Running Title: Moving beyond null hypothesis testing.

Moving beyond traditional null hypothesis testing; Evaluating expectations directly

Van de Schoot, R.,[*1] Hoijtink, H.,[1] Jan-Willem Romeijn[2]

[1] Department of Methods and Statistics, Utrecht University, The Netherlands
[2] Department of Philosophy, Groningen University, The Netherlands
[*] Correspondence:
Dr. Rens van de Schoot
Utrecht University
Department of Methodology and Statistics
P.O. Box 80.140
3508TC, Utrecht, The Netherlands
a.g.j.vandeschoot@uu.nl

Acknowledgement: Supported by a grant from the Netherlands organization for scientific research: NWO-VICI-453-05-002.

Moving beyond traditional null hypothesis testing; evaluating expectations directly

Abstract

In this mini-review, it will be illustrated that testing the traditional null hypothesis is not always an appropriate strategy. Halve in jest, we discuss Aristotle's scientific investigations about the shape of the earth in the context of evaluating the traditional null hypothesis. We conclude that Aristotle was actually interested in evaluating *informative* hypothesis. In contemporary science the situation is not much different. That is, many researchers have no particular interest in the traditional null hypothesis. More can be learned from data by evaluating specific expectations, or so-called informative hypotheses, than by testing the traditional null hypothesis. These *informative* hypotheses will be introduced and an overview of literature is provided on evaluating informative hypothesis.

Keywords: null hypothesis testing, Bayesian analysis, informative hypothesis, inequality constraints

1. Introduction

In the current mini-review, it is argued that testing the traditional null hypothesis is not always an appropriate strategy. That is, many researchers have no particular interest in the hypothesis `nothing is going on' (Cohen, 1990). So why test such a hypothesis if one is not interested in it? The APA stresses in its publication manual that null hypothesis testing should just be a starting point for statistical analyses: "Reporting elements such as effect sizes and confidence intervals are needed to convey the most complete meaning of the results" (APA, 2010, p.33; see also Fidler, 2002). In the current paper we go one step further then reporting effect sizes and confidence intervals and argue that specific expectations should be evaluated directly. As Osborne (2010) stated "The world doesn't need another journal promulgating 20th century thinking, genuflecting at the altar of $p <$ 0.05. I challenge us to challenge tradition" (p.3) and that is exactly what we will do in the current paper. Statistical tools for the evaluation of informative hypotheses are becoming available and are more often used in applications. We provide an overview of the current state of affairs for the evaluation of informative hypotheses. But first we argue, halve in jest, what is 'wrong' with the traditional null hypothesis, and we introduce the *informative* hypothesis.

One important note has to be made, namely researchers like Wagenmakers, Lee, Lodewyckx and Iverson (2008) criticize T-tests rendering no legitimate results and that p-values are prone to misinterpretation. Or, researchers like Coulson, Healey, Fidler, and Cumming (2010) and Fidler (2001), who also explicitly argue against solely reporting p-values and argue for using confidence intervals. Or, researchers like Rosenthal, Rosnow, and Rubin (2000) argue for using focused contrasts which could be used to evaluate expectations directly. However, in the current paper we will focus on developments in statistics that move beyond using confidence intervals, effect sizes and planned contrasts.

2. What is 'wrong' with the traditional null hypothesis?

Cohen (1994) aptly summarized the criticism of traditional null hypothesis testing in the title of his paper "The earth is round ($p < .05$)". Let us elaborate on this for education and entertainment using an example inspired by this title.

The question of the shape of the earth was a recurring issue in scientific debate during the era of Aristotle (384BC-322BC; see Rusell, 1997). By that time, the Greek idea that the earth was round had dominated scientific thinking. 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 and let us tell the story of Aristotle's scientific investigations using different ways of evaluating hypotheses.[1]

[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. 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.

In order to falsify the old Mesopotamian hypothesis, we say that Aristotle might have used an approach based on testing the traditional null hypothesis:

H_0: The shape of the earth is a flat disk,

H_1: The shape of the earth is not a flat disk.

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.

Aristotle would have gathered data about the shape of the earth and found evidence against the null hypothesis, for example: 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. Such observations could not be taken as evidence of a flat earth, and H_0 would have been rejected, leading Aristotle to conclude that the earth cannot be represented by a flat disk.

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:

$H_{0'}$: The shape of the earth is a sphere,

$H_{1'}$: The shape of the earth is not a sphere.

Now, imagine that Aristotle continued his search for data and that he gathered data that yielded evidence against (!) the null hypothesis[2]: while standing on a mountain top, he noticed the Earth's surface has many irregularities and if enough irregularities are observed it could provide just enough evidence to reject the null hypothesis. 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? Not much! Both hypothesis tests reject the traditional null hypotheses H_0 and $H_{0'}$. Following the Neyman –Pearson procedure of hypothesis testing, we can tentatively adopt the alternative hypotheses H_1 and $H_{1'}$. This procedure tells us that the earth is neither a flat disk, nor a sphere and we remain ignorant of the earth's actual shape. This ignorance is a result of the 'catch-all' alternative hypothesis as proposed by Neyman and Pearson (Neyman, & Pearson, 1967). Unfortunately, the catch-all includes all shapes that are non-flat and non-spherical, for example pear-shaped.[3]

[2] 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.

[3] Admittedly, not all methodologists agree on this point. In response to Aristotle's imagined disappointment, Popper would have argued that this insight is all that Aristotelian science, or any science

Rather than using the hypothesis tests given above, we might argue that Aristotle was actually interested in evaluating:

H_A : The shape of the earth is a flat disk,

versus

H_B : The shape of the earth is a sphere.

In such a direct comparison the conclusion will be more informative.

3. What Does This Historical Example Teach Us?

Evaluating specific expectations directly produces more useful results than sequentially testing traditional null hypotheses against catch-all rivals. We argue that researchers are often interested in the evaluation of informative hypotheses and already know that the traditional null hypothesis is an unrealistic hypothesis. This presupposes that prior knowledge is available and if that were not the case, testing the traditional null hypothesis would be appropriate. In most applied articles, however, prior knowledge is available in the form of specific expectations about the ordering of statistical parameters.

Let us illustrate this using an example of Van de Schoot, Van der Velde, 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 (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. Each sociometric status group has been characterised by distinct behavioural patterns which influence the quality of social relations. For example, peer rejection was found to be related to antisocial behaviour, whereas popular adolescents tended to be considered as well-known, attractive, athletic, and socially competent, but can also be anti-social, as was shown by Van de Schoot, Van der Velde et al. (2010).

Suppose we want to compare these five soiometric status groups on the number of committed offences reported to the police last year (minor theft, violence, and so on) and let the groups be denoted by μ1 for the mean on the number of committed offences for the popular group, μ2 for the rejected group, μ3 for the neglected group, μ4 for the controversial group and μ5 for the average group. Different types of hypotheses can be formulated that are used in the procedures described in the remainder of this paper.

First, informative hypotheses can be formulated denoted by H_{I_1}, H_{I_2}, ..., H_{I_N} for

for that matter, can hope for. When it comes to general hypotheses, or hypotheses that are beyond the reach of direct verification, 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!

a set of N hypotheses. These hypotheses contain information about the ordering of the parameters in a model, in our example the five means. Such expectations about the ordering of parameters can stem from previous studies, a literature review or even academic debate. Consider an imaginary hypothesis with inequalities between the five mean scores, $H_{I_1}: \mu_3 < \mu_1 < \mu_5 < \mu_2 < \mu_4$, where the neglected group is expected to commit fewer offences compared to the popular group who in turn are expected to commit fewer offences compared to the average group, and so on. If no information is available about the ordering, this is denoted by a comma. Another expectation could be the hypothesis $H_{I_2}: \mu_3 < \{\mu_1, \mu_5, \mu_2\} < \mu_4$, where the neglected group is expected to commit fewer offences compared to the popular, average and rejected groups. There is no expected ordering between these three groups, but all three are expected to commit fewer offences then the controversial group. The research question would be, which of the two informative hypothesis receives most support from the data.

Second, there is the traditional null hypothesis (denoted by H_0), which states that nothing is going on and all groups have the same score, $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. Third, if no constraints are imposed on any of the means, and any ordering is equally likely, the hypothesis is called a ‘catch all’ alternative hypothesis, or an unconstrained hypothesis (denoted by H_U): $H_U: \mu_1, \mu_2, \mu_3, \mu_4, \mu_5$. In the next section we present an overview of possible alternatives for traditional null hypothesis testing to evaluate one or more informative hypotheses.

4. Evaluating informative hypotheses

In the literature different procedures are described that allow for the evaluation of informative hypotheses. We present an overview of technical papers, software and applications for two types of approaches: (1) hypothesis testing approaches and (2) model selection approaches. Note that we limit ourselves to a discussion of papers where software is available for applied researchers.

4.1 Hypothesis testing approach

There are approaches reported in literature that renders a p-value for the comparison of H_I with H_0 or with H_U. First, an adaptation of the traditional F-test for analysis of variance (ANOVA) has been proposed by Silvapulle, Silvapulle, and Basawa (2002, see also Silvapulle & Sen, 2004), called the F-bar test. It is a confirmatory method to test one single informative hypothesis in two steps, for example:

$$H_0: \mu_3 = \mu_1 = \mu_5 = \mu_2 = \mu_4$$

versus

$$H_{I_1}: \mu_3 < \mu_1 < \mu_5 < \mu_2 < \mu_4,$$

and

$$H_{I_1}: \mu_3 < \mu_1 < \mu_5 < \mu_2 < \mu_4$$

versus

$$H_U: \mu_3, \mu_1, \mu_5, \mu_2, \mu_4,$$

where in the second hypothesis test H_{I_1} serves as the role as the null hypothesis. Software for the F -bar test is described in Kuiper, Klugkist, and Hoijtink, (2010), but applications are to our knowledge not yet reported in literature. Application of the F -bar test is easy using the software,[4] and the results are comparable with a classical F-test. The disadvantage is that only one single informative hypothesis at a time can be evaluated and only for univariate analysis of variances.

Testing informative hypotheses for structural equation models (SEM) has been described in Stoel, Galindo-Garre, Dolan, and Van den Wittenboer (2006), where constraints were imposed on variance terms to obtain only positive values (see also, Gonzalez & Griffin, 2001). A likelihood ratio test is used and the software is available in the statistical package R (R Development Core Team, 2005).[5]

The procedure described in Van de Schoot, Hoijtink and Deković (2010) also makes use of a likelihood ratio test, but goes one step further than Stoel et al. (2006). A parametric bootstrap procedure in combination with inequality constraints imposed on regression coefficients. The methodology consists of several steps to be performed with the aid of commonly used software Mplus (Muthen & Muthen, 2007). [6] Van de Schoot and Strohmeier (in press) introduced the methodology to non-statisticians and showed that using this method results in a power gain. That is, fewer participants are needed to obtain a significant effect compared to a default chi-square test.

4.2 Model Selection approach

A second way of evaluating an informative hypothesis is to use a model selection approach. This is not a test of the model in the sense of hypothesis testing, rather it is an evaluation between statistical models using a trade-off between model fit and model complexity. 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.

There is a variety of model selection procedures commonly used in practical applications, most notably Akaike's Information Criterium (AIC) (Akaike, 1973), the Bayesian Information Criterium (BIC) (Schwarz, 1978) and the Deviance Information Criterium (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). Problems with these standard model selection tools in the context of evaluating informative hypotheses arise because the tools are not equipped to deal with inequality constraints (Mulder, Hoitjink & Klugkist, 2009; Van de Schoot, Romeijn, & Hoijtink, under review). Although the model selection tools differ in their expression, the result always consist of two parts: the likelihood of the best fitting hypothesis within the model is a measure of model fit, and an expression containing the number of (effective) parameters of the model as 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. The problem is that the expression of complexity is based on the number of parameters in the model and can

[4] The software can be downloaded at http://vkc.library.uu.nl/vkc/ms/research/ProjectsWiki/Informative%20hypotheses.aspx

[5] The corresponding scripts can be downloaded from the Web site of Psychological Methods.

[6] The software can be downloaded at staff.fss.uu.nl/agjvandeschoot

not take inequality constraints into account. That is, $H_{I_1} : \mu_3 < \mu_1 < \mu_5 < \mu_2 < \mu_4$ and $H_{I_2} : \mu_3 < \{\mu_1, \mu_5, \mu_2\} < \mu_4$ would receive the same measure for complexity which is unwanted because H_{I_1} is more parsimonious than H_{I_2} due to more restriction imposed on the five means.

Alternative model selection tools have been proposed in the literature. First, an alternative model selection procedure is the Paired-Comparison Information Criterion (PCIC) proposed by Dayton (1998, 2003), with an application in Taylor et al. (2007). The PCIC is an exploratory approach which computes a default model selection tool for all logically possible subsets of group orderings. For the PCIC only the source code for the programming language GAUSS was available (Dayton, 2001), but Kuiper and Hoijtink, (2010) made the PCIC available in a user friendly interface.[5] The disadvantage of the PCIC is that it is an exploratory approach.

Second, the literature also contains one modification of the AIC 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 & Silvapulle, in press) with an application in Hothorn, Vaeth, and Hothorn, (2009). 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. Software for ORIC is described in Kuiper, Klugkist, and Hoijtink, (2010). The ORIC is as to yet only available for analysis of variance models, but a generalization is under construction.

Alternatives for the BIC and the DIC are under construction, see Romeijn, Van de Schoot, and Hoijtink (under review) and Van de Schoot, Hoijtink, Brugman, and Romeijn (under review), respectively.

Finally, one other method of model selection, which receives more and more attention in literature, involves the evaluation of informative hypothesis using Bayes factors. In this method each (informative) hypothesis of interest is provided with a `degree of support' which tells us exactly how much support there is for each of the 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.

The methodology of evaluating a set of inequality constrained hypotheses has proved to be a flexible tool that can deal with many types of constraints. We refer to the book of Hoijtink, Klugkist and Boelen (2008), and the papers of Van de Schoot, Mulder et al. (in press) and Van de Schoot, Hoijtink et al. (2011) as a first step for interested readers. For a philosophical background see Romeijn and Van de Schoot (2008) and for more information on hypotheses elicitation see Van Wesel, Boeije and Hoitjink (under review). Varies papers describe comparisons between traditional null hypothesis testing and Bayesian evaluation of informative hypotheses, see Kuiper and Hoijtink (2010), Hoijtink, Huntjes et al. (2008), Hoijtink, & Klugkist (2007), and Van de Schoot, Hoijtink et al. (2011).

Software is available for:[5]

- AN(C)OVA models (Klugkist, Laudy & Hoijtink, 2005; Kuiper & Hoijtink, 2010; Van Wesel, Hoijtink & Klugkist, 2010) with an application in Van Well, Kolk, & Klugkist (2009);
- Multivariate linear models including time-varying and time-invariant covariates (Mulder, Hoijtink, & Klugkist, 2009; Mulder, Klugkist et al., 2009) with an application in Kammers, Mulder, De Vignemont, and Dijkerman (2009);
- Latent class analyses (Laudy, Boom, & Hoijtink, 2005; Hoijtink, 2001) with applications in Laudy et al. (2005) and Van de Schoot and Wong (in press);
- Order restricted contingency tables (Laudy & Hoijtink, 2007; see also Klugkist, Laudy & Hoijtink, in press) with applications in Meeus, Van de Schoot, Keijsers, Schwartz, & Branje (2010) and Meeus, van de Schoot, Klimstra, and Branje, (in press).

5. Conclusion

Statistics have come a long way since the early beginnings of testing the traditional null hypothesis: ‘nothing is going on’. Developments in statistics allow researchers to directly evaluate their expectations in the form of informative hypotheses specified with inequality constraints. The current mini-review provides the current state of affairs.

Literature

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

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.

Anraku, K. (1999). An information criterion for parameters under a simple order restriction. *Journal of the Royal Statistical Society, series B, 86*, 141-152.

Cohen, J. (1994). The Earth is round (p < .05). *American Psychologist, 49*, 997-1003.

Coulson, M., Healey, M., Fidler,F. & Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Quantitative Psychology and Measurement,, 1,* 1-9.

Dayton, C. M. (1998). Information criteria for paired-comparison problem. *American Statistician, 52,* 144–151.

Dayton, C. M. (2001). SUBSET: Best subsets using information criteria. *Journal of Statistical Software, 6, 1*–10.

Dayton, C. M. (2003). Information criteria for pairwise comparisons. *Psychological Methods, 8*, 61-71

Gonzalez, R. & Griffin, D. (2001). Testing parameters in structural equation modelling: Every "One" matters. *Psychological Methods, 6*, 258 - 269.

Fidler, F. (2001). Computing Correct Confidence Intervals for Anova Fixed-and Random-Effects Effect Sizes. *Educational and Psychological Measurement, 61*, 575-604

Fidler, F. (2002). The Fifth edition of the Apa Publication Manual: Why its Statistics Recommendations are so Controversial. *Educational and Psychological Measurement*, *62,* 749-770

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., 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.

Hothorn, L., Vaeth, M. & Hothorn, T. (2009). Trend tests for the evaluation of exposure-response relationships in epidemiological exposure studies. *Epidemiologic Perspectives & Innovations, 6.*

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 .

Kuiper, R. M. & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychological Methods, 15,* 69-86.

Kuiper, R.M., Hoijtink, H. & Silvapulle, M.J. (in press). An Akaike-type information criterion for model selection under inequality constraints. *Biometrika.*

Kuiper, R. M., Klugkist, I. & Hoijtink, H. (2010). A fortran 90 program for confirmatory analysis of variance. *Journal of Statistical Software, 34*, 1-31.

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. (in press). Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. *Psychological Methods*.

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 con_rmatory latent class analysis in developmental psychology. *European Journal of Developmental Psychology, 2*, 1-15.

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, 81, 1565–1581.*

Meeus, W., van de Schoot, R., Klimstra, T. and Branje, S. (in press). Change and Stability of Personality Types in Adolescence: A Five-Wave Longitudinal Study in Early-to-middle and Middle-to-late Adolescence. *Developmental Psychology*

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.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus: Statistical analysis with latent variables: User's guide.* Los Angeles, CA: Muthén & Muthén.

Neyman, J., & Pearson, E. (1967). Joint Statistical Papers. Cambridge: Cambridge University Press.

Osborne J. W. (2010). Challenges for quantitative psychology and measurement in the 21st century. *Frontiers in Pscychology, 1,* 1-3.

R Development Core Team. (2005). R: A language and environment for statistical computing [Computer software]. Vienna: R Foundation for Statistical Computing.

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.

Romeijn, J-W, Van de Schoot, R, & Hoijtink, H. (under review). One Size Does Not Fit All: derivation of an adapted BIC. *Manuscript submitted for publication.*

Rusell, J. B. (1997). *Inventing the flat Earth: Columbus and modern historians*. Burnham: Greenwood Press.

Schwarz. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

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.

Silvapulle, M. J., & Sen, P. K. (2004). *Constrained statistical inference: Order, inequality, and shape constraints*. London: John Wiley Sons.

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.

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.

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.

Van de Schoot, R., Hoijtink, H., Brugman, D., & Romeijn, J-W (under review). A Prior Predictive Loss Function for the Evaluation of Inequality Constrained Hypotheses. *Manuscript submitted for publication.*

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., Mulder, J., Van Aken, M. A. G., Orobio de Castro, B., Meeus, W. & Romeijn, J.-W. (2011). Evaluating Expectations about Negative Emotional States of Aggressive Boys using Bayesian Model Selection. *Developmental Psychology, 47, 203-212*

Van de Schoot, R., Mulder, J., Hoijtink, H., van Aken, M.A.G. , Dubas, J.S., de Castro, B. O., Meeus, W. & Romeijn, J.-W. (in press). Psychological Functioning, Personality and Support from family: An Introduction Bayesian Model Selection. *European Journal of Developmental Psychology*

Van de Schoot, R., Romeijn, J-W, & Hoijtink, H. (under review). Background Knowledge in Model Selection Procedures. *Manuscript submitted for publication.*

Van de Schoot, R. & Strohmeier, D. (in press). Testing informative hypotheses in SEM Increases Power: An illustration contrasting classical hypothesis testing with a parametric bootstrap approach. *International Journal of Behavioural Development*

Van de Schoot, R. & Wong, T. (in press). Do Antisocial Young Adults Have a High or a Low Level of Self-concept? *Self & Identity*

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.

Van Wesel, F., Boeije, H. and Hoijtink, H. (under review). Elicitation and use of hypotheses for analysis of variance models: Challenging the current practice. *Manuscript submitted for publication.*

Van Wesel, F., Hoijtink, H., & Klugkist, I. (2010). Choosing priors for constrained analysis of variance : methods based on training data. *Scandinavian Journal of*

Statistics. DOI: 10.1111/j.1467-9469.2010.00719.x

Wagenmakers, E.-J., Lee, M. D., Lodewyckx, T., & Iverson, G. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), Bayesian Evaluation of Informative Hypotheses, pp. 181-207. Springer: New York.