2}) > \pi_{3,12}$	. . .	$(\pi_{161}, \pi_{16,2}) > \pi_{3,16}$
M4b: $j = 1, 2, 3$	$(\pi_{1,1}, \pi_{1,2}) > \pi_{1,3}$	. . .	$(\pi_{1,1}, \pi_{1,2}) > \pi_{1,11}$	$\pi_{12,1} > \pi_{3,12} > \pi_{12,2}$	. . .	$\pi_{16,1} > \pi_{3,16} > \pi_{16,2}$
M4c: $j = 1, \ldots, 4$	$(\pi_{1,1}, \pi_{1,2}) > (\pi_{1,3}, \pi_{1,4})$	. . .	$(\pi_{11,1}, \pi_{11,2}) > (\pi_{11,3}, \pi_{11,4})$	$(\pi_{12,1}, \pi_{12,3}) > (\pi_{12,2}, \pi_{12,4})$	. . .	$(\pi_{16,1}, \pi_{16,3}) > (\pi_{16,2}, \pi_{16,4})$
M4d: $j = 1, \ldots, 4$	$\pi_{1,1}, \pi_{1,2}, \pi_{1,3}, \pi_{1,4}$	. . .	$\pi_{11,1}, \pi_{11,2}, \pi_{11,3}, \pi_{11,4}$	$\pi_{12,1}, \pi_{12,3}, \pi_{12,2}, \pi_{12,4}$	. . .	$\pi_{16,1}, \pi_{16,3}, \pi_{16,2}, \pi_{16,4}$

Note The constraints are specified as follows: $\pi_{ij} > \pi_{ij}$, where π is the probability of a 'yes' answer to item i for group j.

8

On the Progression and Stability of Adolescent Identity Formation. A Five-Wave Longitudinal Study in Early-to-middle and Middle-to-late Adolescence

Meeus, W., Van de Schoot, R., Keijsers, L.,
Schwartz, S. J. & Branje, S.

In press for *Child Development*

Abstract

This study examined identity development in a five-wave study of 923 early-to-middle and 390 middle-to-late adolescents thereby covering the ages of 12 to 20. Systematic evidence for identity progression was found: the number of diffusions, moratoriums and searching moratoriums (a newly obtained status) decreased, whereas the representation of the high-commitment statuses (two variants of a (fore)closed identity: "early closure" and "closure", and achievement) increased. We also found support for the individual difference perspective: 63% of the adolescents remained in the same identity status across the five waves. Identity progression was characterized by seven transitions: diffusion → moratorium, diffusion → early closure, moratorium → closure, moratorium → achievement, searching moratorium → closure, searching moratorium → achievement, and early closure → achievement.

8.1 Introduction

Erikson (1968) theorized that one of the main tasks for adolescents is to develop a coherent sense of identity. Marcia's (1966) identity status model has been one of the most important, and widely studied and utilized, elaborations of Erikson's views on identity formation. Marcia distinguished four identity statuses based on the amount of exploration and commitment the adolescent is experiencing or has experienced. Identity diffusion (D) indicates that the adolescent has not yet made a commitment regarding a specific developmental task, and may or may not have explored among different alternatives in that domain. Foreclosure (F) signifies that the adolescent has made a commitment without much prior exploration. In moratorium (M), the adolescent is in a state of active exploration and has not made significant commitments. Identity Achievement (A) signifies that the adolescent has finished a period of active exploration and has made a commitment based on this exploration. Notably, A. S. Waterman (1982) has proposed that adolescents move from diffusion toward achievement as they progress through adolescence. The present study was designed to test this developmental interpretation of the identity status model: are identity statuses stable individual dispositions or do they change over time? We used a five-wave longitudinal dataset to study identity formation from early to late adolescence (ages 12 to 20).

8.1.1 Identity Statuses: Individual Differences or Development?

In his original contribution, Marcia (1966) conceptualized the identity statuses in terms of individual differences: "as individual styles of coping with the psychosocial task of forming an ego identity" (p. 558). In this sense, the statuses represent different individual states or dispositions. Most identity researchers have adopted this perspective and view the statuses as stable individual dispositions. On the other hand, some writers have proposed that

the statuses constitute a developmental sequence. Indeed, A. S. Waterman (1982) proposed a "developmental hypothesis" of the identity status model. This hypothesis involves two assumptions: First, the development of identity has a direction: development represents ". . .changes in identity status that constitute progressive developmental shifts" (A. S. Waterman, 1982, p. 343). Such progressive development involves a movement away from diffusion and toward achievement. The second assumption is that progressive development also involves a specific pattern of transitions between identity statuses: from diffusion into foreclosure or moratorium, and from foreclosure and moratorium into achievement. Consequently, A. S. Waterman (1982) assumes that adolescents starting from diffusion move through foreclosure or moratorium and then into achievement. Most prominent among the identity transitions that Waterman proposes are D $\rightarrow$ M or D $\rightarrow$ F, F $\rightarrow$ M, and M $\rightarrow$ A.

Overviews of studies using identity status classifications have offered limited, but consistent, support for the first assumption within Waterman's developmental hypothesis. Meeus (1996) reported that, in 17 of 25 studies reviewed, the prevalence of achievers was higher in the older age groups, and the prevalence of diffusions was higher in the younger age groups. In their reviews, both Van Hoof (1999) and Berzonsky and Adams (1999) found progressive developmental trends in 7 out of 14 longitudinal studies, and across studies they found a higher prevalence of progressive than regressive shifts. Findings also revealed that, across studies, more than half of participants remained in the same identity status during the course of the study. In a recent meta-analysis of cross-sectional and longitudinal studies, Kroger (2007) found the prevalence of achievements to be about 1.5 times higher in emerging adults (ages 22 to 29) as compared to middle adolescents (aged 15), and the prevalence of diffusions to be about 1.3 times lower.

Support for Waterman's second assumption is very scarce. A limited number of longitudinal studies (Adams & Fitch, 1982; Dellas & Jernigan, 1987; Kroger, 1988; Meeus, Iedema, Helsen & Vollebergh, 1999; A. Waterman,

Geary & Waterman, 1974; A. Waterman & Goldman, 1976; A. Waterman & Waterman, 1971) have tested whether transitions out of some statuses (e.g., diffusion) are more prevalent than transitions into those statuses. The available findings suggest that, in cases where adolescents do change statuses, more adolescents move out of diffusion than into it (in about 30% of the studies), more into than out of achievement (in about 50% of the studies), more out of than into foreclosure (in about 22% of the studies), and more out of than into moratorium (in about 29% of the studies). Over-time stability of the identity statuses was 59%. These patterns suggest that identity status is more likely to remain stable than to change, and that when change does occur, this change tends to be reflected in movement out of diffusion, foreclosure and moratorium, and into achievement.

These findings are not inconsistent with Waterman's second assumption, but they also suggest that the assumption has yet to be subjected to rigorous empirical test. A full test requires that identity status changes are simultaneously tested in a transition table that incorporates all possible transitions between identity statuses; in a two-wave study with 4 identity statuses, this would require a test of a 4 by 4 transition matrix. The studies cited above only tested whether the number of movers into each status was different from the movers out of that status. As a result, systematic empirical tests of the sequential patterns of identity status transitions that Waterman hypothesized remain to be conducted, for instance, that the chances for the transition from diffusion to moratorium are greater than for the transition from diffusion to achievement.

The major aim of the present study is therefore to provide a systematic account of identity status transitions - both progressive and regressive - over time. The prevalence of transitions will also provide information regarding the extent to which identity statuses represent stable individual dispositions or change over time. We now proceed to presenting our conceptualization of identity formation, in terms of the processes that underlie the identity statuses.

8.1.2 A Dimensional Approach: Commitment, In-Depth Exploration, and Reconsideration

Our approach focuses on the management of commitments and posits three dimensions as underlying the process of identity formation. *Commitment* refers to strong choices that adolescents have made with regard to various developmental domains, along with the self-confidence that they derive from these choices. *In-depth exploration* represents the ways in which adolescents maintain their present commitments. It refers to the extent to which adolescents actively explore the commitments that they have already made by reflecting on their choices, searching for information about these commitments, and talking with others about them. *Reconsideration of commitment* refers to the willingness to discard one's commitments and to search for new commitments. Reconsideration refers to the comparison of present commitments with possible alternative commitments when the present ones are no longer satisfactory.

Our model assumes that identity is formed in a process of continuous interplay between commitment, in-depth exploration, and reconsideration. Our model holds that individuals enter adolescence with a set of commitments of at least minimal strength in important ideological and interpersonal identity domains, and that adolescents do not begin the identity development process with a 'blank slate'. The initial commitments build upon the way in which adolescents have resolved the earlier Eriksonian psychosocial crises in childhood and have developed the ego strengths of hope, will, purpose and competence (Erikson, 1968). Numerous studies have offered support for these assumptions. Markstrom, Sabino, Turner and Berman (1997) and Markstrom and Marshall (2007) found clear links between previous Eriksonian ego strengths and identity achievement. Moreover, a number of studies have suggested that early adolescents can possess strong identity commitments (Adams & Jones, 1983; Archer, 1982; Meeus et al., 1999).

During adolescence, individuals manage their commitments in two ways, through in-depth exploration and through reconsideration. In-depth ex-

ploration is a process of continuous monitoring of present commitments and serves the function to make them more conscious and to maintain them. Reconsideration is the process of comparing present commitments to alternative ones and deciding whether they need to be changed. Our model therefore focuses on the dynamic between certainty (exploration in depth) and uncertainty (reconsideration).

So, our model differs from Marcia's model in two respects. First, it differentiates Marcia's concept of exploration into in-depth exploration and reconsideration, which serve to maintain and change commitments, respectively. Secondly, our model has a stronger process orientation than Marcia's model. Marcia views commitments as the outcome of the process of exploration: after exploring various alternative commitments, adolescents choose one or more to which they will adhere. In contrast, our model assumes, as suggested by Grotevant (1987, p. 214), that commitments are formed and revised in an iterative process of choosing commitments and reconsidering them. In addition, our model assumes that adolescents regularly reflect upon their present commitments. In sum, our conceptualization of the process of identity formation implies a twofold management of present commitments. This conceptualization of in-depth exploration and reconsideration resembles the distinction between exploration in depth and exploration in breadth that was originally suggested by Grotevant (1987) and that has been applied by (Luyckx, Goossens & Soenens, 2006) in their dual-cycle model of identity formation.

By including commitment, exploration in depth, and reconsideration in our model, we sought to capture Erikson's (1968) dynamic of *identity versus identity diffusion*. Commitment and in-depth exploration on the one hand, and reconsideration on the other hand, are conceptualized as the two opposing forces within this dynamic: whereas commitment and in-depth exploration imply attempts to develop and maintain a sense of self (i.e., identity coherence or synthesis), reconsideration represents questioning and rethinking this sense of self (identity confusion). To measure this

three-dimensional model of identity formation, we developed the Utrecht-Management of Identity Commitments Scale (U-MICS: Crocetti, Berzonsky & Meeus, 2008) as an extension of the earlier Utrecht-Groningen Identity Development Scale (U-GIDS).

As was the case with Marcia's original dimensions of exploration and commitment, our three-dimensional model can be used to assign participants to identity status categories. For example, using cluster-analytic procedures in a cross-sectional study among 1952 Dutch early and middle adolescents, Crocetti et al. (2008) extracted five statuses from continuous measures of commitment, in-depth exploration, and reconsideration. Four of these statuses very closely resembled Marcia's four statuses. Achievement was represented as a combination of high commitment, high in-depth exploration, and very low reconsideration. Moratorium was represented by a combination of relatively low commitment, moderate in-depth exploration, and relatively high reconsideration; foreclosure as high commitment, relatively low in-depth exploration and very low reconsideration; and diffusion as very low commitment, very low in-depth exploration and very low reconsideration. In addition to these four statuses, a fifth status also emerged - a combination of high commitment, high in-depth exploration, and very high reconsideration. Crocetti et al. (2008) labeled this status as *searching moratorium*. Adolescents in this status have strong commitments and explore them intensively, but they are also very active in considering alternative commitments. Crocetti et al. (2008) found that searching moratoriums were empirically distinguishable from "classical" moratoriums in terms of psychosocial functioning - compared to those in classical moratorium, individuals classified into searching moratorium were characterized by lower levels of depression, anxiety, and aggression, as well as by more favorable relationships with parents. These findings underscore the differences between the two moratorium statuses and suggest that searching moratoriums seek alternative commitments while already possessing strong commitments, whereas classical moratoriums do so with weak or no current commitments.

Within this context, it is important to note that the foreclosed status, as defined by Crocetti et al. (2008), may carry a different psychological meaning depending on the developmental pathways through which adolescents arrive at this status. As a result, in the present study, we differentiate foreclosure into two subtypes - 'early closures' and 'closures'. When adolescents begin in the foreclosed status and remain there over time, they can be considered to be 'early closures', given that they have strong commitments that were established early on, have not tried to consider alternative commitments, and have not engaged in in-depth exploration of their present commitments. Adolescents, however, also can move from moratorium to this status of high commitment, low in-depth exploration, and low reconsideration. In this case they have considered alternative commitments, are not engaged in in-depth exploration of present commitments, and should be labeled simply as 'closures'. Among adolescents with high commitments and low levels of in-depth exploration and reconsideration, we expected that the longitudinal clustering procedures used in the present study would be able to distinguish between closures and early closures.

The differentiation between closures and early closures is intended to highlight the similar profile, but different developmental roots, of these two subtypes of foreclosure. As a result, when we refer to both types of closure, we use the label early closure/closure, ECC. Separate labels - EC for early closure and C for closure - are used to refer to the distinct statuses.

Taken together the findings reported by Crocetti et al. (2008) suggest that our three-dimensional model yields identity statuses that are conceptually quite similar to those proposed by Marcia. Therefore, as is the case with Marcia's statuses, the statuses found with the three-dimensional model can be ordered on an identity status continuum. Diffusion and achievement represent the least and most mature endpoints of the continuum, respectively, with moratorium, searching moratorium and early closure/closure representing intermediate statuses: D - M - SM - ECC - A. Therefore, the

identity status continuum generated by the three-dimensional model offers the potential to study change and stability of identity status.

8.1.3 The Present Study: Aims and Hypotheses

The primary goal of the present study was to evaluate the extent to which identity statuses represent stable individual dispositions versus states into and out of which individuals move over time during adolescence. Both assumptions of Waterman's 'developmental hypothesis' were evaluated here. Support for the first assumption would take the form of decreases in diffusion, and increases in achievement, over time. Support for the second assumption would take the form of the progressive transitions that Waterman proposed: D $\rightarrow$ M, D $\rightarrow$ F, F $\rightarrow$ M, and M $\rightarrow$ A. In the terminology used within our identity model, these transitions would be labeled as D $\rightarrow$ M, D $\rightarrow$ ECC, ECC $\rightarrow$ M, and M $\rightarrow$ A. Based upon prior literature, we expected to find support for the first assumption. Literature examining the second assumption is fairly scarce, so we treated this as an exploratory research question. These issues were examined using data from a five-wave study, including an early-to-middle adolescent cohort and a middle-to-late adolescent cohort, thereby covering the ages from 12 to 20.

We also examined gender differences in identity statuses and identity transitions. In a review of identity status studies between 1966-1995 Kroger (1997) discussed gender differences in overall, interpersonal and ideological identity. As is common in the identity status literature, she defined overall identity as ego strength and ego synthesis that individuals derive from commitments in a combination of life domains, and interpersonal and ideological identity as ego strength and synthesis that individuals derive from interpersonal and educational or work or political commitments, respectively. Kroger reported no gender differences in overall identity, but she did find that females were more often in achievement in interpersonal identity than males, and that in high school samples males seemed to move into the direction of achievement later than females (Kroger, 1997, p. 752 and p. 754,

respectively). Studies conducted since 1995 have replicated these findings with regard to interpersonal identity domains such as friendships (Lewis, 2003) and also have found a higher prevalence of females in achievement, along with a higher number of males in diffusion both in overall identity (Guerra & Braungart-Rieker, 1999) and in ideological identity domains (Schwartz & Montgomery, 2002). This pattern of findings suggests that, since the 1960s, gender differences appear more often in interpersonal identity than in overall and ideological identity, and more often in high school samples than in college/university samples. Moreover, the more recent findings suggest that gender differences also may be more likely to appear in overall and ideological identity from the late 1990s on.

In the present study, we used a Dutch sample where a majority of participants were in high school, and we used and a measure of overall identity that combines interpersonal and ideological domains. In the Netherlands, females may have stronger educational commitments, because they have tended to perform better in school than males since the late 1990s (Statistics Netherlands, 2008b, 2008c). Additionally, Dutch females have been found to have stronger interpersonal commitments than their male counterparts (Meeus & Deković, 1995). Given the age and the nationality of our participants and our use of a combination of interpersonal and ideological domains to tap overall identity, we therefore would expect females, compared to males, to be more strongly represented in achievement and less so in diffusion.

8.2 Method

8.2.1 Participants

Data for this study were collected as part of an ongoing research project on COnflict And Management Of RElationships (CONAMORE; Klimstra, Hale, Raaijmakers, Branje & Meeus, 2009), with a one-year interval between each of the five available waves. The longitudinal sample consisted of

1,313 participants divided into an early-to-middle adolescent cohort (n = 923; 70.3%), who were 12.4 years of age (SD = .59) on average at baseline, and a middle-to-late adolescent cohort (n = 390; 29.7%) with an average age of 16.7 years (SD = .80) during the first wave of measurement. Because both age groups were assessed during five measurement waves, a total age range from 12 to 20 years was available. The early to middle adolescent cohort consisted of 468 boys (50.7%) and 455 girls (49.3%), and the middle to late adolescent cohort consisted of 169 boys (43.3%) and 221 girls (56.7%). In both the younger and older cohorts, the vast majority of adolescents (85.1% and 84.3%, respectively) indicated that they were living with both their parents. The remainder of adolescents lived with their mother (7.9% and 7.2% in the younger and older cohort, respectively) or elsewhere (e.g., with their father, with one biological parent and one stepparent, or with other family members). The composition of the two cohorts did not significantly differ with regard to ethnicity. In the younger cohort, 83.4% identified themselves as Dutch, and 16.6% indicated that they belonged to ethnic minorities (e.g., Surinamese, Antillean, Moroccan, Turkish). In the older cohort, 87.4% of participants were Dutch, and 12.6% were ethnic minorities. In the year when the current study was initiated (2001), 21% of all Dutch early to middle adolescents, and 22% of the Dutch middle to late adolescents, belonged to ethnic minority groups (Statistics Netherlands, 2008a). Thus, ethnic minorities were slightly underrepresented in our sample. With regard to education, all participants initially were in junior high and high schools. Given the Dutch educational system, most participants switched schools at least once during the study. Specifically, participants in the younger cohort switched from junior high school to high school, whereas most of the participants in the older cohort switched from high school to college/university. Because of the sample recruitment procedure, 100% of our middle to late adolescents were in high school or college, whereas national demographic statistics (Statistics Netherlands, 2008a), StatisticsNetherlands2008b, StatisticsNetherlands2008c) reveal that

96% of the Dutch middle to late adolescents was in some form of education during the time period covered by the current study (i.e., 2001-2005).

Sample attrition was 1.2% across waves: in waves 1, 2, 3, 4, and 5 the number of participants was 1,313, 1,313, 1,293, 1,292 and 1,275, respectively. Missing values were estimated in SPSS, using the expectation maximization (EM) procedure. Little's Missing Completely At Random (MCAR) Test produced a normed χ^2 (χ^2/df) of 1.55, which, according to (Bollen, 1989), indicates that the data were likely missing at random, and that it is safe to impute missing values.

8.2.2 Procedure

Participating adolescents were recruited from various high schools in the province of Utrecht, The Netherlands. Participants and their parents received an invitation letter describing the research project and goals and inviting them to participate. More than 99% of the families who were approached signed the informed consent form. During regular annual assessments, participating adolescents completed questionnaires at school or at home. Confidentiality of responses was guaranteed. Adolescents received €10 (approximately US $13) for each wave in which they provided data.

8.2.3 Measures

Identity. Identity was assessed using the U-MICS (Crocetti et al., 2008). The U-MICS consists of 13 five-point Likert-scale items (1 = *completely untrue* to 5 = *completely true*), measuring identity in three dimensions: commitment, in-depth exploration, and reconsideration of commitment. Within the ideological and interpersonal domains, the U-MICS consists of 5 items measuring commitment, 5 items measuring in-depth exploration of present commitments, and 3 items measuring reconsideration of commitment. Sample items for the ideological and interpersonal domain include respectively "My education/best friend makes me feel confident about myself" (commitment), "I often think about my education/best friend" (in-depth

exploration), and "In fact, I'm looking for a different education/a new best friend" (reconsideration of commitment). Although the U-MICS measures identity in different domains, the instrument can be used to assess overall identity. Crocetti et al. (2008) included both ideological (education) and interpersonal (best friend) domains and demonstrated the internal factorial validity of the three-dimensional model across domains. In the present study, Cronbach's alphas for commitment, in-depth exploration and reconsideration of commitment ranged in both cohorts from .91 to .93, .88 to .89, and .84 to .94 across waves, respectively.

8.2.4 Analytic Strategy

To address our research questions, we utilized two applications of the general latent class model: latent class analysis (LCA) and latent transition analysis (LTA). Because we wanted to compare the prevalence of the various identity statuses between the early-to-middle and middle-to-late cohorts, we assumed measurement invariance across cohorts, that is we restricted the profiles of the five identity classes on the three identity dimensions to be the same across cohorts. LCA is a person-centered analytic strategy that is a confirmatory version of cluster analysis. LCA groups individuals into classes based on empirically distinct patterns of scores on the variables (in this case the three identity dimensions) used to create the classes. LCA of continuous variables is sometimes referred to as latent profile analysis. For simplicity, we use the term LCA here. Like confirmatory factor analysis, LCA generates both measurement and structural parameters (Nylund, Asparouhov & Muthén, 2007). The continuous scores for each of the identity variables within each class represent the measurement parameters, whereas the structural parameters refer to the class membership probabilities assigned to groups of individuals. Unlike cluster analysis, LCA offers fit statistics and significance tests to determine number of classes, it assigns class membership based on class probabilities, thereby taking uncertainty of membership, or error, into account. LCA has been found to be superior to cluster analysis in several

Monte Carlo studies (Reinke & Ialongo, 2008). In the present study, we applied LCA to test whether the five hypothesized identity statuses would emerge in each of the 5 waves of measurement.

LTA represents a longitudinal extension of LCA (for a recent overview of LTA, see Kaplan, 2008). LTA calculates patterns of stability and change over time in the form of movement or transitions between classes (identity statuses in this case). Like LCA, LTA models use class-specific parameters (the continuous scores for each of the identity variables within each class) as measurement parameters, and class probabilities as structural parameters to estimate the number of participants in each of the classes. To model change over time, LTA adds a second set of structural parameters, latent transition probabilities, to the latent class model. In a two-wave LTA, for example, transition probabilities refer to the probability of moving into class Y in wave 2 conditional on having been in class X in wave 1. These transition probabilities range between 0 and 1. In sum, then, LTA offers two types of structural parameters: (a) varying numbers of participants in class across waves, indicating increase or decrease in class size over time, and (b) transitions of individuals between classes that carry these changes of class size. Therefore LTA is appropriate for evaluating both assumptions of Waterman's developmental hypothesis, the hypothesized decrease of diffusion and increase of achievement, and the hypothesized identity transitions that carry this increase or decrease over time.

LTA results can be converted into contingency tables summarizing the prevalence of classes (identity statuses) across waves. We use Bayesian Model Selection using (in)equality constraints between the parameters of interest to evaluate these contingency tables. For a more detailed description of this method, readers are referred to Laudy and Hoijtink (2007). Using constraints may express prior information explicitly. This way we can evaluate the likelihood of certain patterns of increases and decreases in identity status membership. Moreover, expected differences in prevalence

of identity statuses between early-to-middle and middle-to-late adolescents and males and females can be evaluated.

The results of the Bayesian Model Selection are expressed in terms of Bayes factors (BFs), representing the amount of evidence in favor of the model at hand compared to another model, and posterior model probabilities (PMPs), representing the probability that the model at hand is the best among a set of finite models after observing the data. Posterior model probabilities of a model are computed by dividing its BF by the sum of all BFs.

8.3 Results

We present our results in 4 steps. First we apply cross-sectional LCA to explore the number of classes (identity statuses) within each of the 5 waves. Second, we select the best-fitting five-wave LTA model in a number of successive steps. Because we wanted to compare the prevalence of the various identity statuses between the early-to-middle and middle-to-late cohorts, we assumed measurement invariance across cohorts by restricting the profiles of the five identity classes on the three identity dimensions to be the same across cohorts. Third, we apply Bayesian evaluations of the contingency tables generated by the final LTA model. The Bayesian evaluations are intended to address four research questions:

1. Is there a differential increase and decrease of identity statuses across time?

2. Are there differences in the prevalence of statuses across time between the early-to-middle and middle-to-late adolescents?

3. Likewise, are there differences in the prevalence between and males and females?

4. Is there a differential increase and decrease of identity statuses over time between males and females?

Fourth, we globally describe the sequence of identity statuses in five-wave identity status trajectories (for instance MMAAA).

8.3.1 Cross-Sectional Latent Class Analysis

For each of the five waves, we estimated a set of cross-sectional LCA's on the entire sample, including all three identity dimensions simultaneously. Analyses were performed using Mplus. We used four criteria to determine the number of latent classes (Nagin, 2005). First, a solution with k classes should result in improvement of model fit compared to a solution with k - 1 classes, indicated by a decrease of the Bayesian Information Criterion (BIC). Second, adding an additional class should lead to a *significant* increase of fit, as indicated by the bootstrap Lo-Mendel-Rubin likelihood ratio test (BLRT) (Nylund et al., 2007). Third, entropy - a standardized measure of classification of individuals into classes, based upon the posterior class probabilities - of the final class solution should be acceptable. Entropy values range from 0 to 1, with values of .70 or higher indicating good classification accuracy (Reinecke, 2006). Fourth, we evaluated the content of the classes in the various solutions. If an additional class in a solution with k classes was found to be a slight variation of a class already found in a solution with k - 1 classes, we would choose the most parsimonious solution.

As expected, we found the five-class solution to be superior to the one- to four-class solutions on both fit indices across waves. BIC of the five-class solutions was at least 56.21 lower than that of one-, two-, three-, or four-class solutions and only in wave 5 did the BLRT indicate that the five-class solution did not fit significantly better than the four-class solution ($p = .14$). Entropy (E) for the five-class solution ranged between .74 and .81, indicating good classification accuracy. Adding a sixth class did not provide additional unique information, given that the sixth class was small ($10 > n > 27$ in each wave) and appeared to represent a variation of one of the other classes. Therefore, we decided to use a five-class model in the LTA's.

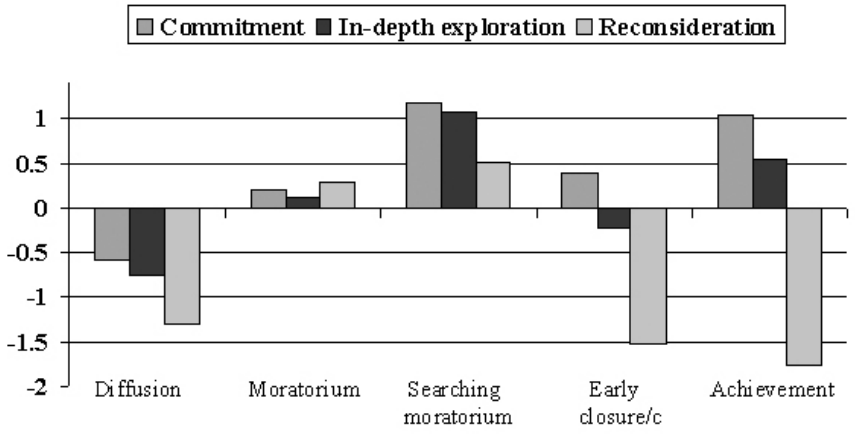


Figure 8.1: Profiles of the identity statuses on the three identity dimensions across waves. *Note*: For reasons of presentation the scores of the identity dimensions of the statuses were centered at a common scale point (2.75). Early closure/c = Early closure/closure.

8.3.2 Five-Wave Latent Transition Analysis

As part of the LTA, we assumed measurement invariance in the five-class LCA solutions across measurement waves. That is, we restricted the profiles of the five identity classes on the three identity dimensions to be equivalent across five waves. We also restricted the variances of the three identity dimensions to be equivalent across classes across waves. Assuming measurement invariance ensures that the profiles of the classes are the same across waves, and allows for a straightforward interpretation of transition probabilities (see Nylund, Muthén. B., Bellmore & Graham., 2006). Figure 8.1 displays the profiles of the statuses. Four of the classes (achievement, moratorium, early closure/closure, and diffusion) strongly resemble Marcia's original statuses. The fifth status, searching moratorium, combines very strong commitment with high levels of in-depth exploration and very high levels of reconsideration.

We developed the final LTA model in two steps. We will describe these steps and then present the results of the final model. In both steps, we

selected the model with the lowest BIC value. The BLRT is not available for LTA models.

LTA step 1: Non-stationary versus stationary transition probabilities. In the first modeling step, we compared a model with non-stationary transition probabilities between adjacent waves to a model with stationary transition probabilities. A model with non-stationary transition probabilities assumes that the likelihoods of transitions between classes are different between waves. In contrast, a model with stationary transition probabilities assumes that the probabilities are equal across waves. Results indicated no significant differences in the transition probabilities across time. The BIC for the LTA model with stationary transition probabilities (32,895) was lower than BIC of the model with non-stationary transition probabilities (33,112). This suggests that adolescents make transitions between identity statuses at the same pace across the four transitions points. As a result, there seems to be a very regular pattern of identity development. Entropy of the stationary model was very good: .85.

LTA step 2: Are there age and gender differences in identity status transitions? We added covariates to the model with stationary transition probabilities to describe heterogeneity in transitions between statuses. In the first model, we included cohort as covariate to test whether transitions into and out of identity statuses were different between the early-to-middle and middle-to-late adolescents. The second model tested whether transitions were different for males and females. The first model comparison indicated no significant differences in the transition probabilities between the cohorts. The BIC for the LTA without cohort as covariate (32,895) was lower than BIC of the model with covariate (32,901). The second model comparison indicated significant gender differences in transition probabilities. The BIC for the LTA with gender (32,894) was lower than BIC of the model without gender (32,895), indicating that the model with gender was 2.72 times more likely than the model without gender (Nagin, 1999). So, rate of change into and out of identity statuses was not different for early-to-middle and middle-to-

late adolescents, but was for males and females. Below, we present follow-up Bayesian analyses to clarify the gender differences.

Increase and decrease of identity statuses over time. Table 8.1 displays the cell sizes for each the five identity statuses for waves 1, 2, 3, 4, and 5 based on the final LTA model. Findings for the whole sample are in the upper panel of the Table. The Table indicates a systematic decrease in diffusion (D), moratorium (M) and searching moratorium (SM) over time, along with a systematic increase in early closure/closure (ECC) and achievement (A). The Table also suggests that early closure/closure is the most prevalent status: between 50.6% and 55.2% of the sample was classified into that status across waves. A majority of the adolescents had relatively strong commitments along with relatively low levels of in-depth exploration and very low levels of reconsideration. The systematic pattern of increases and decreases in status membership is also found across both cohort and gender (panels 2, 3, 4 and 5 of Table 8.1, respectively).

We applied Bayesian Model Selection (Laudy & Hoijtink, 2007) to the upper panel of Table 8.1 to test which of three alternative models of increase and decrease of identity status best fit the data. Model 1 assumed no increase or decrease of identity statuses across five waves, whereas Model 2 assumed a decrease of D, M, and SM and an increase of ECC and A. In Model 3, the unconstrained model, the distribution of statuses over time was allowed to vary freely. The results are in Table 8.2. First, Models 1 and 2 were compared with the unconstrained (Model 3). The BFs for Models 1 and 2 imply that after observing the data, these models are approximately 270 and 7,500 times as likely, respectively, as the unconstrained (Model 3). The second comparison revealed that Model 2 is 27.64 times as likely as Model 1. Posterior model probabilities of models 1, 2, and 3 are .03, .97 and $< .001$, respectively. Note that we assume that before observing the data each model is equally likely. In sum, Model 2, assuming decreases in D, M and SM and increase in ECC and A, was by far the best-fitting model. This model appears to support the first assumption of Waterman's 'developmental hypothesis'.

Table 8.1: Size of Identity Status Classes for the Whole Sample, Early-to-Middle and Middle-to-Late Adolescents, and Males and Females. Findings Based on the Final Stationary 1-year Interval Model

	Identity status									
	Diffusion		Moratorium		Searching moratorium		Early closure/c		Achievement	
Wave	n	%	n	%	n	%	n	%	n	%
Total sample ($N = 1313$)										
1	117	8.9	215	16.4	78	5.9	665	50.6	238	18.1
2	106	8.1	218	16.6	67	5.1	676	51.5	246	18.7
3	97	7.4	211	16.1	60	4.6	684	52.1	261	19.9
4	90	6.9	186	14.2	41	3.1	713	54.3	283	21.6
5	82	6.2	183	13.9	23	1.8	725	55.2	300	22.
Early-to-middle adolescence ($N = 923$)										
1	86	9.3	159	17.2	70	7.6	455	49.3	153	16.6
2	76	8.2	162	17.6	64	6.9	463	50.2	158	17.1
3	68	7.4	146	15.8	57	6.2	481	52.1	171	18.5
4	66	7.2	127	13.8	36	3.9	505	54.7	189	20.5
5	61	6.6	131	14.2	23	2.5	510	55.3	198	21.5
Middle-to-late adolescence ($N = 390$)										
1	31	7.9	56	14.4	8	2.1	210	53.8	85	21.8
2	30	7.7	56	14.4	3	0.8	213	54.6	88	22.6
3	29	7.4	65	16.7	3	0.8	203	52.1	90	23.1
4	24	6.2	59	15.1	5	1.3	208	53.3	94	24.1
5	21	5.5	52	13.3	0	0.0	215	55.1	102	26.2
Males ($N = 637$)										
1	69	10.8	136	21.4	55	8.6	292	45.8	85	13.3
2	61	9.6	141	22.1	45	7.1	300	47.1	90	14.1
3	59	9.3	142	22.3	41	6.4	298	46.8	97	15.2
4	57	8.9	124	19.5	29	4.6	311	48.8	116	18.2
5	51	8.0	115	18.1	17	2.7	328	51.5	126	19.8
Females ($N = 676$)										
1	48	7.1	79	11.7	23	3.4	373	55.2	153	22.6
2	45	6.7	77	11.4	22	3.3	376	55.6	156	23.1
3	38	5.6	69	10.2	19	2.8	386	57.1	164	24.3
4	33	4.9	62	9.2	12	1.8	402	59.5	167	24.7
5	31	4.6	68	10.1	6	0.9	397	58.7	174	25.7

Note. Early closure/c = Early closure/closure.

Table 8.2: Bayesian Model Selection: Comparison of Various Sets of Models

Models	Model comparisons		
	BF of M1/M2 versus M3	BF of M1 versus M2	PMP
RQ1: Increase and decrease of identity statuses?			
M1. No increase or decrease of D, M, SM, ECC, and A	271.16	1	.03
M2. Decrease of D, M and SM and increase of ECC and A	7,497.52	27.64	.097
M3. Unconstrained	1[1]		< .001
RQ2: Prevalence of identity statuses different between cohorts?			
M1. No difference in prevalence	.004	1	< .001
M2. Systematic difference in prevalence	211.50	52,875	.99
M3. Unconstrained	1		< .01
RQ3: Prevalence of identity statuses different between males and females?			
M1. No difference in prevalence	< .01	1	< .001
M2. Systematic difference in prevalence	510.73	51,073	.99
M3. Unconstrained	1		< .001
RQ4: More decrease of D, M and SM and increase of ECC and A in males then females?			
M1. No gender-specific increase or decrease	.004	1	< .001
M2. Gender-specific increase or decrease	5.29	1,322.5	.84
M3. Unconstrained	1		.16

Note. ECC = Early closure/closure. BF = Bayes factor. PMP is posterior model probability.

[1] Models with BF = 1 are reference category.

Transitions between identity statuses. Transition probabilities of identity status change across one-year intervals, as found in the final stationary model, are in Table 8.3. The transition probabilities of identity change between waves 1 and 5 are displayed on the right-hand side of the Table. The 4-year probabilities were calculated using the contingency tables of waves 1 and 5 as generated by the final LTA model. We added these longer-term transition probabilities as a way of elucidating change in identity status across a longer period of time. As expected given the consistency of identity status transition probabilities across time, the transitions during 1-year intervals strongly parallel the transitions during 4-year intervals. Transitions with a relatively high frequency during 1-year intervals were also highly likely during the 4-year interval. Not surprisingly, stability of identity statuses was greater during 1-year intervals than during the 4-year interval, and transitions between identity statuses were more likely to have occurred across four years than across one year.

Seven specific findings warrant mention here. First, 1-year stability is always more likely than change in identity status. This is also true for four-year stability of M, ECC, and A, and with two exceptions for the four-year stability of D and SM, as compared to 4-year identity status change. Notably, 1-year and 4-year stability probabilities for ECC and A are very substantial. Second, very few adolescents shifted from M, ECC, and A into D or SM; during 1-year intervals 4 percent or fewer of the adolescents made this transition, and during the 4-year interval, 5 percent or fewer did so. Third, transitions into moratorium are limited: during 1-year intervals, 9% of diffusions and 23% of searching moratoriums moved into moratorium, whereas during the 4-year interval 11% of diffusions and achievers, and 26% of searching moratoriums, moved into moratorium. Fourth, the percentage of transitions into early closure/closure is substantial: between 8 and 19% of the adolescents in D, M, SM and A moved into early closure/closure during 1 year, and between 22 and 45% did so over the four years of the study. The likelihood of moving from D to ECC was very substantial, 19 and 45% during

Table 8.3: Transition Probabilities of Identity Status Change During 1-year Intervals (n + 1) and 4-year Intervals (n + 4) Across Five Waves. Findings of the Final Stationary Model

Identity status in year n	Identity status in year n + 1[a]					Identity status in year n + 4				
	D	M	SM	ECC	A	D	M	SM	ECC	A
Diffusion (D)	.70	.09	.00	.19	.02	.39	.11	.00	.45	.05
Moratorium (M)	.03	.71	.04	.13	.09	.05	.39	.04	.30	.22
Searching moratorium (SM)	.00	.22	.50	.08	.19	.03	.26	.18	.22	.32
Early closure/c (ECC)	.02	.03	.00	.90	.05	.03	.06	.00	.80	.11
Achievement (A)	.01	.05	.02	.11	.81	.01	.11	.00	.26	.62

Note. Early closure/c = Early closure/closure.

[a] For a stationary model, all transitions probabilities are the same across waves.

1 and 4 years, respectively. Given the distinction we have made between early closure and closure in the introduction, the transition out of diffusion should be labeled as D → EC, given that adolescents who make this transition have never reconsidered identity alternatives or explored present commitments in depth. Similarly, remaining in the early closure status should be labeled as EC → EC. However, adolescents who move from SM, M, or A into ECC have considered identity alternatives or have explored present commitments in depth. Therefore these transitions should be labeled as SM → C, M → C, and A → C. These refer to adolescents who once maintained high levels of in-depth exploration (in the case of regression from achievement to closure) or reconsideration (in the case of movement out of either of the moratorium statuses) and now report low levels of both in-depth exploration and reconsideration. Fifth, few adolescents move from D to A: 2 and 5% during 1 and 4 years, respectively. Sixth, transitions from EC, M, and especially SM into A are quite prevalent: from 5 to 19% and 11 to 32% over 1 and 4 years, respectively. This suggests that the likelihood of a diffused adolescent reaching achievement during adolescence is very low. Seventh, most of the transitions are progressive: from D, M and SM into the direction of ECC and A. But there are also "regressive" transitions: notably from A into M and C, and from SM into M: 5 and 11%, 11 and 26%, and 22 and 26% during 1 and 4 years, respectively.

In addition, to check whether the stability of the findings is affected by the difference in sample size between cohorts, we estimated a replication of our final stationary model, controlling for sample size differences between cohorts by weighting the cohorts equally in the model. This was done by assigning the weight of 1 to each of the 923 cases of the younger cohort and the weight of 2.366 to each of the 390 cases in the older cohort. The replication confirmed all the earlier reported findings and suggests that they were not affected by difference in sample size between cohorts.

Identity status trajectories. Inspection of the five wave identity status trajectories revealed two general patterns. First, 822 adolescents were in

the same identity status in waves 1 and 5. The vast majority of the 822 participants (93.7%, or 63% of the total sample) stayed in the same status in all waves. Second, 491 adolescents were in different statuses in waves 1 and 5: 78.2% of them made only one status transition, 20.4% made two transitions, and 1.4% made three or more transitions during the five waves of the study. So the majority of identity status changers made only one transition. We also found that 11% of the change trajectories were two-transition trajectories in which adolescents passed through SM or M as transitory identity statuses.

These findings partially support the second assumption of Waterman's developmental hypothesis. We indeed found three of the four progressive identity status trajectories that Waterman hypothesized, notably D $\rightarrow$ M, D $\rightarrow$ EC, and M $\rightarrow$ A. We did not find general support for Waterman's assumption that adolescents starting from diffusion move through more than two identity statuses to reach achievement. We elaborate further on these issues in the discussion.

Age differences. In the second step of LTA modeling, we did not find differences between age groups in rate of change into and out of identity statuses. Table 8.1, second and third upper panels, summarizes this similar and regular identity change for both cohorts. This table also shows systematic cohort differences in the prevalence of the statuses in waves 1 to 5. In all waves, the number of diffusions, moratoriums, and searching moratoriums is higher in the younger age group, whereas the number of early closures/closures and achievers is lower. There are a few exceptions to this general pattern: in waves 3 and 4, the prevalence of moratoriums was lower, and in waves 4 and 5, the prevalence of early closures/closures was higher in the younger age group. We applied Bayesian Model Selection to evaluate which of three alternative models of the prevalence of the identity statuses in waves 1 and 5 in both cohorts provided the best fit to the data. Model 1 assumed no difference in prevalence between the cohorts, whereas Model 2 assumed a higher prevalence of D, M and SM in early-to-middle adolescence and a higher prevalence of ECC and A in middle-to-late adolescence. Model

3, the unconstrained model, did not specify the distribution of statuses across cohorts. Table 8.2 presents the findings. The BFs imply that Model 1 is 250 times less likely than Model 3, and that Model 2 is 211.50 times more likely than Model 3. Moreover, Model 2 is 52,875 times as likely as Model 1. Posterior model probabilities of models 1, 2, and 3 are $< .001$, .99 and $< .01$, respectively. Replication of the same Bayesian models for waves 2, 3 and 4 revealed similar differences between age groups. We do not include a full report of these models to save space. These findings show that the middle-to-late adolescents are generally in more "progressive" identity statuses than the early-to-middle adolescents.

Finally, Table 8.1 makes clear that the patterns of increases and decreases in identity status memberships unfold quite systematically from early to middle and middle to late adolescence. Diffusion decreases from 9.3 to 6.6% in early-to-middle adolescence, and from 7.9 to 5.5% in middle-to-late adolescence. For moratorium, the percentages decrease from 17.2 to 14.2% and from 14.4 to 13.3%, and for searching moratorium the percentages decrease from 7.6 to 2.5 and from 2.1 to 0%, whereas the percentages for achievement rise from 16.6 to 21.5 and from 21.8 to 26.2%. Percentages for the early closure/closure status rise from 49.3 to 55.3 in the younger cohort but are fairly stable in the older cohort (53.8 to 55.1). So, for diffusion, moratorium, and searching moratorium, a systematic decrease is evident from early to late adolescence; for achievement we observe a systematic increase; and for the early closure/closure identity status, there appears to be an increase followed by stabilization at a relatively high level.

Gender Differences. In the second step of LTA modeling, we found gender differences in identity status transitions. Inspection of the separate 1-year transition tables indicated the general pattern of identity transitions to be the same for males and females. The seven primary results that we found for the transitions in the whole sample appeared to generalize across gender. However, one systematic gender difference appeared: in 4 of the 5 statuses (M, SM, ECC, and A), 1-year stability was more likely in females than in

males - percentage differences were 1%, 12%, 2%, and 4%, respectively. To interpret this finding, we estimated two sets of Bayesian models on the two lower panels of Table 8.1. In the first set, we evaluated gender differences in the prevalence of identity statuses in waves 1 and 5. Model 1 assumed no difference in prevalence between males and females, whereas Model 2 assumed a higher prevalence of D, M and SM in males and a higher prevalence of ECC and A in females. Model 3, the unconstrained model, did not specify any distribution of the statuses for either gender. Table 8.2 presents the findings. The BFs indicated that Model 1 is approximately 100 times less likely than Model 3, and that Model 2 is about 510 times more likely than Model 3. Moreover, Model 2 was 51,073 times more likely than Model 1. Posterior model probabilities of models 1, 2, and 3 were $< .001$, .99 and $< .001$ respectively - and as a result, we retained Model 2. Replication of the same Bayesian models for waves 2, 3 and 4 revealed the same patterns of differences between males and females. We include no full report of the models to save space. These findings show that females are generally classified into more "progressive" identity statuses than males.

In the second set of Bayesian models, we evaluated three alternative models of gender differences in increases and decreases of identity status membership between waves 1 and 5. Model 1 assumed no gender differences in likelihood of increase or decrease in the identity statuses over time, whereas Model 2 assumed a larger likelihood of decrease of D, M, and SM in males than in females and a larger likelihood of increase of ECC and A in males than in females. Model 3, the unconstrained model, did not specify any pattern of increase or decrease of identity statuses for either gender. Table 8.2 presents the findings from these model comparisons. The BFs imply that Model 1 was 250 times less likely to represent the data adequately than Model 3, and that Model 2 was about 5 times more likely than Model 3. Moreover, Model 2 was 1,322.5 times more likely to represent the data than Model 1. Posterior model probabilities for models 1, 2, and 3 were $< .001$, .84, and .16, respectively. These findings show that decreases in D, M and

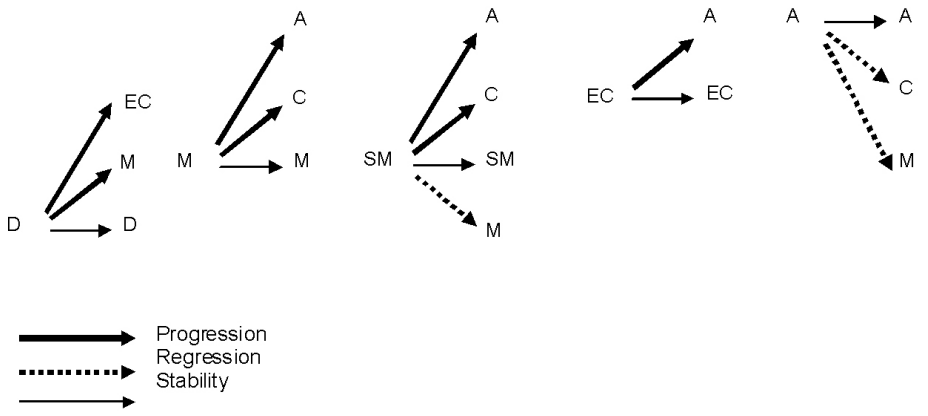


Figure 8.2: Transitions of identity progression, regression and stability. Note: D → EC, D → M and M → A were hypothesized by Waterman (1982) as progressive transitions, and A → M as regressive transition. Percentages of the number of participants that changed statuses between wave 1 and wave 5 can be found in the right-hand side of Table 8.3 (for example, .11 signifies that 11 percent of the participants who were in diffusion in wave 1 moved to moratorium in wave 5).

SM, and increases in ECC and A, are more likely in males than females. Taken together, these findings indicate that females are in more advanced and more stable identity statuses than males, but that males may catch up during later adolescence.

Conclusion. Figure 8.2 summarizes the main findings of the final LTA model. The figure incorporates only those identity transitions with 4-year probabilities of .10 or above. The figure is based on an additional Bayesian Model Selection, in which we constrained the probabilities of the 4-year transitions below .10 (see Table 8.3) to be zero and the probabilities of the other transition to be greater than zero. This model was clearly superior to a model in which the transitions were allowed to vary freely; posterior model probabilities of both models were .99 and .01, respectively. A comparable model comparison of the probabilities of the 1-year transitions replicated these findings. The figure clearly shows that progressive identity transitions outnumber regressive transitions. Identity progression is represented by seven

transitions: D → M, D → EC, M → C, M → A, SM → C, SM → A and EC → A. Identity regression is represented by three transitions: A → M, A → C, and SM → M.

8.4 Discussion

8.4.1 Identity Formation: Progression and Stability

The present study was designed to test the extent to which identity status is a stable individual disposition or whether it changes over time. Our findings revealed a steady increase of A and ECC, and a steady decrease of D, M and SM, in both cohorts. These findings support the first assumption of Waterman's (1982) developmental hypothesis and document identity progression: growth of the high-commitment statuses (ECC and A), decrease in the number of adolescents who do not address identity issues (D), and decrease in the proportion of adolescents who are in the process of finding their identity (M and SM). Taken together, these findings clearly demonstrate identity maturation during adolescence, and they converge with the results of Klimstra et al. (2009), who have provided evidence for systematic personality maturation between early and late adolescence. Support for Waterman's second hypothesis was also substantial, but less consistent: we found support for three of the four progressive identity transitions that Waterman hypothesized, notably D to EC, D to M and M to A, and for one of the four regressive transitions he proposed: notably A to M. On the other hand, we also found substantial support for the individual difference perspective: 63% of the adolescents remained in the same identity status in wave 1 and 5. This percentage is remarkably similar to the 59% found in the longitudinal identity status studies reviewed in the introduction. So our study replicates earlier findings on change and stability of identity. We offer two interpretations for the high stability of identity status that we found. The first interpretation assumes that changes in identity may be more prevalent in emerging adulthood than in adolescence due to more frequent

and intense consideration of adult roles. Some support for this position has been obtained in a recent meta-analysis of personality change by Roberts, Walton and Viechtbauer (2006), who concluded that personality change is much more prevalent in young adulthood than in adolescence. The second assumption states that the universality of the identity conflict and subsequent identity formation has been wrongly and too commonly assumed. Support for this interpretation has been found in the review by Kroger (2007), who showed across cross-sectional and two-wave longitudinal studies no more than 25% of participants to be in moratorium. Future longitudinal studies from early adolescence into emerging adulthood are needed to more thoroughly test these alternative interpretations.

Another major finding is that Marcia's (1966) original four statuses - achievement, moratorium, foreclosure (here ECC), and diffusion - indeed emerged empirically as identity statuses at all 5 waves along with a new status: searching moratorium. Our findings strongly suggest that searching moratorium is an early and middle adolescent status that disappears in late adolescence. So, in late adolescence (corresponding to the age group that Marcia used in his own research), we empirically extracted Marcia's four identity statuses. It is also of interest to note that the five-wave LTA model was characterized by high classification accuracy ($E = .85$).

Interestingly, we found the early closure/closure status to be the most prevalent in both cohorts. A majority (80%) of the 725 adolescents who were in this status in wave 5 can be considered to be early closures: they stayed in this status in all waves ($n = 530$), or made the transition from diffusion to early closure ($n = 53$). They are labeled as early closures because they have not considered identity alternatives and have continued to maintain relatively strong commitments over time. A minority (20%) of the adolescents with an early closure/closure profile transition from moratorium, searching moratorium, or achievement. They have considered identity alternatives and therefore cannot be defined as early closures. We have labeled them as closures. At present they possess relatively strong

commitments and maintain them in an automatic way. They did not, however, always do so. We will discuss this issue further in the paragraph on limitations and suggestions for further research.

8.4.2 Identity Transitions

Figure 8.2 summarizes the main findings of the final LTA model. The transitions depicted in this table are predominantly progressive. Identity progression is represented by seven transitions: D → M, D → EC, M → C, M → A, SM → C, SM → A, and EC → A. Identity regression is represented by three transitions: A → M, A → C, and SM → M. The seven progressive transitions indicate different changes in the identity configurations of adolescents: starting to think about alternative commitments (D → M), making stronger commitments (D → EC), making stronger or much stronger commitments along with decreases in considering alternative commitments (M → C and M → A, respectively), making much more secure (SM → C) or much more secure and active (SM → A) commitments, and making more active and strong commitments (EC → A). The three regressive identity transitions indicate discarding commitments and starting to consider alternative ones (A → M), moving from strong and active commitments to more rigid and less active commitments (A → C), and discarding commitments while maintaining high levels of reconsideration (SM → M).

The analyses of status transitions imply that the high-commitment statuses, A and ECC, are more likely than the other statuses to serve as endpoints of identity formation in adolescence. Stability of A and ECC is very substantial, and recent studies have documented that both statuses show more positive profiles of psychosocial adjustment compared to diffusions or moratoriums. Achievers and adolescents in foreclosure or early closure/closure tend to be characterized by relatively low levels of depression (Meeus, 1996), anxiety (Berman, Weems & Stickle, 2006), substance use (Luyckx, Goossens, Soenens & Vansteenkiste, 2005), aggression (Crocetti et al., 2008), and relatively high levels of emotional stability and self-

esteem (Luyckx et al., 2005). In contrast to achievement and early closure/closure, searching moratorium and diffusion appear to be almost exclusively transitional statuses, given that many adolescents move out of these statuses and few (if any) move into them.

The transitional analyses suggested only a limited amount of change. Few participants shifted from diffusion into achievement or from achievement to diffusion (Table 8.3). Secondly, analyses of the identity trajectories revealed that the majority of adolescents who change identity status across five years make only one transition. This makes clear that changes in identity status tend to be decisive, and that there is an extremely low probability of additional identity status transitions.

The analyses of identity transitions and trajectories shed new light on moratorium and reinforce the distinction between moratorium and searching moratorium. First, our findings particularly underscore the transitional nature of searching moratorium, as adolescents move primarily out of this status. In fact, there were no searching moratoriums in the last waves for the middle-to-late adolescent cohort. On the other hand, our analyses showed moderate stability among adolescents in moratorium. This finding suggests that a considerable number of moratoriums might be unable to move out of this unstable identity configuration, and should be considered as "characterological moratoriums," as Côté and Schwartz (2002, p. 584) have suggested. An important implication of Ct and Schwartz's suggestion is that, for characterological moratoriums, back-and-forth movement between moratorium and achievement, as has been found to occur in adulthood for some people (Stephen, Fraser & Marcia, 1992), is improbable as an identity trajectory in adolescence. However, moratorium and searching moratorium may nonetheless function as transitory statuses. When adolescents make more than one identity transition, they tend to pass through moratorium or searching moratorium. "Classical" moratorium, however, may also be characterological, whereas searching moratorium appears to be exclusively transitory. In addition, our findings show that searching moratorium offers

a better starting point to reach achievement than does moratorium: 32% versus 22% (4-year transition probabilities), respectively.

As noted earlier, our findings partially support the second assumption of Waterman's developmental hypothesis. Waterman (1982) hypothesized that progressive identity development would occur through the transitions D → EC, D → M, EC → M, and M → A. We found evidence for the progressive transitions D → EC, D → M and M → A. The explanation for the non-occurrence of the transition EC → M is that, as noted before, the early closure status functions very often as the final identity status observed in adolescence. Adolescents in this status do not think a lot about their present commitments and are not active in searching for alternative commitments. Therefore, they do not appear likely to give up their commitments and consider adopting alternative life choices. With regard to regressive transitions, we found support for one of the four regressive identity status transitions that Waterman hypothesized: from achievement to moratorium. Achievers are adolescents with the highest level of in-depth exploration. They are very active in processing information about their commitments. This orientation may give rise to loss of present commitments and search for alternative commitments if one's present commitments are deemed unsatisfactory (Luyckx et al., 2006). We did not find support for the three other regressive transitions that Waterman hypothesized: ECC → D, M → D, and A → D. Our findings clearly suggest that it is almost impossible to go back to a state of disinterest in identity work (D) once individuals have held strong commitments (ECC and A) or have been active in considering alternative commitments (M). In sum, consistent with earlier research (Berzonsky & Adams, 1999; Kroger, 2007; Van Hoof, 1999), we found more progression than regression. At the same time, our findings also make clear that regression in identity is something that cannot be ignored.

8.4.3 Gender Differences

We found considerable gender differences in patterns of identity formation. As expected, females were more likely to be achieved, and less likely to be diffused, than males. In addition, we also found that females were more likely to be classified into the early closure/closure status, and less likely to appear in both moratoria, than males. These findings suggest that, at least at present, females may be "further ahead" of males when overall identity is measured as a combination of interpersonal and ideological (especially educational) domains in adolescence and the early part of emerging adulthood. The explanation might be that Dutch females combine their classic stronger interpersonal commitments (Meeus & Deković, 1995) with stronger educational commitments, given that girls at present often outperform boys in school (Statistics Netherlands, 2008b, 2008c). Our findings are consistent with results of recent studies in the United States that have reported gender differences in interpersonal identity (Lewis, 2003), ideological identity (Schwartz & Montgomery, 2002) and overall identity (Guerra & Braungart-Rieker, 1999). We also found, however, that males tend to "catch up" during adolescence. This is consistent with the review by Kroger (1997), and suggests that earlier physical and cognitive maturation in girls may account for some of this pattern. Recent studies have shown that girls reach puberty between one and two years earlier than boys (Beunen et al., 2000), and that, in early adolescence, girls tend to be up to a full year ahead of boys in several aspects of brain development (Giedd, Blumenthal, Jeffries, Castellanos & Zijdenbos, 1999; Colom & Lynn, 2004). Therefore, girls might reach the mature identity statuses earlier than boys, whereas boys catch up during adolescence.

8.4.4 Limitations and Suggestions for Further Research

A number of limitations of the present study warrant discussion. The first limitation involves our sole reliance on self-report questionnaires. Although questionnaires are the most appropriate instruments by which to gather

information on subjective processes, such as identity development, the biases involved in self-reports may have come into play. Future research could try to overcome these biases by focusing on the micro processes underlying identity development. This could be done by tapping identity on a day-to-day basis.

A second limitation of our study is its descriptive nature. We did not test explanations of identity progression during adolescence - for example, why certain identity transitions have a higher probability than others. Given that identity status transitions in adolescence seem to be quite decisive, future research should try to specify the conditions that predict the timing of these transitions. Longitudinal designs that include a focus on the link between identity transitions and life transitions, that is transitions in the educational and occupational career and in the formation of intimate relationships, may be a fruitful option here.

Although our model is conceptually distinct from Marcia's model in two respects, the findings of our study are remarkably similar to those of studies using Marcia's paradigm. Notably we found similar percentages of change and stability of identity statuses, moratorium to be among the least stable statuses over time, and that the high-commitment statuses are associated with the most positive adjustment profile. Notably different is the very high prevalence of early closure/closure in our study as compared to that of foreclosure in the earlier studies using Marcia's paradigm. Obviously, this finding requires replication and expansion, for instance by including more domains in an overall identity measure, such as dating relationships, work, religion, and politics. It is also important for future research to examine whether the high prevalence of early closure/closure in our study is due to our use of reconsideration instead of Marcia's original measure of exploration in breadth. Adding a measure of exploration in breadth to the U-MICS could clarify this.

Despite these limitations, the present study has contributed significantly to our understanding of the process of identity formation over time. It is the first five-wave study of a broad-range sample of early-to-middle and

middle-to-late adolescents to show how identity develops between the ages 12 to 20 and to elucidate which identity transitions are most likely to characterize these changes. The makeup of our sample suggests that our findings may be generalizable to individuals who are in various types of education during adolescence. Findings of our study may be less generalizable to adolescents who enter the labor force very early and to adolescents from ethnic minority groups. We also clearly demonstrated that statuses with a very clear resemblance to those from Marcia's model emerged in all waves in both early-to-middle and middle-to-late adolescence. It is hoped that these findings inspire more longitudinal research on identity development.

PART IV

Remaining Issues

9

Summary (in Dutch)

De rode draad in deze dissertatie is de informatieve hypothese. Onderzoekers hebben bepaalde verwachtingen over hoe de werkelijkheid er uit ziet. Het doel van veel onderzoekers is het evalueren van een aantal van deze verwachtingen om te bepalen welke de beste is. Deze verwachtingen kunnen zijn geformuleerd in termen van wat ik een *informatieve* hypothese zal noemen. Dit is een statistische hypothese waarbij ongelijkheidsrestricties zijn gespecificeerd tussen parameters, bijvoorbeeld de ordening tussen drie groepsgemiddelden: $\mu_1 < \mu_2 < \mu_3$, waarbij het teken '$<$' aangeeft dat het eerste gemiddelde (aangegeven met μ_1) lager is dan het tweede gemiddelde (μ_2) dat weer lager is dan het derde gemiddelde (μ_3).

Hoofdstuk 1 van het proefschrift biedt een introductie voor de andere hoofdstukken met een uitgebreide inleiding over wat informatieve hypothesen precies zijn. Ook laat ik zien waarom informatieve hypothesen niet geëvalueerd kunnen worden met klassieke nul hypothese toetsing, een methode die vrijwel alle onderzoekers gebruiken. De resterende proefschrift hoofdstukken zijn vervolgens in drie delen opgesplitst.

Deel 1 biedt een filosofische benadering van informatieve hypothesen, waarbij hoofdstukken 2 en 3 beschrijven waarom het evalueren van informatieve hypothesen beter is dan klassieke nul hypothese toetsing (Hoofdstuk 2) of klassieke model selectie (Hoofdstuk 3). Let op: wanneer ik de

terminolgie *klassieke* nul hypothese toetsing gebruik, dan verwijs ik naar de veel gebruikte klassieke toetsing met behulp van p-waardes waar de nul hypothese $\mu_1 = \mu_2 = \mu_3$ wordt getoetst. En wanneer ik de terminolgie *klassieke* model selectie gebruik, dan verwijs ik naar de veel gebruikte model selectie maten Akaike's informatie criterium, de AIC, (Akaike, 1973, 1981)), de Bayesian informatie criterium, de BIC, (Schwarz, 1978), of de Deviance informatie criterium, de DIC, (Spiegelhalter et al., 2002).

Deel 2 biedt een statistische benadering van informatieve hypothesen. Eerst laat ik zien hoe informatieve hypothesen geevalueerd kunnen worden met behulp van Bayesiaanse model selectie (Hoofdstuk 4). Dit is een methode die al in diverse statistische papers wordt beschreven (zie bijvoorbeeld Hoijtink, Klugkist & Boelen, 2008; Mulder, Klugkist et al., 2009) en al in enkele toegepaste papers wordt gebruikt (zie bijvoorbeeld Meeus, Van de Schoot, Keijsers et al., 2010; Van Well et al., 2009). Hoofdstuk 5 introduceert een nieuwe methode om een informatieve hypothese te toetsten met behulp van een parametrische bootsrap methode. Deze methode biedt de mogelijkheid om informatieve hypothesen te evalueren in (complexe) structurele modellen (ook wel SEM modellen genoemd), iets wat nog niet eerder mogelijk was. Ten slotte wordt in Hoofdstuk 6 een nieuwe model selectie maat afgeleid, de Prior informatie criterium (PIC). Dit is een Bayesiaanse model selectie maat die een sterke link heeft met de Deviance informatie criterium (DIC) van Spiegelhalter et al. (2002).

In Deel 3 worden twee toepassingen gepresenteerd waar een inhoudelijke vraag met behulp van informatieve hypothesen wordt onderzocht. In Hoofdstuk 7 wordt onderzocht hoe delinquent gedrag en zelfbeeld van adolescenten met elkaar samenhangen en of er twee subgroepen bestaan van adolescenten die delinquent gedrag vertonen: een subgroep die tevens een hoog zelfbeeld heeft en een subgroep die juist een laag zelfbeeld heeft. Hoofdstuk 8 gaat over de ontwikkeling van identiteit van adolescenten over de tijd heen en de verwachting daarbij dat over de tijd heen bepaalde

identiteitstypes minder vaak voor gaan komen terwijl andere types toenemen in prevalentie.

Omdat het concept van informatieve hypothesen relatief nieuw is, geef ik hierna een inleiding voor niet statistici over wat deze informatieve hypothesen precies zijn, waarom deze hypothesen beter gebruikt kunnen worden dan klassieke nul hypothese en hoe dit gedaan kan worden met behulp van Bayesiaanse model selectie. Deze inleiding is eerder gepubliceerd het tijdschrift 'De Psyholoog' (Van de Schoot et al., 2009).

Rechtstreeks Verwachtingen Evalueren of de Nul Hypothese Toetsen?

Van de Schoot, R., Hoijtink, H., & S. Doosje

Published, 2009, in *De Psycholoog, 4*, 196-203

9.1 Introductie

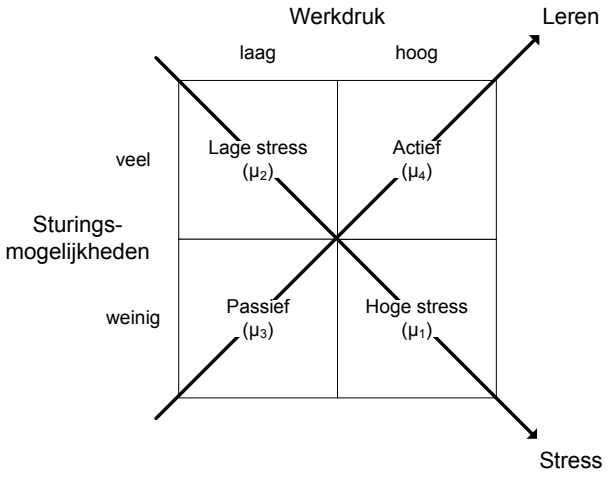
Een praktiserend NIP-psycholoog houdt zijn of haar wetenschappelijke literatuur bij en blijft op de hoogte van recente ontwikkelingen op zijn of haar vakgebied. In dit artikel willen we praktiserend psychologen op de hoogte houden van een minder voor de hand liggende ontwikkeling, namelijk op het gebied van statistiek. In de wetenschappelijke literatuur die een praktiserende psychologen vaak leest, wordt door vrijwel alle onderzoekers klassieke nul hypothese toetsing (NHT) gebruikt om antwoord te geven op de onderzoeksvraag. In dit artikel zullen we laten zien waarom NHT niet per se de beste keuze hoeft te zijn om de onderzoeksvraag te beantwoorden. Zie ook het stuk van Eric-Jan Wagenmakers in De Psycholoog van juli/augustus

2008. We laten zien wat de consequenties zijn als het mis gaat met NHT en introduceren vervolgens een recent ontwikkelde methode die een veelbelovend alternatief biedt voor NHT, namelijk Bayesiaanse Model Selectie (BMS) (Hoijtink & Boom, 2008; Klugkist et al., 2005). Aan de hand van een voorbeeld leggen we kort uit hoe BMS werkt en welke voordelen deze methode biedt ten opzichte van NHT. Om te laten zien hoe NHT 'faalt' en BMS beter werkt, presenteren we een relevant voorbeeld uit de arbeids- en gezondheidspsychologie. Op deze manier wordt duidelijk hoe BMS in de (wetenschappelijke) praktijk gebruikt kan worden.

9.2 Informatieve Hypothesen

Onderzoekers hebben bepaalde verwachtingen over hoe de werkelijkheid er uit ziet. Verwachtingen en hypothesen kunnen gebaseerd zijn op eerder (literatuur-) onderzoek, wetenschappelijk debat of zelfs (subjectieve) menings- verschillen. Het laatste kan bijvoorbeeld als de ene onderzoeker overtuigd is van het effect van een nieuwe interventie of nieuwe behandelmethode die nog niet eerder is onderzocht en een andere onderzoeker niet. Het is natuurlijk wel van belang dat alleen nuttige verwachtingen met elkaar worden vergeleken en niet alle mogelijke verwachtingen. Wij houden juist een pleidooi voor het rechtstreeks evalueren van de voorkennis die een onderzoeker heeft.

Het doel van veel onderzoekers is het evalueren van een aantal van deze verwachtingen om te bepalen welke de beste is. Met andere woorden welke verwachting de meeste steun krijgt van de verzamelde data. Verwachtingen zijn geformuleerd in termen van wat wij *informatieve* hypothesen zullen noemen. Dit omdat er *a priori*, dat is voordat er data zijn verzameld, *informatie* bestaat. Bijvoorbeeld over de ordening tussen twee (of meer) groepsgemiddelden: $\mu_1 < \mu_2$, waarbij het teken '<' aangeeft dat het eerste gemiddelde (μ_1) lager is dan het tweede gemiddelde (μ_2). Wij zullen laten zien dat onderzoekers deze verwachtingen wel *willen* evalueren, maar dit niet zo maar *kunnen* doen. Het is namelijk vrijwel onmogelijk om met



Figuur 9.1: Het interactie model van Karasek

NHT complexe informatieve hypothesen te evalueren. Als onderzoekers dit toch proberen omdat er geen alternatieven voor handen zijn, dan ontstaan er enkele problemen die we nader zullen toelichten aan de hand van een voorbeeld.

9.3 Voorbeeld: Werkdruk, Sturingsmogelijkheden en Verkoudheid

In deze sectie presenteren we een voorbeeld dat we eerst met behulp van NHT evalueren en daarna met BMS. Karasek (1979) stelde dat de gezondheid van werknemers wordt bepaald door combinaties van de mate van werkdruk ('job demands') en de beschikbare sturingsmogelijkheden ('job control'). Daarnaast zijn er twee onderliggende mechanismen die van invloed zijn op de werkdruk en sturingsmogelijkheden, namelijk leren en stress, zie Figuur 9.1.

Karasek voorspelde dat met name een combinatie van hoge werkdruk en weinig sturingsmogelijkheden (een 'hoge stress' werksituatie) het risico

op gezondheidsklachten zou vergroten ten opzichte van een 'lage stress' werksituatie' (lage werkdruk, veel sturingsmogelijkheden) en ten opzichte van een 'actieve' (hoge werkdruk, veel sturingsmogelijkheden) en een 'passieve' (lage werkdruk, weinig sturingsmogelijkheden) werksituatie. In Figuur 9.1 zijn de vier werksituaties weergegeven: *hoge stress*, *lage stress*, *passief*, en *actief*.

Het interactiemodel van Karasek veronderstelt dat de hoge werkdruk een toestand van fysiologische opwinding teweegbrengt, bijvoorbeeld door een verhoogde hartslag en adrenalineproductie, die door de gebrekkige sturingsmogelijkheden niet kan worden omgezet in een effectieve copingrespons (Buunk, de Jonge, Ybema & Wolff, 1998). Omdat er aanwijzingen zijn dat het interactiemodel een goede verklaring biedt van cardiovasculaire klachten (Schnall, Landsbergis & Baker, 1994), zouden we kunnen veronderstellen dat dit ook geldt voor andere ziektebeelden, zoals het risico om verkouden te worden. Karasek's interactiemodel is hiervoor echter niet geheel ondersteund omdat alleen een hoofdeffect van werkdruk (Hao, Duan & Zhang, 2002; Mohren, Swaen, Borm, Bast & Galama, 2001) of van sturingsmogelijkheden (Doosje, Goede, Doornen, Goldstein & Van de Schoot, 2010) werd gevonden. De empirische steun voor de interacties die Karasek veronderstelt is daarom beperkt.

In het voorbeeld voor dit artikel gebruiken we de dataset beschreven in het artikel van (Doosje et al., 2010). De onderzoeksvraag is hoe de vier typen werksituaties die Karasek beschrijft, zie Figuur 9.1, verschillen met betrekking tot het aantal keren verkouden zijn geweest in het afgelopen halfjaar. Daarover hebben we drie verwachtingen opgesteld naar aanleiding van eerder onderzoek.

Verwachting A:

Vanuit de oorspronkelijke theorie van Karasek (1979) zouden we verwachten dat de groep *hoge stress* (μ_1) het ongezondst is ten opzichte van de groepen *lage stress* (μ_2), *passief* (μ_3) en *actief* (μ_4). De groep *hoge stress* heeft dan

een hoger gemiddelde op het aantal keren verkouden zijn in het afgelopen halfjaar dan de overige drie groepen. De informatieve hypothese ziet er dan zo uit, waarbij '>' verwijst naar een hoger gemiddelde en dus vaker verkouden zijn en '=' naar een gelijk gemiddelde:

$$H_A : \mu_1 > \{\mu_2 = \mu_3 = \mu_4\}$$

Verwachting B:

Als werkdruk de meest relevante variabele is, zoals Mohren et al. (2001) en Hao et al. (2002) hebben gevonden, dan is een hoge werkdruk gerelateerd aan het ongezondst zijn en dus vaker verkouden zijn. De groepen *actief* (μ_4) en *hoge stress*(μ_1) zouden dan een hoger gemiddelde hebben dan de andere twee groepen:

$$H_B : \{\mu_1 = \mu_4\} > \{\mu_2 = \mu_3\}$$

Verwachting C:

Zoals is gesuggereerd door Doosje et al. (2010) spelen zowel werkdruk als sturingsmogelijkheden een rol bij verkouden worden. In dat geval zou de groep *hoge stress* (μ_1) een hoger gemiddelde hebben op de variabele verkouden zijn gevolgd door de groep *passief* (μ_3), gevolgd door de groep *actief* (μ_4). De groep *lage stress* (μ_2) zou dan het minst vaak verkouden zijn omdat zij weinig stress en veel sturingsmogelijkheden hebben. De bijbehorende hypothese ziet er dan zo uit:

$$H_C : \mu_1 > \mu_3 > \mu_4 > \mu_2$$

Om erachter te komen welke van deze drie verwachtingen het meest waarschijnlijk is, is in het artikel van Doosje et al. (2010) een variantie analyse (ANOVA) uitgevoerd. In Tabel 9.1 zijn de groepsgemiddelden weergegeven. Er bleken significante verschillen te bestaan tussen de vier groepen ($F(3) = 9.51, p < .001$) en post-hoc analyse met Bonferroni correctie laat zien dat sommige maar niet alle groepen onderling van elkaar verschillen,

zie Tabel 9.1. Als twee gemiddeldes dezelfde letter hebben, dan is het verschil significant. Zo hebben μ_1 en μ_2 beide de letter A en verschillen significant van elkaar ($p < .05$).

Voor de ANOVA is de volgende nul hypothese getoetst: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$. Er werden significante groepsverschillen gevonden en de nul hypothese is verworpen. Merk op dat dit tot nu toe nog steeds geen informatie geeft over welke van de informatieve hypothese (H_A, H_B, H_C) de beste is. Om hierover toch een uitspraak te doen kan gekeken worden naar de ordening van de gemiddelden uit Tabel 9.1:

$$\mu_3 > \mu_1 > \mu_4 > \mu_2 \ .$$

Er zijn ook post-hoc toetsen uitgevoerd en als we bij niet significant resultaat de ordening van de gemiddeldes aanpassen, dan wordt de ordening:

$$\{\mu_3 = \mu_1\} > \{\mu_4 = \mu_2\} \ .$$

Het is nu op basis van dit resultaat erg lastig om te kiezen tussen de drie informatieve hypothesen H_A, H_B, en H_C. Geen van de hypothesen worden namelijk volledig ondersteund door de gevonden ordening van de data. De resultaten geven slechts in meer of mindere mate steun voor elk van de informatieve hypothesen. Voor H_A geldt dat μ_1 groter is dan μ_2 en μ_4, maar niet groter is dan μ_3. Alleen de gemiddeldes μ_1 en μ_4 voor H_B zijn groter dan μ_2, maar dit geldt niet voor μ_3. Beide hypothesen worden dus niet echt ondersteund door de data. Hypothese C komt echter dicht in de buurt, zeker als gekeken wordt naar de ordening op basis van de groepsgemiddelden. Als echter naar de significante resultaten wordt gekeken, dan klopt ook deze hypothese niet meer.

Hypothese C lijkt dus op het eerste gezicht de meeste steun te krijgen, maar de vraag is of dit ook werkelijk zo is. Het is nog lastiger, of zelfs onmogelijk om te zeggen hoeveel waarschijnlijker de ene hypothese is ten opzichte van een andere. Met andere woorden, nul hypothese toetsing geeft in dit voorbeeld geen bevredigend antwoord op de onderzoeksvraag, in de volgende sectie gaan we nader in op wat er precies mis gaat.

Tabel 9.1: Groepsgemiddelden en standaarddeviaties (SD) voor de vier groepen van het Karasek model

	Hoge stress (μ_1)	Lage stress (μ_2)	Passief (μ_3)	Actief (μ_4)
Gemiddelde	2.37^{ab}	2.17^{ac}	2.42^{cd}	2.18^{bd}
SD	1.03	0.93	0.98	0.97
n	289	594	517	638

noot: Gemiddelden met dezelfde letter verschillen significant van elkaar ($p < .05$)

9.4 Wat Gaat er Mis?

Er is door de tijd heen veel literatuur verschenen met kritiek op het gebruik van NHT en het gebruik van p-waarden (Cohen, 1990, 1992, 1994; Balluerka et al., 2005; Krantz, 1999; Rozenboom, 1960; Sterne & Smith, 2001; M. D. Lee & Wagenmakers, 2005). Wij zullen ons voornamelijk richten op waar het mis gaat bij het evalueren van informatieve hypothesen met behulp van NHT.

Bij NHT is de hypothese die daadwerkelijk getoetst wordt de bekende nul hypothese *er is niks aan de hand* versus het alternatief *er gebeurt iets, maar we weten niet wat.* In het eerste voorbeeld van de vorige sectie was de onderzoeksvraag welke informatieve hypothese het meest waarschijnlijk was H_A, H_B of H_C:

$$\begin{aligned} H_A: &\quad \mu_1 > \{\mu_2 = \mu_3 = \mu_4\} \ , \\ H_B: &\quad \{\mu_1 = \mu_4\} > \{\mu_2 = \mu_3\} \ , \\ H_C: &\quad \mu_1 > \mu_3 > \mu_4 > \mu_2 \ . \end{aligned}$$

De hypothesen die daadwerkelijk getoetst worden met NHT zijn echter:

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 = \mu_3 = \mu_4 \ , \\ H_1: &\quad \text{niet } H_0 \ . \end{aligned}$$

Merk op dat deze nul (H_0) en alternatieve hypothese (H_1) niet hetzelfde zijn als de informatieve hypothesen H_A, H_B en H_C die de onderzoekers eigenlijk

wilden evalueren. Als de nul hypothese en de alternatieve hypothese geen onderdeel zijn van de onderzoeksvraag, dan is er geen directe relatie tussen de hypothesen waar een onderzoeker in geïnteresseerd is en de hypothesen die daadwerkelijk getoetst worden met NHT. De resultaten van NHT geven in dat geval geen antwoord op de onderzoeksvraag.

Daar komt bij dat onderzoekers vaak helemaal niet geïnteresseerd zijn in de nul hypothese. Onderzoekers hebben namelijk vrijwel altijd verwachtingen over hoe de relatie tussen variabelen eruit zou moeten zien. Het is dan raar dat een nul hypothese 'er is niks aan de hand' wordt getoetst aangezien de onderzoeker van te voren al weet dat er 'iets' aan de hand is. Het is dan veel logischer om de informatieve hypothesen rechtstreeks te evalueren in plaats van de nul hypothese te toetsen.

Als dan toch een nul hypothese wordt getoetst, dan wordt de traditionele p-waarde gebruikt om deze nul hypothese te verwerpen of niet te verwerpen. Het omslagpunt van deze dichotome beslissing ligt bij de welbekende waarde van $p < .05$. Deze drempelwaarde van .05 is niet alleen willekeurig gekozen (zie bv: Cohen, 1994; Rozenboom, 1960), maar laat alleen ruimte voor de conclusie dat een nul hypothese wel of niet wordt verworpen met niks daar tussenin. Dit kan leiden tot vreemde beslissingen, bijvoorbeeld in het geval dat een p-waarde $p = .051$ of $p = .049$ is. In het eerste geval wordt de nul hypothese niet verworpen en in het tweede geval wel. Het mag duidelijk zijn dat beide situaties niet veel van elkaar verschillen. Het is dan vreemd dat de conclusie voor beide situaties totaal anders is.

Wanneer de nul hypothese wordt verworpen, dan weten we eigenlijk nog steeds niks over de informatieve hypothesen aangezien de alternatieve hypothese geen informatie bevat over de ordening tussen de gemiddelden. Ook een visuele inspectie van bijvoorbeeld de groepsgemiddelden is niet altijd voldoende en is in ieder geval subjectief. Hoe kan een onderzoeker dan toch uitspraken doen over de informatieve hypothesen? Het resultaat zou geen dichotome ja/nee beslissing moeten zijn, maar een kans per hypothese

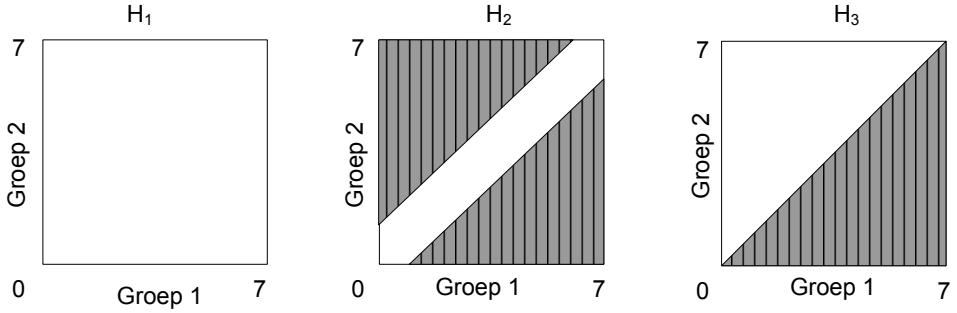
dat deze de beste is. In de volgende sectie presenteren we een methode, Bayesiaanse model selectie, die hiertoe wel in staat is.

9.5 Bayesiaanse Model Selectie

Omdat veel artikelen over BMS lastig te lezen zijn voor een niet-statisticus, geven wij een zeer korte en vereenvoudigde introductie. Hiervoor gebruiken we een tweede voorbeeld dat is gebaseerd op Doosje et al. (2010) met slechts twee groepen en 1 variabele, zodat we ook grafisch kunnen weergeven wat er gebeurt. Voor een uitgebreidere introductie en voor een overzicht van publicaties zie Hoijtink, Klugkist and Boelen (2008) en voor een meer technische introductie Klugkist et al. (2005).

Stel dat we de groep *lage stress* (μ_1) willen vergelijken met de groep *hoge stress* (μ_2) op het aantal keren verkouden zijn geweest in het afgelopen halfjaar. En stel dat we de volgende drie hypothesen hebben: (H_1) er is geen verwachting over de twee groepen; (H_2) beide groepen hebben dezelfde score; en (H_3) de groep *lage stress* is minder vaker verkouden dan groep de groep *hoge stress*. De hypothesen zien er dan zo uit:

$$\begin{aligned} H_1: &\quad \mu_1 \ , \ \mu_2 \ , \\ H_2: &\quad \mu_1 \ = \ \mu_2 \ , \\ H_3: &\quad \mu_1 \ < \ \mu_2 \ . \end{aligned}$$



Figuur 9.2: Voorkennis vertaald in Informatieve Hypothesen

Om erachter te komen welke van de drie hierboven beschreven hypothesen het meest waarschijnlijk is, gaan we deze evalueren met zogenaamde posterior model kansen (PMK). Om deze PMK's uit te rekenen zijn drie ingrediënten nodig, namelijk (1) de voorkennis die een onderzoeker heeft, (2) de likelihood (waarschijnlijkheid) van de data en (3) de steun in de data voor elk van de hypothesen.

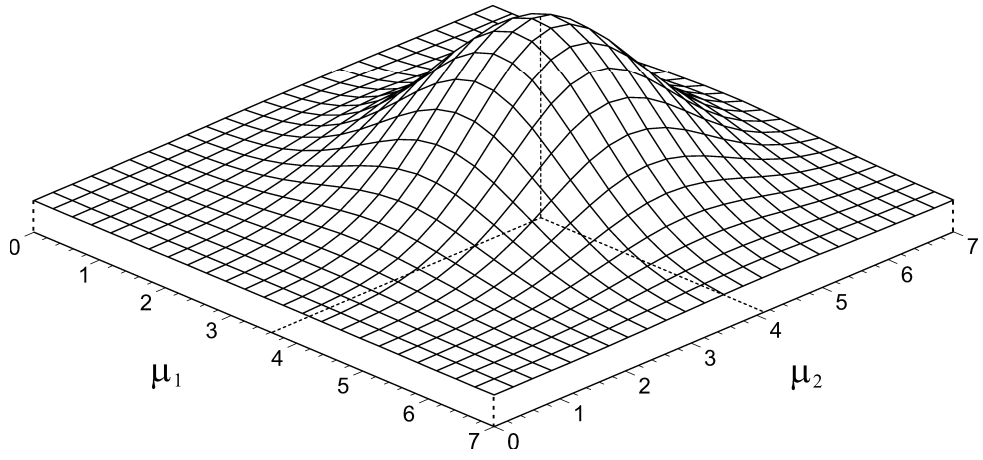
Het eerste ingrediënt is de kennis die er is over de ordening van de gemiddelde scores van de groepen *lage stress* (μ_1) en *hoge stress* (μ_2), voordat de data zijn gezien. Dit zijn de opgestelde hypothesen welke in Figuur 9.2 grafisch zijn weergegeven. Het vierkant vertegenwoordigt alle mogelijke combinaties van gemiddelden die beide groepen kunnen hebben op de variabele verkouden zijn.

Voor elke hypothese bepalen we nu wat de toegestane ruimte is binnen dit vierkant. Met andere woorden, we bepalen welke combinaties van gemiddelden toegestaan zijn voor elk van de opgestelde hypothesen. Voor Hypothese 1 is de gehele ruimte mogelijk, alle combinaties van gemiddelde scores op *verkouden* zijn toegestaan. Voor Hypothese 2 is een groot deel van het vierkant niet toegestaan, er zijn namelijk alleen combinaties van gemiddelden mogelijk waar beide groepen aan elkaar gelijk zijn, dit is de diagonaal van Figuur 9.2. Voor Hypothese 3 is het gedeelte van het vierkant toegestaan waar de groep *hoge stress* hoger scoort op het risico om verkouden

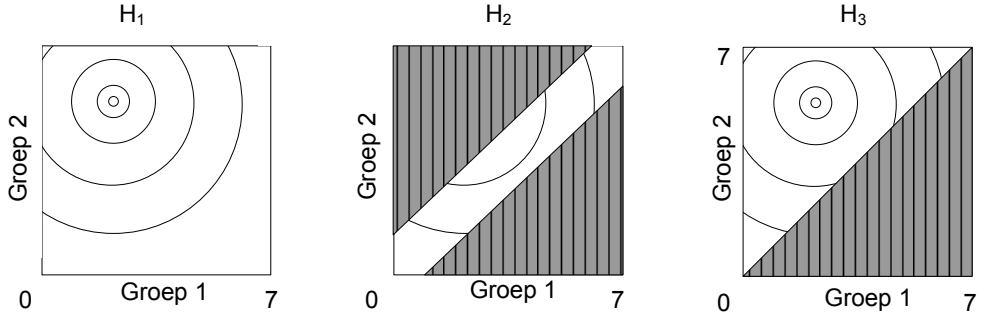
te zijn dan de groep *lage stress*. Dit is de witte driehoek in Figuur 9.2. Binnen deze ruimte is elke combinatie van gemiddelden even waarschijnlijk. Dit wordt ook wel een uniforme prior verdeling genoemd die over de toegestane ruimte is gelegd, zie voor meer informatie hierover Klugkist et al. (2005).

Het tweede ingrediënt is de informatie die er aanwezig is in de data over de waarschijnlijkheid van mogelijke combinaties in de populatie. Dit wordt ook wel likelihood van de data genoemd en kan gezien worden als een landschap met daarin een piek op de plek waar de meest waarschijnlijke combinatie van de gemiddelden zich bevindt, zie Figuur 9.3. Als we een uitspraak willen doen over het gemiddelde van de groepen *hoge stress* en *lage stress* in de populatie, dan zijn de gemiddelden die in de data zijn geobserveerd, het meest waarschijnlijk. Stel dat bijvoorbeeld $\mu_1 = 3.6$ en $\mu_2 = 4.1$ de gemiddeldes zijn die in de data set zijn geobserveerd, dan is de kans dat in de populatie groep 1 een gemiddelde heeft van 4.6 en groep 2 een gemiddelde van 3.1, maximaal is. Merk op dat dit fictieve waardes zijn gebaseerd op een hypothetische data set. De maximale waarschijnlijkheid, maximum likelihood genoemd, is de piek van de curve in Figuur 9.3. Combinaties van gemiddelden die verder afliggen van deze piek zullen steeds minder waarschijnlijk zijn in de populatie en leiden tot een steeds lagere curve in Figuur 9.3. De kans dat bijvoorbeeld in de populatie de groepsgemiddelden een waarde hebben van $\mu_1 = 6.2$ en $\mu_2 = 1.8$ is veel kleiner, wat te zien is aan de lagere curve op dit punt in de grafiek.

Het derde ingrediënt is de berekening van de hoeveelheid steun die er is in de data voor elk van de hypothesen. Dit wordt gedaan door de gemiddelde hoogte van de likelihood uit te rekenen binnen de toegestane parameter ruimte. Om dit grafisch weer te geven leggen we Figuur 9.2 op Figuur 9.3 wat resulteert in Figuur 9.4. In deze laatste figuur is te zien hoeveel er van de likelihood in de toegestane ruimte van het vierkant ligt. Voor elk van de drie hypothesen kan vervolgens uitgerekend worden hoe groot de gemiddelde hoogte van de likelihood is in deze toegestane ruimte. Een groot deel van het lagere gebied van de likelihood valt bijvoorbeeld buiten de



Figuur 9.3: Likelihood van de data



Figuur 9.4: De voorkennis en data met elkaar gecombineerd

toegestane ruimte die bij Hypothese 3 hoort. Dit lagere gebied wordt echter wel meegenomen in de berekening van Hypothese 1 omdat hier het gehele oppervlak van het vierkant toegestaan is. Hierdoor zal de gemiddelde hoogte van de likelihood voor Hypothese 3 een stuk hoger zijn dan voor Hypothese 1. De gemiddelde hoogte voor Hypothese 2 zal juist heel erg klein zijn ten opzichte van Hypothese 1 en 3, omdat een groot gedeelte van de likelihood inclusief de piek van de curve niet in het toegestane gebied van Hypothese 2 ligt.

Tabel 9.2: Resultaten BMS voor Voorbeeld 2

Hypothese	PMK
H_1	.31
H_2	.08
H_3	.61

De drie ingrediënten die we hiervoor besproken hebben, worden omgezet in posterior model kansen (PMK's). Wanneer de gemiddelde hoogte van de likelihood groter is, dan is er dan meer steun van de data voor de hypothese wat vervolgens resulteert in een hogere PMK. Een PMK houdt niet alleen rekening met hoe goed de hypothese bij de data past, maar ook hoe complex de hypothese is. Dit resulteert in één enkel getal per hypothese op een kansschaal en kan geïnterpreteerd worden als de waarschijnlijkheid per hypothese dat deze de beste is van alle hypothesen die onderzocht worden. Een gebruiker van BMS hoeft alleen de hypothesen te specificeren in termen van restricties tussen de statistische parameters, zoals $\mu_1 < \mu_2$, en de dataset aan te leveren, de bijbehorende software levert de uitkomsten van de analyses (zie $http://www.fss.uu.nl/ms/informativehypothesis$).

De berekening van de PMK's geschiedt aan de hand van Bayes Factors, uitgevonden door Thomas Bayes in 1764 (Bayes, 1764) en is verder ontwikkeld in 1774 door Laplace (zie: Laplace's 1774 Memoir on Inverse Probability in: Stigler, 1986). Het was pas in de 20ste eeuw dat de Bayesiaanse benadering opnieuw ontdekt werd door o.a. Ramsey, de Finetti, Jeffreys, en Jaynes (voor een overzicht zie: Corfield & Williamson, 2001). Pas op het eind van de vorige werden computers snel genoeg om de berekeningen ook daadwerkelijk uit te voeren (zie b.v.: Bayarri & Berger, 2000; Kass & Raftery, 1995; Raftery, 1995). Het omzetten van de drie ingrediënten in Bayes Factors en daarna in PMK's is uitgebreid beschreven in het boek van Hoijtink, Klugkist and Boelen (2008).

De resultaten van voorbeeld 2 (zie Tabel 9.2) laten zien dat Hypothese 3 de meeste ondersteuning krijgt door de informatie die in de data set zit

en dus de beste hypothese is vergeleken met de andere twee hypothesen. De conclusie is dan dat de verwachting de *hoge stress* groep hoger scoort dan de de *lage stress* groep op verkouden zijn het beste is met een waarschijnlijkheid van .61. De kans dat deze conclusie niet correct is, is 1 - .61 = .39. Het is nu aan de onderzoeker om te beslissen of een waarschijnlijkheid van .61 en een foutmarge van .31 een interessante conclusie oplevert.

Het lijkt misschien dat er een grote fout marge is, merk echter wel op dat Hypothese 3 ongeveer 8 keer zo waarschijnlijk is als Hypothese 2. Dit kan op zichzelf al een bevredigend resultaat zijn. Dat Hypothese 3 'slechts' 2 keer zo waarschijnlijk is als een model zonder enige beperkingen (Hypothese 1), is niet eens zo slecht. Dit omdat het enige verschil tussen beide hypothesen slechts 1 restrictie is. De conclusie dat Hypothese 3 de beste hypothese is in deze model selectie competitie is dus geoorloofd. Het zou overigens best kunnen zijn dat bepaalde resultaten van BMS een ongeveer even grote PMK opleveren, bijvoorbeeld bij een PMK's van .49 en .50. In dit geval moet de onderzoeker terug naar de tekentafel en moet er gezocht worden naar een betere verwachting die wellicht meer ondersteuning krijgt van de data. Deze nieuwe hypothese kan dan worden toegevoegd aan de reeds bestaande set van hypothesen. Dit geldt natuurlijk ook als een andere onderzoeker een andere hypothese er op na houdt en zijn of haar eigen verwachting wil toevoegen.

Wat hebben we nu gedaan? We hebben de voorkennis voordat we data hebben verzameld, vertaald in een set van hypothesen. We hebben daarna uitgerekend hoe waarschijnlijk deze hypothesen zijn nadat we de data hebben gezien. Hierdoor is duidelijk gemaakt welke verwachting het beste wordt ondersteund door de data en ook hoeveel onzekerheid hierover bestaat. In de volgende sectie passen we deze methode toe op het uitgebreidere Karasek voorbeeld.

Tabel 9.3: Resultaten van BMS voor Voorbeeld 1

Hypothese	PMK
H_A	.11
H_B	.01
H_C	.88

9.6 Voorbeeld Opnieuw Geanalyseerd

De verwachtingen van voorbeeld 1 zijn met behulp van BMS geanalyseerd. Ter herinnering: Verwachting A stelt dat de groep *hoge stress* het vaakst verkouden is ten opzichte van de groepen *lage stress*, *passief* en *actief*; Verwachting B stelt dat de groepen *actief* en *hoge stress* het vaakst verkouden zijn dan de andere twee groepen; Verwachting C stelt dat de groep *hoge stress* het vaakst verkouden is gevolgd door respectievelijk de groep *passief*, *actief* en *lage stress*.

In Tabel 9.3 is voor elk van de drie verwachtingen aangegeven hoe waarschijnlijk deze is. Hypothese C heeft de hoogste waarschijnlijkheid en heeft een kans van .88 dat dit de beste hypothese is en een kans van .12 dat dit niet zo is.

Geconcludeerd kan worden dat de groep *hoge stress* (μ_1) het vaakst verkouden is. Mensen die gekarakteriseerd worden door veel stress op het werk en maar weinig sturingsmogelijkheden hebben, lopen dus het grootste risico op verkoudheid. Deze groep wordt gevolgd door mensen die een lage werkdruk ervaren maar tevens ook weinig sturingsmogelijkheden hebben. Heb je echter veel sturingsmogelijkheden dan heb je minder kans om verkouden te zijn, en heb je ook nog eens weinig stress op het werk, dan ben je relatief het gezondst en ben je het minst vaak verkouden.

9.7 Conclusie

Met klassieke nul hypothese toetsting moet een hele stapel output geëvalueerd worden om een onderzoeksvraag te beantwoorden: F-toets, post-hoc toetsen, groepsgemiddelden, etc. Deze stapel output kan makkelijk leiden tot verwarrende resultaten. Daarnaast geven de resultaten van NHT geen direct antwoord op de onderzoeksvraag en kunnen informatieve hypothesen niet direct met elkaar vergeleken worden, iets dat met BMS wel kan. Ook wanneer het niet voor de hand ligt welke hypothese de meeste steun krijgt van de data, bijvoorbeeld wanneer de ordening van de groepsgemiddelden niet geheel overeen komt met elk van de hypothesen, geeft BMS nog steeds een interpreteerbaar resultaat. Zelfs bij veel complexere onderzoeksvragen dan in dit artikel besproken, bijvoorbeeld met meerdere (on)afhankelijke variabelen, meerdere meetmomenten over de tijd heen, covariaten, meer groepen, enz., geeft BMS nog steeds een enkel getal per hypothese.

Bayesiaanse model selectie (BMS) resulteert in makkelijk te interpreteren resultaten en geeft een precies antwoord op de onderzoeksvraag. Dat is namelijk per verwachting/hypothese de kans dat deze hypothese de beste hypothese is en dus de meeste steun krijgt van de data. BMS is daardoor een veelbelovend alternatief voor NHT en zal steeds vaker opduiken in de wetenschappelijk literatuur.

References

References

Adams, G. & Fitch, S. A. (1982). Ego stage and identity status development: A cross-sequential analysis. *Journal of Personality and Social Psychology*, *43*, 574-583.

Adams, G. & Jones, R. (1983). Female adolescents identity development: Age comparisons and perceived child-rearing experience. *Developmental Psychology*, *19*, 249-256.

Agnew, R. (1992). Foundation for a generalism strain theory of crime and delinquency. *Criminology*, *30*, 47-87.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (p. 267 - 281). Budapest: Akademiai Kiado.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, *16*, 3 - 14.

Andrews, D. W. K. (1996). Admissibility of the likelihood ratio test when the parameter space is restricted under the alternative. *Econometrica*, *64*, 705-718.

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, *68*, 399-405.

Anraku, K. (1999). An information criterion for parameters under a simple order restriction. *Journal of the Royal Statistical Society, series B*, *86*, 141-152.

Archer, S. (1982). The lower age boundaries of identity development. *Child Development*, *53*, 1551-1556.

Balluerka, N., Gómez, J. & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology*, *1*, 55-70.

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

Batson, C. D., Lishner, D. A., Cook, J. & Sawyer, S. (2005). Similarity and nurturance: Two possible sources of empathy for strangers. *Basic and Applied Social Psychology*, *27*, 15-2.

Baumeister, R. F., Boden, J. M. & Smart, L. (1996). Relation of threatened egotism to violence and aggression: The dark side of high self-concept. *Psychological Review*, *103*, 5-33.

Baumeister, R. F., Campbell, J. D., Krueger, J. I. & Vohs, K. D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? *American Psychological Society*, *4*, 11-44.

Bayarri, M. J. & Berger, J. O. (2000). P-values for composite null models. *Journal of the American Statistical Association*, *95*, 1127-1142.

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370-418.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289-300.

Berger, J. & Pericchi, L. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*,

109-122.

Berger, J. & Pericchi, L. R. (2004). Training samples in objective bayesian model selection. *The Annals of Statistics*, *32*, 841-869.

Berger, V. W. & Ivanova, A. (2002). The bias of linear rank tests when testing for stochastic order in ordered categorical data. *Journal of Statistical Planning and Inference*, *107*, 237 - 247.

Berman, S. L., Weems, C. F. & Stickle, T. R. (2006). Existential anxiety in adolescents: Prevalence, structure, association with psychological symptoms and identity development. *Journal of Youth and Adolescence*, *35*, 303-310.

Berzonsky, M. & Adams, G. (1999). Reevaluating the identity status paradigm: Still useful after 35 years. *Developmental Review*, *19*, 557-590.

Beunen, G., Thomis, M., Maes, H. H., Loos, R., Malina, R. M. & Claessens, A. L. (2000). Genetic variance of adolescent growth in stature. *Annals of Human Biology*, *27*, 173-186.

Boden, J. M., Fergusson, D. M. & Horwood, J. (2007). Self-esteem and violence: testing links between adolescent self-esteem and later hostility and violent behaviour. *Social Psychiatry and Psychiatric Epidemiology*, *42*, 881-891.

Bollen, K. (1989). *Structural equation modeling with latent variables*. New York: Wiley.

Box, G. E. P. (1980). Sampling and bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, series A*, *143*, 383 - 430.

Burnham, K. P. & Anderson, D. R. (1998). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods Research*, *33*, 261 - 304.

Bushman, B. J. & Baumeister, R. F. (1998). Threatened egotism, narcissism, self-esteem, and direct and displaced aggression: Does self-love or self-hate lead to violence? *Journal of Personality and Social Psychology*, *75*, 219-229.

Buunk, B. P., de Jonge, J., Ybema, J. F. & Wolff, C. J. (1998). Psychosocial aspects of occupational stress. In H. Thierry, P. J. Drenth, P. J. Willems & C. J. de Wolff (Eds.), *Handbook of work and organizational psychology.* Hove, England: Psychology Press/Erlbaum (UK) Taylor & Francis.

Campbell, J. D. (1990). Self-esteem and clarity of the self-concept. *Journal of Personality and Social Psychology*, *59*, 538-59.

Carroll, A., Houghton, S., Wood, R., Perkins, C. & Bower, J. (2007). Multidimensional self-concept, age and gender differences in australian high school students involved in delinquent activities. *School Psychology International*, *28*, 237-256.

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, *90*, 131-1321.

Chongcharoen, S., Singh, B. & Wright, F. (2002). Powers of some one-sided multivariate tests with the population covariance matrix known up to a multiplicative constant. *Journal of Statistical Planning and Inference*, *107*, 103 - 121.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Coie, J. D. & Dodge, K. A. (1998). Aggression and antisocial behavior. In W. Damon & N. Eisenberg (Eds.), *Handbook of child psychology* (Vol. 3, p. 779-862).

Cole, P. G., Chan, L. K. S. & Lytton, L. (1989). The perceived competence of juvenile delinquents and non-delinquents. *Journal of Special Education*,

23, 294-302.

Colom, R. & Lynn, R. (2004). Testing the developmental theory of sex differences in intelligence on 12-18 year olds. *Personality and Individual Differences*, *36*, 75-82.

Corfield, D. & Williamson, J. (Eds.). (2001). *Foundations of bayesianism* (Vol. 24). London: Kluwer Academic Publishers.

Côté, J. & Schwartz, S. (2002). Comparing psychological and sociological approaches to identity: identity status, identity capital, and the individualization process. *Journal of Adolescence*, *25*, 571-586.

Crocetti, E., Berzonsky, M. & Meeus, W. (2008). A person-centered approach to identity styles. *Manuscript submitted for publication.*.

Dayton, C. M. (2003). Information criteria for pairwise comparisons. *Psychological Methods*, *8*, 61-71.

De Bruyn, E. E. J., Vermulst, A. A. & Scholte, R. H. J. (2003). The nijmegen problem behaviour list: Construction and psychometric characteristics. *Manuscript submitted for publication*.

De Graaf, N. D., De Graaf, P. M. & Kraaykamp, G. (2000). Parental cultural capital and educational attainment in the netherlands: A refinement of the cultural capital perspective. *Sociology of Education*, *73*, 92-111.

Deković, M., Noom, M. J. & Meeus, W. (1997). Expectations regarding development during adolescence: Parental and adolescent perceptions. *Journal of Youth and Adolescence*, *26*, 253 - 272.

Deković, M., Wissink, I. & Meijer, A. M. (2004). The role of family and peer relations in adolescent antisocial behaviour: comparison of four ethnic groups. *Journal of Adolescence*, *27*, 497 - 514.

Dellas, M. & Jernigan, L. P. (1987). Occupational identity status development, gender comparisons, and internal-external control in first-year air force cadets. *Journal of Youth and Adolescence*, *16*, 587-600.

Derkse, W. (1993). *On simplicity and elegance*. Delft: Eburon.

Diamantopoulou, S., Rydell, A.-M. & Henricsson, L. (2008). Can both low and high self-esteem be related to aggression in children? *Social*

Development,, *17*, 682-698.

Dishion, T. J. & McMahon, R. J. (1998). Parental monitoring and the prevention of child and adolescents problem behavior: A conceptual and empirical foundation. *Clinical Child and Family Psychology Review*, *1*, 61 - 75.

Dodge, K. A. (1985). Advances in cognitive-behavioral research and therapy. In P. C. Kendall (Ed.), (p. 73-110). Orlando, FL: Academic.

Donnellan, M. B., Trzesniewski, K. H., Robins, R. W., Moffitt, T. E. & Caspi, A. (2005). Low self-esteem is related to aggression, antisocial behavior, and delinquency. *Psychological Science*, *16*, 328-335.

Doosje, S., Goede, M. D., Doornen, L. V., Goldstein, J. & Van de Schoot, R. (2010). Humorous coping styles, job characteristics and job-related affect as predictors of the incidence of upper respiratory infection. *South African Journal of Industrial Psychology*.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Erikson, E. (1968). *Identity, youth and crisis.* New York: Norton.

Fandrem, H., Strohmeier, D. & Roland, E. (2009). Bullying and victimization among native and immigrant adolescents in norway. the role of proactive and reactive aggressiveness. *The Journal of Early Adolescence*, *6*, 898-923.

Fergusson, D. M. & Horwood, L. J. (2002). Male and female offending trajectories. *Development and Psychopathology*, *14*, 159-177.

Feshbach, N. (1975). Empathy in children: Some theoretical and empirical considerations. *Counseling Psychologist*, *5*, 25-30.

Forster, M. R. (2002). The new science of simplicity. In A. Zellner, H. A. Kreuzenkamp & M. Mc Aleer (Eds.), *Simplicity, inference, and modelling* (p. 83-119). Cambridge: Cambridge University Press.

Forster, M. R. & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, *45*, 1-35.

Galindo-Garre, F. & Vermunt, J. (2004). The order-restricted association model: Two estimation algorithms and issues in testing. *Psychometrika*, *69*, 641 - 654.

Galindo-Garre, F. & Vermunt, J. (2005). Testing log-linear models with inequality constraints: A comparison of asymptotic, bootstrap, and posterior predictive p-values. *Statistica Neerlandica*, *59*, 82 - 94.

Gelfand, A. E., Smith, A. F. M. & Lee, T. (1992). Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association*, *87*, 523-532.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman&HallCRC.

Gelman, A., Meng, X. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, *6*, 733 -807.

Gerris, J. R. M., Houtmans, M. J. M., Kwaaitaal-Roosen, E. M. G., Schipper, J. C., Vermulst, A. A. & Janssens, J. M. A. M. (1998). *Parents, adolescents, and young adults in dutch families: A longitudinal study.* (Tech. Rep.). The Netherlands: Institute of Family Studies, University of Nijmegen.

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X. & Zijdenbos, A. (1999). Brain development during childhood and adolescence: A longitudinal mri study. *Nature Neuroscience*, *2*, 861-863.

Gonzalez, R. & Griffin, D. (2001). Testing parameters in structural equation modelling: Every "One" matters. *Psychological Methods*, *6*, 258 - 269.

Gottfredson, M. R. & Hirschi, T. (1990). *A general theory of crime.* Stanford University Press Stanford, California.

Greenberger, E. & Chen, C. (1996). Perceived family relationships and depressed mood in early and late adolescence: A comparison of european and asian american. *Developmental Psychology*, *32*, 707 - 716.

Grotevant, H. (1987). Toward a process model of identity formation. *Journal of Adolescent Research*, *2*, 203-222.

Guerra, A. & Braungart-Rieker, J. (1999). Predicting career indecision in college students: The role of of identity formation and parental relationship factors. *Career Development Quarterly*, *47*, 255-266.

Halpern, J. Y. (2003). *Reasoning about uncertainty.* MIT Press.

Hamaker, E. L., van Hattum, P., Kuiper, R. & Hoijtink, H. (2009). Handbook of advanced multilevel analysis. In K. R. . J. Hox (Ed.), (chap. Model Selection Based on Information Criteria in Multilevel Modeling). Taylor and Francis.

Han, S. (1977). A globally convergent method for nonlinear programming. *Journal of Optimization Theory and Applications*, *22/23*, 297 - 309.

Hao, Q., Duan, Z. & Zhang, A. (2002). Relations between the occupational stress of nurses and their salivary immunoglobulin a level. *Chinese Nursing Research*, *16*, 207-208.

Harter, S. (1983). Developmental perspectives on the self-concept. In M. Heatherington (Ed.), *Handbook of child psychology: Social and personality development.* New York: Wiley.

Harter, S. (1987). The determinants and mediational role of global self-worth. In N. Eisenberg (Ed.), *Contemporary topics in developmental psychology* (p. 219-242). Wiley: New York.

Harter, S. (1990). Causes, correlates, and the functional role of global self-worth: A life-span perspective. In R. J. Sternberg & J. J. Kolligan (Eds.), *Competence considered.* (p. 67-69). New Haven, CT: Yale University Press.

Harter, S. (1999). *The construction of the self. a developmental perspective.* New York, London: The Guilford press.

Henderson, L., Goodman, N. D., Tenenbaum, J. B. & Woodward, J. F. (2010). The structure and dynamics of scientific theories: A hierarchical bayesian perspective. *Philosophy of Science*, *77*, 172-200.

Hoijtink, H. (1998). Constrained latent class analysis using the gibbs sampler and posterior predictive p-values: Applications to educational testing. *Statistica Sinica*, *8*, 691-712.

Hoijtink, H. (2000). Posterior inference in the random intercept model based on samples obtained with markov chain monte carlo methods. *Computational Statistics*, *3*, 315-336.

Hoijtink, H. (2001). Confirmatory latent class analysis: Model selection using bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research*, *36*, 563-588.

Hoijtink, H. & Boom, J. (2008). Latent class models specified using inequality constraints. In H. Hoijtink, I. Klugkist & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses.* New-York: Springer.

Hoijtink, H., Huntjes, R., Reijntjes, A., Kuiper, R. & Boelen, P. (2008). An evaluation of bayesian inequality constrained analysis of variance. In H. Hoijtink, I. Klugkist & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypothesis* (p. Chap. 5). New York : Springer.

Hoijtink, H. & Klugkist, I. (2007). Comparison of hypothesis testing and baysian model selection. *Quality and Quantity*, *41*, 73-91.

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

Hoijtink, H. & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the gibbs sampler and posterior predictive checks. *Psychometrika*, *62*, 171-190.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65-70.

Hothorn, L., Vaeth, M. & Hothorn, T. (2009). Trend tests for the evaluation of exposure-response relationships in epidemiological exposure studies. *Epidemiologic Perspectives & Innovations*, *6*.

Howard, G. S., Maxwell, S. E. & Fleming, K. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and bayesian analysis. *Psychological Methods*, *5*, 315-

332.

Hsu, J. (1996). *Multiple comparisons.* Chapman and Hall, London.

Jang, S. J. & Thornberry, T. P. (1998). Self-esteem, delinquent peers, and delinquency: A test of the self-enhancement hypothesis. *American Sociological Review*, *63*, 586-598.

Jaynes, E. T. (2003). *Probability theory, the logic of science.* Cambridge: Cambridge University Press.

Jennissen, R. P. W. & Blom, M. (2005). *Allochtone en autochtone verdachten van verschillende delicttypen nader bekeken.* (Tech. Rep.). Beschikbaar via: www.wodc.nl (Opgevraagd: 03-10-2007). Den Haag: WODC/CBS.

Jolliffe, D. & Farrington, D. P. (2004). Empathy and offending: A systematic review and meta-analysis. *Aggression and Violent Behaviour*, *9*, 441-476.

Jongmans, M. J., Smits-Engelsman, B. C. M. & Schoenmaker, M. M. (2003). Consequences of comorbidity of developmental coordination disorders and learning disabilities for severity and pattern of perceptualmotor dysfunction. *Journal Of Learning Disabilities*, *36*, 528-537.

Kammers, M., Mulder, J., De Vignemont, F. & Dijkerman, H. (2009). The weight of representing the body: Addressing the potentially indefinite number of body representations in healthy individuals. *Experimental Brain Research*, *Published on-line, 22 sept. 2009*.

Kaplan, D. (2008). An overview of markov chain methods for the study of stage-sequential developmental processes. *Developmental Psychology.*, *44*, 457-467.

Karasek, R. A. (1979). Job demands, job decision latitude and mental strain: Implications for job redesign. *Administrative Science Quarterly*, *24*, 285-308.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Kato, B. & Hoijtink, H. (2006). A bayesian approch to inequality constrained linear mixed models: Estimation and model selection. *Statistical Modelling*, *6*, 231-249.

Khalil, G., Saikali, R. & Berger, L. (2002). More powerful tests for the sign testing problem. *Journal of Statistical Planning and Inference*, *107*, 187 - 205.

Kieseppä, I. A. (2001). Statistical model selection criteria and the philosophical problem of underdetermination. *British Society for the Philosophy of Science*, *52*, 761 - 794.

Klimstra, T. A., Hale, W. W., Raaijmakers, Q. A. W., Branje, S. J. T. & Meeus, W. (2009). Maturation of personality in adolescence. *Journal of Personality and Social Psychology*, *96*, 898-912.

Klugkist, I. & Hoijtink, H. (2007). The bayes factor for inequality and about equality constrained models. *Computational Statistics and Data Analysis*, *51*, 6367-6379.

Klugkist, I., Laudy, O. & Hoijtink, H. (2005). Inequality constrained analysis of variance: A bayesian approach. *Psychological Methods*, *10*, 477 - 493.

Klugkist, I., Laudy, O. & Hoijtink, H. (2010). Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. *Psychological Methods*.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *94*, 1372-1381.

Kroger, J. (1988). A longitudinal study of ego identity status interview domains. *Journal of Adolescence*, *11*, 49-64.

Kroger, J. (1997). Gender and identity: The intersection of structure, content, and context. *Sex Roles*, *36*, 747-770.

Kroger, J. (2007). *The status of identity: Developmental perspectives.* (Tech. Rep.). Paper presented at the 14th Annual Conference of the Society for Research on Identity Formation, Washington, DC.

Kuiper, R. M. & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychological Methods*, *15*, 69-86.

Kuiper, R. M., Klugkist, I. & Hoijtink, H. (2010). A fortran 90 program for confirmatory analysis of variance. *Journal of Statistical Software*, *34*, 1-31.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79 - 86.

Laub, J. H., Nagin, D. S. & Sampson, R. J. (1998). Trajectories of change in criminal offending: good marriages and the desistance process. *American Sociological Review*, *63*, 225-238.

Laudy, O., Boom, J. & Hoijtink, H. (2005). Bayesian computational methods for inequality constrained latent class analysis. In A. V. der Ark & M. A. C. K. Sijtsma (Eds.), *New development in categorical data analysis for the social and behavioral sciences* (p. 63-82). Erlbaum: Londen.

Laudy, O. & Hoijtink, H. (2007). Bayesian methods for the analysis of inequality constrained contingency tables. *Statistical Methods in Medical Research*, *16*, 123-138.

Laudy, O., Zoccolillo, M., Baillargeon, R., Boom, J., Tremblay, R. & Hoijtink, H. (2005). Applications of confirmatory latent class analysis in developmental psychology. *European Journal of Developmental Psychology*, *2*, 1-15.

Lee, C. C. & Yan, X. (2002). Chi-squared tests for and against uniform stochastic ordering on multinomial parameters. *Journal of Statistical Planning and Inference*, *107*, 267 - 280.

Lee, M. D. & Pope, K. J. (2006). Model selection for the rate problem: A comparison of significance testing, bayesian, and minimum description length statistical inference. *Journal of Mathematical Psychology*, *50*, 193-202.

Lee, M. D. & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on trafimow (2003). *Psychological Review*, *112*, 662-668.

Leenders, I. & Brugman, D. (2005). Moral/non-moral domain shift in young adolescents in relation to delinquent behaviour. *British Journal of Developmental Psychology*, *23*, 65 - 79.

Leucari, V. & Consonni, G. (2003). Compatible priors for causal bayesian networks. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, a. A. S. D. Heckerman & M.West (Eds.), *Bayesian statistics* (p. 597-606). Oxford: Clarendon Press.

Lewis, H. (2003). Differences in ego identity among college students across age, ethnicity, and gender. *Identity*, *3*, 159-189.

Luyckx, K., Goossens, L., Soenens, B. & Vansteenkiste, M. (2005). Identity statuses based upon four rather than two identity dimensions: Extending and refining marcias paradigm. *Journal of Youth and Adolescence*, *34*, 605-618.

Luyckx, K., Goossens, L. & Soenens, B. (2006). A developmental contextual perspective on identity construction in emerging adulthood: Change dynamics in commitment formation and commitment evaluation. *Developmental Psychology*, *42*, 366-380.

Lynch, J. W., Kaplan, G. A. & Salonen, J. T. (1997). Why do poor people behave poorly? variation in adult health behaviours and psychosocial characteristics by stages of the socioeconomic life course. *Social Science & Medicine*, *44*, 809-819.

Lynch, S. (2007). *Introduction to applied bayesian statistics and estimation for social scientists.* New-York: Springer.

Marcia, J. E. (1966). Development and validation of ego-identity status. *Journal of Personality and Social Psychology*, *3*, 551-558.

Markstrom, C. & Marshall, S. (2007). The psychosocial inventory of ego-strengths: Examination of theory and psychometric properties. *Journal of Adolescence*, *30*, 63-79.

Markstrom, C., Sabino, V., Turner, B. & Berman, R. (1997). The psychosocial inventory of ego-strengths: Development and validation of a new eriksonian measure. *Journal of Youth and Adolescence*, *26*, 705-732.

Marsh, H. W. & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science*, *1*, 133-163.

Mason, D. (2003). Transcendental meditation in criminal rehabilitation and crime prevention. *Journal of Offender Rehabilitation*, *36*, 27-30.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147-163.

Meeus, W. (1996). Studies on identity development in adolescence: An overview of research and some new data. *Journal of Youth and Adolescence*, *25*, 569-598.

Meeus, W. & Deković, M. (1995). Identity development, parental and peer support in adolescence: Results of a national dutch survey. *Adolescence*, *30*, 931-945.

Meeus, W., Iedema, J., Helsen, M. & Vollebergh, W. (1999). Patterns of adolescent identity development: Review of literature and longitudinal analysis. *Developmental Review*, *19*, 419-461.

Meeus, W., Van de Schoot, R., Keijsers, L., Schwartz, S. J. & Branje, S. (2010). On the progression and stability of adolescent identity formation. a five-wave longitudinal study in early-to-middle and middle-to-late adolescence. *Child Development*.

Meeus, W., Van de Schoot, R., Klimstra, T. & Branje, S. (2010). Change and stability of personality types in adolescence: A five-wave longitudinal study in early-to-middle and middle-to-late adolescence. *Manuscript submitted*.

Modecki, K. L. (2009). its a rush: Psychosocial content of antisocial decision making. *Law and Human Behavior*, *33*, 183-193.

Mohren, D. C., Swaen, G. M., Borm, P. J., Bast, A. & Galama, J. M. (2001). Psychological job demands as a risk factor for common cold in a dutch working population. *Journal of Psychosomatic Research*, *50*, 21-27.

Mounts, N. S. & Steinberg, L. (1995). An ecological analysis of peer influence on adolescent grade point average and drug use. *Developmental Psychology*, *31*, 915 - 922.

Mulder, J., Hoijtink, H. & Klugkist, I. (2009). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*, 887-906.

Mulder, J., Klugkist, I., Van de Schoot, R., Meeus, W., Selfhout, M. & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *53*, 530-546.

Murphy, C. M., Stosny, S. & Morrel, T. M. (2005). Change in self-esteem and physical aggression during treatment for partner violent men. *Journal of Family Violence*, *20*, 201-210.

Muthén, L. K. & Muthén, B. O. (2007). *Mplus: Statistical analysis with latent variables: User's guide.* Los Angeles, CA: Muthén & Muth.

Nagin, D. S. (1999). Analyzing developmental trajectories: a semi-parametric group based approach. *Psychological Methods*, *4*, 139-157.

Nagin, D. S. (2005). *Group-based modeling of development.* Cambridge, MA: Harvard University Press.

Neumark-Sztainer, D., Story, M., French, S. A. & Resnick, M. D. (1997). Psychosocial correlates of health compromising behaviors among adolescents. *Health Education Research*, *12*, 37-52.

Nylund, K., Asparouhov, T. & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling*, *14*, 535-569.

Nylund, K., Muthén. B., A., and Nishina, Bellmore, A. & Graham., S. (2006). Stability and instability of peer victimization during middle school: Using latent transition analysis with covariates, distal outcomes, and modeling extensions. *Manuscript submitted for publication.*.

O'Hagan, A. (1995). Fractional bayes factors for model comparison. journal of the royal. *Journal of the Royal Statistical Society, Series B*, *57*, 99 - 138.

Orobio De Castro, B., Slot, N. W., Bosch, J. D., Koops, W. & Veerman, J. W. (2003). Negative feelings exacerbate hostile attributions of intent in highly aggressive boys. *Journal of Clinical Child and Adolescent Psychology*, *32*, 56-65.

Orobio de Castro, B., Veerman, J. W., Koops, W., Bosch, J. D. & Monshouwer, H. J. (2002). Hostile attribution of intent and aggressive behavior: A meta-analysis. *Child Development*, *73*, 916-934.

Perez, J. M. & Berger, J. (2002). Expected posterior prior distributions for model selection. *Biometrika*, *89*, 491 - 511.

Perlman, M. D. & Wu, L. (2002a). A class of conditional tests for a multivariate one-sided alternative. *Journal of Statistical Planning and Inference*, *107*, 155 - 171.

Perlman, M. D. & Wu, L. (2002b). A defense of the likelihood ratio test for one-sided and order-restricted alternatives. *Journal of Statistical Planning and Inference*, *107*, 173 - 186.

Phalet, K. & Schönpflung, U. (2001). Intergenerational transmission of collectivism and achievement values in to acculturation contexts. the case of turkish families in germany and turkish and moroccan families in the netherlands. *Journal of Cross-Cultural Psychology*, *32*, 186 - 201.

Piko, B. F., Fitzpatrick, K. M. & Wright, D. R. (2005). A risk and protective framework for understanding youths externalizing problem behavior in two different cultural settings. *European Child & Adolescent Psychiatry*, *14*, 95-103.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Popper, K. (1963). *Conjectures and refutations*. London: Routledge and Keagan Paul.

Press, S. J. (2005). *Applied multivariate analysis: Using bayesian and frequentist methods of inference (2nd ed)*. Malabar, FL: Krieger.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111-163.

Reinecke, J. (2006). Longitudinal analysis of adolescents deviant and delinquent behavior. *Methodology*, *2*, 100-112.

Reinke, W. & Ialongo, N. (2008). Empirically derived subtypes of child academic and behavior problems: Co-occurrence and distal outcomes. *Journal of Abnormal Child Psychology,*, *36*, 759-770.

Rigby, K. & Slee, P. (1993). Dimensions of interpersonal relation among australian children and implications for psychological well-being. *Journal of Social Psychology*, *133*, 33-42.

Ritov, Y. & Gilula, Z. (1993). Analysis of contingency tables by correspondence models subject to order constraints. *Journal of the American Statistical Association*, *88*, 1380 - 1387.

Roberts, B., Walton, K. & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life-course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1-25.

Robertson, T., Wright, F. T. & Dykstra, R. L. (1988). *Order restricted statistical inference*. New York : Wiley.

Romeijn, J.-W., Van de Schoot, R. & Hoijtink, H. (2010). One size does not fit all: Derivation of an adapted bic. *Manuscript in preparation*.

Rosenberg, M., Schooler, C. & Schoenbach, C. (1989). Self-esteem and adolescent problems: Modeling reciprocal effects. *American Sociological Review*, *54*, 1004-1018.

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

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

Roverato, A. & Consonni, G. (2004). Compatible prior distributions for dag models. *Journal of the Royal Statistical Society, Series B*, *66*, 47-62.

Rozenboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416-428.

Rusell, J. B. (1997). *Inventing the flat earth: Columbus and modern historians*. Burnham: Greenwood Press.

Salmivalli, C. (2001). Feeling good about oneself, being bad to others? remarks on self-esteem, hostility, and aggressive behavior. *Aggression and Violent Behavior*, *6*, 375-393.

Salmivalli, C., Kaukiainen, A., Kaistaniemi, L. & Lagerspetz, K. M. J. (1999). Self-evaluated self-esteem, peer-evaluated self-esteem, and defensive egotism as predictors of adolescents participation in bullying situations. *Personality and Social Psychology Bulletin,*, *25*, 1268-1278.

Sampson, A. R. & Singh, H. (2002). Min and max scorings for two sample partially ordered categorical data. *Journal of Statistical Planning and Inference*, *107*, 219 - 236.

Schnall, P., Landsbergis, P. & Baker, D. (1994). Job strain and caridovascular disease. *Annual Review of public health*, *15*, 381-411.

Schoenberg, S. (1997). Constrained maximum likelihood. *Computational Economics*, *10*, 251 - 266.

Scholte, R. H. J., Van Lieshout, C. F. M. & Van Aken, M. A. G. (2001). Relational support in adolescence: Factors, types, and adjustment. *Journal of Research in Adolescence*, *11*, 71-94.

Schultz, L., Bonawitz, E. & Griffiths, T. (2007). Can being scared cause tummy aches? naive theories, ambiguous evidence, and preschoolers causal inferences. *Developmental Psychology*, *43*, 1124-1139.

Schwartz, S. & Montgomery, M. (2002). Similarities or differences in identity development? the impact of acculturation and gender on identity

process and outcome. *Journal of Youth and Adolescence*, *31*, 359-372.

Schwarz. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Sen, P. K. & Silvapulle, M. J. (2002). An appraisal of some aspects of statistical inference under inequality constraints. *Journal of Statistical Planning and Inference*, *107*, 3 - 43.

Silvapulle, M. J. & Sen, P. K. (2004). *Constrained statistical inference: Order, inequality, and shape constraints*. London: John Wiley Sons.

Silvapulle, M. J., Silvapulle, P. & Basawa, I. V. (2002). Tests against inequality constraints in semiparametric models. *Journal of Statistical Planning and Inference*, *107*, 307 - 320.

Sober, E. (2002). Bayesianism, its scope and limits. In R. Swinburne (Ed.), *Bayes theorem* (p. 21-38). Oxford: Oxford University Press.

Sober, E. (2006). Parsimony. In J. Pfeifer & S. Sarkar (Eds.), *The philosophy of science: An encyclopedia* (Vol. 2, p. 530-541). New York: Routledge.

Spencer, S., Josephs, R. & Steele, C. (1993). Low self-esteem: The uphill struggle for self-integrity. In R. Baumeister (Ed.), *Self-esteem. the puzzle of low self-regard.* (p. 21-36). Baumeister, R.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society, series B*, *64*, 583-639.

Statistics Netherlands. (2008a). *Statline: Bevolking naar herkomst en generatie.* (Tech. Rep.). Voorburg, The Netherlands: Statististics Netherlands.

Statistics Netherlands. (2008b). *Statline: Mbo; leerlingen en geslaagden.* (Tech. Rep.). Voorburg, The Netherlands: Statististics Netherlands.

Statistics Netherlands. (2008c). *Statline: Vo; leerlingen en geslaagden.* (Tech. Rep.). Voorburg, The Netherlands: Statistics Netherlands.

Stephen, J., Fraser, E. & Marcia, J. E. (1992). Moratorium-achievement (mama) cycles in lifespan identity development: Value orientations and reasoning system correlates. *Journal of Adolescence*, *15*, 283-300.

Sterne, J. A. C. & Smith, G. D. (2001). Sifting the evidence - what's wrong with significane tests? *Physical Therapy*, *8*, 1464-1471.

Stigler, S. M. (1986). Laplace's 1774 memoir on inverse probability. *Statistical Science*, *1*, 359-363.

Stoel, R. D., Galindo-Garre, F., Dolan, C. & Van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, *4*, 439 - 455.

Strohmeier, D., Fandrem, H., Spiel, C. & Stefanek, E. (2009). The goal to be accepted by friends as underlying function of aggressive behavior in immigrant adolescents. *Manuscript submitted*.

Swann, W. B., Chang-Schneider, C. & McClarty, K. L. (2007). Do peoples self-views matter? self-concept and self-esteem in everyday life. *American Psychologist*, *62*, 84-94.

Taylor, S., Zvolensky, M., Cox, B., Deacon, B., Heimberg, R., Ledley, D. et al. (2007). Robust dimensions of anxiety sensitivity: Development and initial validation of the anxiety sensitivity index-3. *Psychological Assessment*, *19*, 176-188.

Thornberry, T. P. & Krohn, M. D. (2003). *Taking stock of delinquency: An overview of findings from contemporary longitudinal studies.* New York: Kluwer Academic/Plenum Publishers.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from bayes theorem. *Psychological Review*, *3*, 526-535.

Treffers, P. D. A., Goedhart, A. W., Veerman, J. W., Bergh, B. R. H. Van den, Ackaert, L. & Rycke, L. (2002). *Competentiebelevingsschaal voor adolescenten (cbsa), handleiding (manual of the self perception profile for adolescents - dutch version).* Lisse: Swets & Zeitlinger.

Trzesniewski, K. H., Donnellan, M. B., Moffitt, T. E., Robins, R. W., Poulton, R. & Caspi, A. (2006). Low self-esteem during adolescence predicts poor health, criminal behavior, and limited economic prospects

during adulthood. *Developmental Psychology*, *42*, 381-390.

Tsonaka, R. & Moustaki, I. (2007). Parameter constraints in generalized linear latent variable models. *Computational Statistics and Data Analysis*, *51*, 4164 - 4177.

Van Buuren, S. & Oudshoorn, C. G. M. (1999). *Flexible multivariate imputation by mice (no. pg/vgz/99.054).* (Tech. Rep.). Leiden: TNO Preventie en Gezondheid.

Van Buuren, S. & Oudshoorn, C. G. M. (2005). *Multivariate imputation by chained equations (no. pg/vgz/00.038).* (Tech. Rep.). Leiden: TNO Preventie en Gezondheid.

Van de Schoot, R., Hoijtink, H. & Deković, M. (2010). Testing inequality constrained hypotheses in sem models. *Structural Equation Modeling*, *17*, 443-463.

Van de Schoot, R., Hoijtink, H. & Doosje, S. (2009). Rechtstreeks verwachtingen evalueren of de nul hypothese toetsen? nul hypothese toetsing versus bayesiaanse model selectie. [Directly evaluating expectations or testing the null hypothesis: Null hypothesis testing versus bayesian model selection.]. *De Psycholoog*, *4*, 196-203.

Van de Schoot, R., Hoijtink, H., Mulder, J., Van Aken, M. A. G., Orobio de Castro, B., Meeus, W. et al. (2010). Evaluating expectations about negative emotional states of aggressive boys using bayesian model selection. *Developmental Psychology*.

Van de Schoot, R., Romeijn, J.-W. & Hoijtink, H. (2010). A prior predictive loss function for the evaluation of inequality constrained hypotheses. *Submitted for publication*.

Van de Schoot, R., Velden, F. van der, Boom, J. & Brugman, D. (2010). Can at risk young adolescents be popular and antisocial? Sociometric status groups, antisocial behavior, gender and ethnic background. *Journal of Adolescence*.

Van de Schoot, R. & Wong, T. (2010). Do antisocial young adults have a high or a low level of self-concept? *Self and Identity*.

Van Aken, M. A. G. & Dubas, J. D. (2004). Personality type, social relationships, and problem behaviour in adolescence. *European Journal of Developmental Psychology*, *1*, 331-348.

Van Aken, M. A. G., Van Lieshout, C. F. M., Scholte, R. H. J. & Haselager, G. J. T. (2002). Personality types in childhood and adolescence: Main effects and person relationship transactions. In L. Pulkkinen & A. Caspi (Eds.), *Pathways to successful development: Personality over the life course.* Cambridge: Cambridge University Press.

Van Hoof, A. (1999). The identity status field re-reviewed: An update of unresolved and neglected issues with a view on some alternative approaches. *Developmental Review*, *19*, 497-556.

Van Well, S., Kolk, A. M. & Klugkist, I. (2009). The relationship between sex, gender role identification, and the gender relevance of a stressor on physiological and subjective stress responses: Sex and gender (mis)match effects. *International Journal of Psychophysiology*, *32*, 427-449.

Vermeiren, R., Bogaerts, F., Ruchkin, V., Deboutte, D. & Schwab-Stone, M. (2004). Subtypes of self-concept and self-concept in adolescent violent and property offenders. *Journal of Child Psychology and Psychiatry*, *45*, 405-411.

Visser-Van Balen, H., Laak, J. J. F. ter, Treffers, P. D. A., Sinnema, G. & Geenen, R. (2007). Construction and validation of the self-perception profile for young adults dutch version. In H. V.-V. Balen (Ed.), *Growing up with short stature. psychosocial consequences of hormone treatment.* (p. 83-104). Utrecht: Utrecht University, PhD thesis.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*, 212-213.

Walker, L. J., Gustafson, P. & Frimer, J. A. (2007). The application of bayesian analysis to issues in developmental research. *International Journal of Behavioral Development*, *4*, 366-373.

Walker, L. J., Gustafson, P. & Hennig, K. (2001). The consolidation/transition model in moral reasoning development. *Developmental Psychology*, *37*, 187-197.

Waterman, A., Geary, P. & Waterman, C. (1974). Longitudinal study of changes in ego identity status from the freshman to the senior year at college. *Developmental Psychology*, *10*, 387-392.

Waterman, A. & Goldman, J. (1976). A longitudinal study of ego identity development at a liberal arts college. *Journal of Youth and Adolescence*, *5*, 361-369.

Waterman, A. & Waterman, C. K. (1971). A longitudinal study of changes in ego identity status during the freshman year at college. *Developmental Psychology*, *5*, 167-173.

Waterman, A. S. (1982). Identity development from adolescence to adulthood: An extension of theory and a review of research. *Developmental Psychology*, *18*, 358.

Webster, G. D., Kirkpatrick, L. A., Nezlek, J. B., Smith, C. V. & Paddock, E. L. (2007). Different slopes for different folks: Self-esteem instability and gender as moderators of the relationship between self-esteem and attitudinal aggression. *Self and Identity*, *6*, 74-96.

Wong, T. & Van de Schoot, R. (2010). Reporting violent victimization to the police: The role of the sex of the offender. *Submitted for publication*.

Acknowledgement (Dankwoord)

Allereerst wil ik mijn promotor, Herbert Hoijtink, en mijn co-promotor, Jan-Willem Romeijn, enorm bedanken:

Herbert, je hebt me ontzettend veel geleerd, waaronder: mijn formulevrees overwinnen; me enthousiasmeren voor (Bayesiaanse) statistiek; ook al zul je het niet geloven, je hebt me zelfs 'nee' leren zeggen tegen tal van interessante extra projecten en andere afleidingen; ten slotte is nog het vermelden waard dat je mijn relatie hebt weten te redden door me te verbieden mijn promotie en bruiloft op dezelfde dag te vieren (grapje. . .).

Jan-Willem, door jou ben ik gaan 'nadenken' en ik kijk met veel plezier terug op de vele filosofische discussies die we de laatste jaren gevoerd hebben (waaronder telefonisch meer dan een uur terwijl jij in zuid- afrika was). Je hebt me geleerd elke zin, zelfs elk woord, op een gouden weegschaal te leggen en als ik dan dacht dat een paper indienbaar was, werd vervolgens de hele procedure nog eens herhaald. . . . en herhaald. . . . en ng eens herhaald. Ik kom graag nog vaak in Groningen bij je over de vloer om onze samenwerking voort te zetten.

Niet officieel verbonden aan mijn PhD-traject, maar toch van onschatbare waarde, wil ik Irene Klugkist bedanken. Irene, je hebt me verleid om berhaupt met dit project te starten. Tijdens de afgelopen jaren heb je me

altijd voorzien van adviezen, zowel inhoudelijk als persoonlijk over hoe in de academische wereld te werken, dank daarvoor!

Ook dank voor mijn kamergenootjes, Rebecca Kuiper en Floryt van Wesel, met wie ik de afgelopen 3,5 jaar intensief heb 'samengeleefd'. Het was vast niet altijd even gemakkelijk met mij op de kamer, maar ik heb het enorm naar mijn zin gehad en veel steun en hulp van jullie mogen ervaren. Binnen het VICI-project waren we met vijf promovendi werkzaam, waarvan Joris Mulder en Carel Peeters nog niet vermeld. Waarschijnlijk was ik de minst statistisch onderlegde van het stel. Mijn dank is dus groot voor het beantwoorden van alle 'stomme' vragen die ik de afgelopen jaren heb gesteld aan jullie. Joris en Rebecca, dank voor het mogen gebruiken van de door jullie ontwikkelde software en het vele geduld dat jullie hadden om mij uit te leggen hoe het programma(tje) werkte.

Voor de verschillende dissertatie hoofdstukken heb ik samengewerkt met collega onderzoekers en ik wil hen dan ook bedanken voor de goede samenwerking: Wim Meeus, Marcel van Aken, Bram Orobio de Castro, Maja Dekovic, Daan Brugman, Judith Dubas, Thessa Wong, Joris Mulder en Sibe Doosje.

Het schrijven van een dissertatie is een hele klus en ik ben mijn dank verschuldigd aan Annette Mills voor de vele (Engelse) taalcorrecties, Maaike van Rossum voor het vertalen van word bestanden naar Tex bestanden, Nelleke de Weerd voor de begeleiding bij het drukken en ten slotte Wenneke van de Schoot-Hubeek voor de vele uren die jij hebt besteed aan mijn dissertatie.

Wenneke, voor jou een aparte paragraaf om je te bedanken. Tijdens onze bachelor psychologie hebben we elkaar leren kennen, tijdens onze master zijn we gaan samenwonen, en tijdens mijn promotie zijn we verloofd en getrouwd. Dat laatste in dezelfde maand dat mijn dissertatie af moest zijn en dat is een zware periode geweest. Toch heb jij me te allen tijde gesteund en geholpen; je hebt elk hoofdstuk gelezen, geprezen en gecorrigeerd (zelfs een formule fout gevonden in het meest technische hoofdstuk). Mijn dank is zeer groot

en sorry voor de avonden dat ik wazig voor me uit aan het staren was en niet luisterde naar al je verhalen...

Ten slotte wil ik graag mijn moeder, mijn tweede vader Louis, en mijn zusje Elske, bedanken voor alle steun en getoonde interesse de afgelopen jaren. Ik weet dat jullie een goede poging hebben ondernomen om alles te begrijpen waar ik mee bezig ben geweest en dat waardeer ik enorm. Als jullie dit lezen ben ik inmiddels begonnen als universitair docent waarbij ik de onderwijs-interesse van mijn moeder en de statistische-interesse van mijn overleden vader met elkaar combineer. Ik heb er zin in!

Short C.V.

Rens van de Schoot was born on July 9, 1979, in Eindhoven, The Netherlands. He completed pre-university education (HAVO) in Eindhoven at the Eckart College and studied Medical imaging techniques at the Fontys Hoge School. He worked for two years full time on the x-ray department of the university hospital in Utrecht (UMC). After this, he completed his Psychology bachelor with a minor in juvenile delinquency and he graduated cum laude for the research master Development and Socialization of Children and Adolescents at the graduate school for social sciences at Utrecht university. After that, he started working as PhD-student at the department of Methods and Statistics. During his PhD, he was chair for the university board of PhD-students and vice-chair for the scientific committee of the Dutch Institute for Psychologists (NIP). Besides working at his dissertation, he did many consultations within and outside academia. Moreover, he worked on a project together with the Netherlands Centre for Graduate and Research Schools entitled "Ph.D. trajectories and labor market mobility: A survey of doctoral graduates in the Netherlands". He finished his dissertation after only working three years on his main project and he was able to publish several papers during his PhD, listed below. Currently he is assistant professor at the department Methods and Statistics, Utrecht university.

List of Publications

Journal articles:

1. **Van de Schoot, R.** & Wong, T. (in press). Do Antisocial Young Adults Have a High or a Low Level of Self-concept? *Self & Identity.*

2. **Van de Schoot, R.**, Hoijtink, H., Mulder, J., Van Aken, M., Orobio de Castro, B., Meeus, W. & Romeijn, J.-W. (in press). Evaluating Expectations about Negative Emotional States of Aggressive Boys using Bayesian Model Selection. *Developmental Psychology.*

3. **Van de Schoot, R.**, van der Velden, F., Boom, J. & Brugman, D. (in press). Can at Risk Young Adolescents be Popular and Antisocial? Sociometric Status Groups, AntiSocial behaviour, Gender and Ethnic Background. *Journal of Adolescence.*

4. **Van de Schoot, R.**, Hoijtink, H. & Deković, M. (2010). Testing Inequality Constrained Hypotheses in SEM Models. *Structural Equation Modeling, 17*, 443-463.

5. **Van de Schoot, R.**, Hoijtink, H. & Doosje, S. (2009). Rechtstreeks Verwachtingen Evalueren of de Nul Hypothese Toetsen? Nul Hypothese Toetsing versus Bayesiaanse Model Selectie [Directly Evaluating

Expectations or Testing the Null Hypothesis: Null Hypothesis Testing versus Bayesian Model Selection.]. *De Psycholoog 4,* 196-203.

6. Boelen, P. A., **Van de Schoot, R.**, Hout, M. van den, Keijser, J. de, & Bout, J. van den (2010). Prolonged Grief Disorder, Depression, and Posttraumatic Stress-Disorder are Distinguishable Syndromes. *Journal of Affective Disorders, 125*, 374-378.

7. Meeus, W., **Van de Schoot, R.**, Keijsers, L., Schwartz, S. J. & Branje, S. (in press). On the Progression and Stability of Adolescent Identity Formation. A Five-Wave Longitudinal Study in Early-to-middle and Middle-to-late Adolescence. *Child Development.*

8. Doosje, S., De Goede, M. P.M., Van Doornen, L. J.P. & **Van de Schoot, R.** (in press). Humorous coping styles, job characteristics and job-related affect as predictors of the incidence of upper respiratory infection. *South African Journal of Industrial Psychology.*

9. Mulder, J., Klugkist, I., **Van de Schoot, R.**, Meeus, W., Selfhout, M. & Hoijtink, H. (2009). Bayesian Model Selection of Informative Hypotheses for Repeated Measurements. *Journal of Mathematical Psychology, 53,* 530-546.

10. Romeijn, J.-W. & **Van de Schoot, R.** (2008). A Philosopher's View on Bayesian Evaluation of Informative Hypotheses. In H. Hoijtink, I. Klugkist, & P. Boelen *(ed.). Bayesian Evaluation of Informative Hypotheses, New-York: Springer, p.* 329-358.

Under Review:

11. **Van de Schoot, R.**, Mulder, J., Hoijtink, H., Van Aken, M.A.G. , Dubas, J. S., Orobio de Castro, B., Meeus, W. & Romeijn, J.-W. (2009). Psychological Functioning, Personality and Support from

family: An Introduction to Bayesian Model Selection. *Manuscript submitted for publication.*

12. **Van de Schoot, R.**, Hoijtink, H., Brugman, D. & Romeijn, J.-W. (2010). A Prior Predictive Loss Function for the Evaluation of Inequality Constrained Hypotheses. *Manuscript submitted for publication.*

13. **Van de Schoot, R.**, Romeijn, J.-W. & Hoijtink, H. (2010). Background Knowledge in Model Selection Procedures. *Manuscript submitted for publication.*

14. Wong, T. & **Van de Schoot, R.** (2009). Reporting Violent Victimization to the Police: The Role of The Sex of the Offender. *Manuscript submitted for publication.*

15. Van Rossum, M., **Van de Schoot**, **R.** & Hoijtink, H. (2010). Inequality Constrained Hypotheses for ANOVA. *Manuscript submitted for publication.*

16. De Graaf, H., **Van de Schoot**, **R.**, Woertman, L. & Meeus, W. (2010). Parental Support and Romantic and Sexual Development: A Three Wave Longitudinal Study. *Manuscript submitted for publication.*

17. Meeus, W., **Van de Schoot, R.**. Klimstra, T. & Branje, S. (2010). Change and Stability of Personality Types in Adolescence: A Five-Wave Longitudinal Study in Early-To-Middle and Middle-To-Late Adolescence. *Manuscript submitted for publication.*

18. Lüftenegger, M., Schober, B., **Van de Schoot, R.**, Wagner, P., Finsterwald, M. & Spiel, C. (2010). Lifelong Learning as a Goal - Do Autonomy and Self-Regulation in School Result in Well Prepared Pupils? *Manuscript submitted for publication.*

19. Van Loey, N.E., Van Beeck, E.F. , Faber, A.W., **Van de Schoot, R.** & Bremer, M. (2010). Objective and Subjective Evaluation, Two Different Perspectives on Health Related Quality of Life: A Multicentre Prospective Cohort Study. *Manuscript submitted for publication.*

Other publications:

20. Sonneveld, H., Lockhorst, D., & **Van de Schoot,R.** (2009). Het rendement van MaGW/NWO promotierendementen. Rapport voor NWO. Nederlands Centrum voor de Promotieopleiding IVLOS, Universiteit Utrecht

21. Sonneveld, H., Yerkes, M., & **Van de Schoot, R.** (in press). Ph.D. Trajectories and Labor Market Mobility: A Survey of Doctoral Graduates in the Netherlands. Report for Netherlands Centre for Graduate and Research Schools in the Netherlands (Subsidized by the Dutch Ministry of Education, Culture and Science).

22. **Van de Schoot, R.**, Yerkes, M., & Sonneveld, H. (2009). Codebook of Netherlands Survey of Doctorate Recipients (subsidized by the Dutch Ministry of Education, Culture and Science).

23. **Van de Schoot, R.**, & Hoijtink, H. (2008). Tutorial for Inequality Constrained Latent Class Modeling (LCM version 1.0). Technical paper; download at www.fss.uu.nl/ms/informativehypotheses.