Universiteit Leiden

Faculteit der Sociale Wetenschappen

Decision Trees:

Amelioration, Simulation, Application

Master's Thesis

Michelle van der Geest

Master's Thesis Methodology and Statistics Master

Methodology and Statistics Unit, Institute of Psychology,

Faculty of Social and Behavioral Sciences, Leiden University

Date: August 2018

Student number: 1754378

Supervisor: Dr. Dusseldorp

Abstract

The focus of most RCTs is to find out which treatment works best on average. Patients, however, benefit most when receiving a treatment that works best based on their own pre-treatment characteristics. When a subgroup of the population benefits the most from a treatment different from the treatment that benefits another part of the population most, a qualitative subgroup-interaction is present. QUINT was developed to find these interactions. Nevertheless, these interactions are relatively often not found. This makes QUINT inferior to another tree-based method, MOB. The present study aims to improve QUINT in order to find those interactions more often. An adapted version of QUINT is compared to MOB, to see if the methods are now equally effective. The simulation study shows that QUINT now performs better than MOB in terms of a lower Type I error rate (0.323 versus 0.589) and similar proportions correctly assigned (0.738 or 0.803 versus 0.793) and Type II error rates (0.216 versus 0.251). To demonstrate and justify the new version of QUINT, an application study is performed. This study shows that the adapted version is at least as good as the current version of QUINT. A limitation of the simulation study is the small sample sizes used. Future research could address this limitation as well as add an extra evaluation criterion to the simulation study and compare QUINT to other tree-based methods designed to find treatment-subgroup interactions. In conclusion, the adaptation of QUINT appears to be successful.

Table of contents

Introduction

Evidence-based medicine requires researchers and medical practitioners to ask scientific questions, observe, analyse and record evidence (Peile, 2004). Part of asking scientific questions is conducting experiments. To this aim, randomized controlled trails (RCTs) are used. The focus of most RCTs is to find out which treatment works best on average. According to Epstein and Sherwood (1996), it is difficult to retrieve information about a patient’s individual outcome from RCTs. A patient’s outcome may depend heavily on characteristics of the physician and on characteristics of the patient himself (Albisser, 2000).

When subgroups of patients, which differ in their characteristics, vary in the efficacy of one or more treatments, there is differential treatment efficacy. This interaction between patient characteristics and treatment efficacy can be either quantitative or qualitative. In a quantitative interaction, one treatment is always better than the other, but how much better varies with the patient characteristics. In a qualitative interaction, one treatment is better for some patients, while the other treatment is better for other patients.

Since the article by Epstein and Sherwood (1996) several tree-based methods have been developed that can find subgroups of patients with the use of data from RCTs. Most of those methods make no distinction between quantitative and qualitative interactions, such as STIMA (simultaneous threshold interaction modeling algorithm; Dusseldorp, Conversano, & Van Os, 2010; Dusseldorp & Meulman 2004), Interaction Trees (Su, Tsai, Wang, Nickerson, & Li, 2009), MOB (Model-based recursive partitioning; Zeileis, Hothorn, & Hornik, 2008), Virtual Twins (Foster, Taylor, & Ruberg, 2011), and SIDES (subgroup identification based on differential effect search; Lipkovich, Dmitrienko, Denne, & Enas, 2011). However, a patient and its medical practitioner are usually most interested in which treatment works best for the patient rather than the efficacy of the treatment compared to another treatment. To this end QUINT (qualitative interaction trees) was developed: a tree-based method that searches for qualitative interactions. Figure 1 shows an example of QUINT used by Dusseldorp, Doove and Van Mechelen (2015) for comparing two mutually exclusive treatments for depression in patients with breast cancer. The figure shows that the total group of patients is partitioned into three classes: P_1, P_2 and P_3 (see the leaves). A certain group of patients, the group with low dispositional optimism, many negative social interactions and a low treatment extensiveness, is better off receiving treatment 1 (partition class 1), a nutrition-based treatment, than treatment 0. In contrast, patients with a low dispositional optimism score and with few negative social interactions and patients with a dispositional optimism score between 18.5 and 21.5, are better off receiving treatment 0 (partition class 2), an education-based treatment, instead of treatment 1. A third group (partition class 3) is indifferent to whether it receives treatment 0 or treatment 1. In this group (leaves 3 and 5) the nutrition-based treatment is

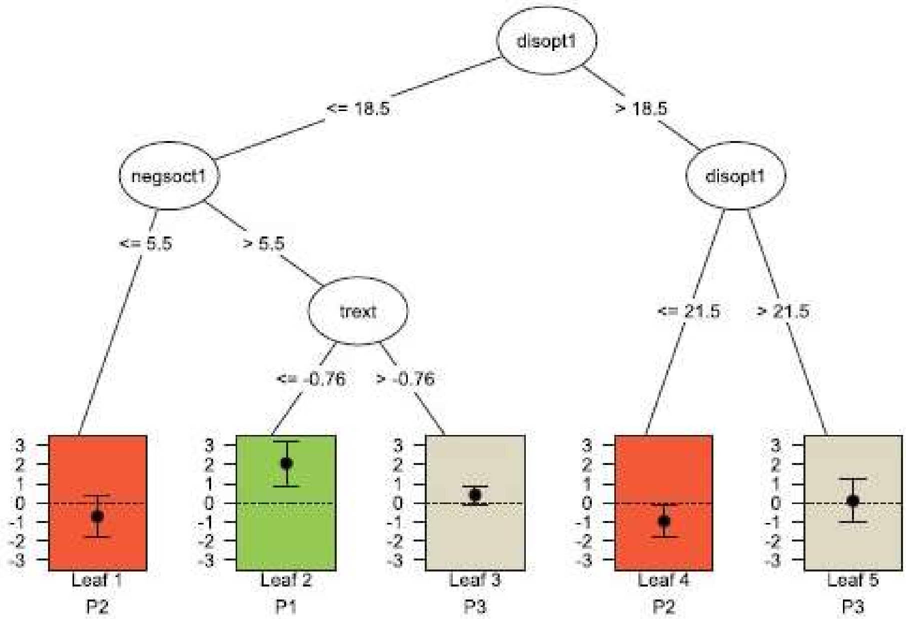


Figure 1. Example of a pruned qualitative interaction tree for the outcome Improvement in depression using the Breast Cancer Recovery Project data, as produced by the package quint. The splitting variables are: disopt (dispositional optimism), negsoct1 (negative social interaction), and trext (treatment extensiveness index). Each leaf of the tree is assigned to one of the three subgroups denoted in the figure by P_1, P_2, and P_3, respectively, and visualized by different colors of the leaves (green, red, and grey). P_1 means treatment 1 is best, P_2 means treatment 0 is best and P_3 means the treatments are equally effective. The vertical axis of the leaves pertains to the effect size *d*. Reprinted from Dusseldorp, Doove and Van Mechelen (2015), p. 3.

equally effective as the education-based treatment. Without a search for qualitative interactions, every patient were to receive the nutrition-based treatment, since this treatment is best overall.

In the present study, QUINT (Dusseldorp & Van Mechelen, 2014) is compared to MOB (Zeileis et al., 2008) on the effectiveness in assigning patients to the most effective treatment. Both methods build trees suitable for two alternative treatments and were compared to each other earlier (Sies & Van Mechelen, 2016; Van der Geest, 2017). According to Sies and Van Mechelen (2016) QUINT performed worse than MOB, especially when comparing the Type II error rates. However, Sies and Van Mechelen (2016) used models with both a qualitative interaction as well as a quantitative interaction. Since QUINT was built specifically to find qualitative interactions, we performed a pilot simulation study based on the beforementioned study with the emphasis on qualitative interactions (Van der Geest, 2017). The results of this pilot study showed that the current version of QUINT (version 1.2) has a large Type II error rate. A closer inspection of the results revealed that one of the current stopping rules of QUINT (version 1.2) was too conservative. This stopping rule is based on the qualitative interaction criterion. Practically, this means a tree is allowed to grow when the treatment outcomes in each of the leaves differ by a critical minimum absolute value (d_{min}). To improve the Type II error rate of QUINT, we propose to apply the qualitative interaction criterion at a later moment in the pruning process.

The present study consists of three stages. In the first stage, a new version of QUINT (version 2.0) is presented with an adjusted stopping rule. In the second stage a simulation study is performed to evaluate the effectiveness of MOB and QUINT (version 2.0). The following research questions are investigated:

1. Do QUINT (version 2.0) and MOB differ in their proportion of patients assigned to the best treatment?

2. Do QUINT (version 2.0) and MOB differ in their Type I error rate and Type II error rate?

In the third stage, the results of the adapted version of QUINT (version 2.0) are compared to the results of QUINT (version 1.2). The latter results were described in Formanoy et al. (2016) in a study regarding the efficacy of a physical or social environmental intervention in reducing the need for recovery from work for office workers. For this third stage, the following research question is formulated:

3. Do the subgroups found by QUINT (version 2.0) differ from the subgroups found by QUINT (version 1.2)?

Amelioration

Motivation

A requirement to grow a tree by QUINT is that there are at least two subgroups of patients, P_1 in which patients are better off receiving treatment 1 than treatment 0 and P_2 in which patients are better off receiving treatment 0 than treatment 1, having a certain difference in means of outcome *Y*. This difference in means can either be unstandardized or standardized (the latter being Cohen's effect size *d*). Dusseldorp & Van Mechelen (2014) showed that a standardized difference in means ($d_{\min}$) with an absolute value of 0.3 or higher accompanies an acceptable Type I error. Hence a minimal absolute value of 0.3 for $d_{\min}$ is required to grow a tree. This $d_{\min}$ needs to be present in P_1 as well as P_2. Otherwise, the effect size in the leaf with an absolute value of $d_{\min}$ below 0.3 is too low to affirm that one treatment is better than the other. Hence, it is then not allowed to claim a qualitative treatment-subgroup effect exists.

In the present version of QUINT (1.2), the qualitative interaction requirement is assessed at the first split. This could lead to the rejection of a tree in the earliest stages of fitting the tree. The change in the QUINT algorithm encompasses delaying the inspection of the qualitative interaction requirement to the pruning stage. Hopefully this will result in fewer incorrect rejections of trees and hence increase the power of QUINT.

Adaptation

To do this, two changes are made in the tree-growing stage. First, the check on the qualitative interaction is removed. Before, there was a check on the effect size of the differences in treatment outcomes in the two leaves of the tree after the first split. This hindered QUINT from returning trees in which the qualitative interaction shows only at a later stage such as in Figure 2. The effect sizes in the two leaves of this tree after the first split would be too small (i.e. 0.24 and -0.21) to comply with the check.

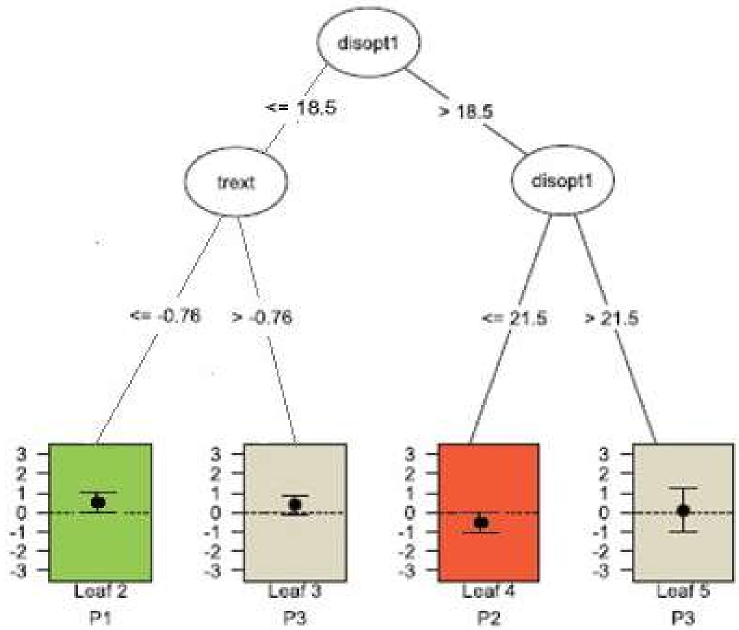


Figure 2. Example of a tree with a qualitative interaction that QUINT (version 1.2) does not find.

Second, with this check removed, the tree-growing stage does not stop immediately in the specific situation when all patients actually belong in the root node. After each split in the tree-growing stage, a value is calculated that takes into account the difference in treatment outcome between treatment 0 and treatment 1 and the sample sizes of class 1 and 2. Ideally, this value is as high as possible (on a scale from 0 to 4). The tree-growing stage continues as long as this value, *C*, keeps increasing. With *C* being 0 after the first split, a second split is made. However, this was not possible with the qualitative interaction check present in the tree-growing stage. To prevent QUINT from making a second split in this specific situation, an adaptation is made. The new algorithm of the tree-growing stage can be found in Appendix A (with the removed qualitative interaction in red and the added adaptation in green).

In the pruning stage, no part of the algorithm is removed. Instead, two features are added. One feature deals with the above situation that all patients are actually in the root node (i.e. there is no subgroup-treatment interaction)[1]. The adaptation stops the pruning stage and returns the same tree as the tree-growing stage does. The other feature is a check on the qualitative interaction that was removed from the tree-growing stage. This time, the effect sizes of the differences in treatment outcomes in all the leaves present in the pruned tree are used to check for the presence of a qualitative interaction. If all absolute values of the effect sizes of the leaves assigned to the first treatment or all absolute values of the effect sizes of the leaves assigned to the second treatment are smaller than the qualitative interaction condition requires, all patients should receive the same treatment and the tree is pruned back to the root node. The new algorithm of the pruning stage can be found in Appendix B (with the added adaptation in green).

Simulation

Data generation

The rows of each generated data set represent the patients receiving a treatment, and the columns are the following attributes of the patients:

[1] This is necessary for the simulation study. In the simulation study, every tree reaches the pruning stage. Normally, when using patient data in which *C* is 0, the pruning stage is not being used. If someone were to accidentally use the pruning stage in this situation, however, no strange error message is returned.

- the pre-treatment characteristics of the patients, X_j (with $j = 1, ..., J$),
- the treatment alternative A to which the patient is randomly assigned (with $A = 0$ being assigned to treatment 0 and $A = 1$ being assigned to treatment 1),
- the true optimal treatment (g^{opt}).

Data were generated according the following true model:

$Y_i = \mu(A, X) = 1.0 + 0.25X_1 + 0.25X_2 - 0.25X_5 - d[A - g^{opt}(\boldsymbol{X})]^2 + \epsilon_i$,

where Y denotes the outcome variable, i stands for the individual, d equals Cohen's d which was a design factor, and ϵ_i represents the error term, having a standard normal distribution. From this model, four true scenarios were created, differing in the definition of g^{opt}. Both the data generation and the scenarios are based on Sies and Van Mechelen (2016) and are exactly the same as in Van der Geest (2017). The scenarios are:

Scenario (A) $g^{opt}(X) = I(X_1 > -0.545)]I(X_2 < 0.545)$,

Scenario (B) $g^{opt}(X) = I(X_1 < -0.545)I(X_2 > 0.545) \mid I(X_1 < -0.545)I(X_2 < 0.545)I(X_3 < 0.545)$,

Scenario (C) $g^{opt}(X) = I(X_1 > X_2^2)$,

Scenario (D) $g^{opt} = 1$.

The above-mentioned definitions used for g^{opt} make use of a maximum of three characteristics of a patient: X_1, X_2 and X_3. In Scenario (A) and (C) X_1 and X_2 are included, in Scenario (B) also X_3 is included and in Scenario (D) no patient characteristics are included. In Scenario (D) all patients are better of receiving treatment 1. This scenario is used to estimate the Type I error of the methods. In Scenario (A), (B) and (C) sometimes treatment 0 is better and sometimes treatment 1 is better: these scenarios involve qualitative interactions. However, only Scenarios (A) and (B) are tree-based interactions (i.e., using thresholds). Since Scenario (B) includes the most person characteristics, this scenario is regarded complex and Scenario (A) and (C) are considered simple.

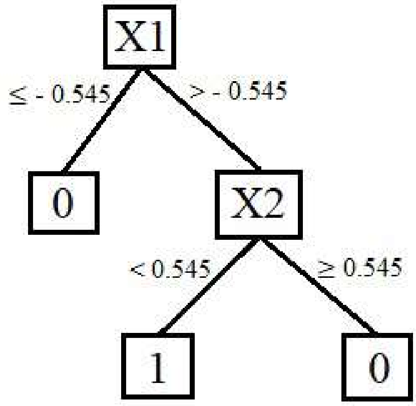


Figure 3. Decision tree of Scenario A

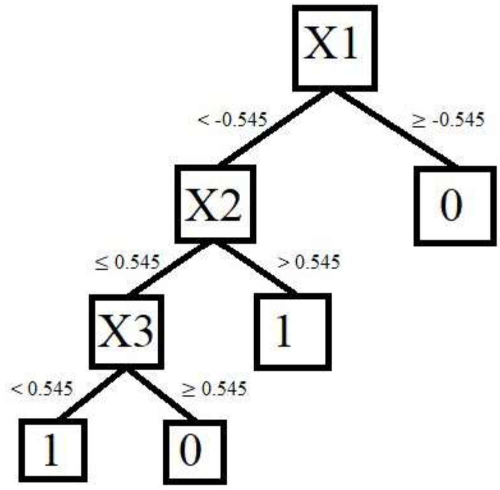
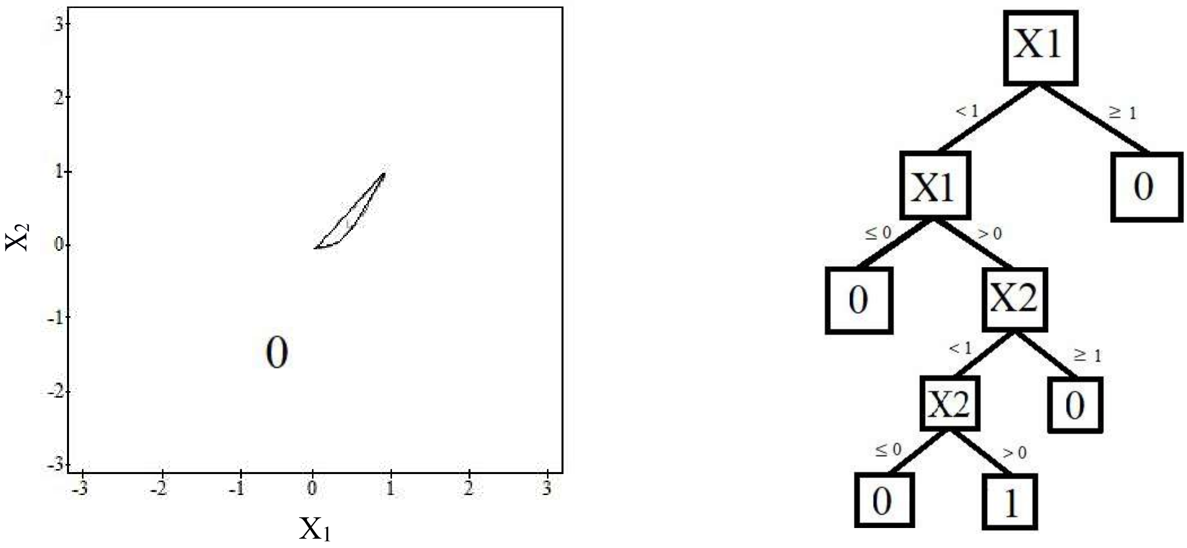


Figure 4. Decision tree of Scenario B

Figures 3 and 4 depict Scenario (A) and (B) as a decision tree. For Figure 3, when pre-treatment characteristic X_1 is -0.545 or lower, treatment 0 is the best alternative. When X_1 is higher than -0.545 and X_2 is lower than 0.545, treatment 1 is the best alternative. Treatment 0 is again the best alternative for the situation when X_1 is higher than -0.545 and X_2 is equal to or higher than 0.545. Figure 4 is the mirror image of Figure 3 made more complex by adding pre-treatment characteristic X_3. Figure 5 visualizes Scenario (C) in two ways: in a grid and as a decision tree. When pre-treatment characteristic X_1 and X_2 are equal to or lower than 0 or equal to or higher than 1, treatment 0 is the best alternative. When X_1 and X_2 are between 0 and 1, the best treatment depends on the exact combination of the values. Therefore, it is not possible to show the exact decision tree. In most cases, the bottom right leaf in Figure 5(b) is split further. Looking at Figure 5(a), one may get the impression that treatment 1 is almost never preferred above treatment 0. However, since the pre-treatment characteristics have mean 0 and standard deviation 1, a pre-treatment characteristic has a 34% chance of ranging between 0 and 1. As said earlier, in Scenario (D) all patients are better of receiving treatment 1. Therefore, no decision tree visualizes this scenario.

The pre-treatment characteristics of the patients are generated from a multivariate



(a) (b)

Figure 5. Scenario C shown in a grid (a) and as a decision tree (b). In (a) the area inside the curved shape is assigned treatment 1. Everything outside the shape is assigned treatment 0.

normal distribution *N*(0, 1) and a varying correlation $\rho_{jj'}$. The treatment alternative *A* has a Bernouilli distribution with θ = 0.50. This means patients are randomly assigned to the treatments and each patient is equally likely to be assigned to treatment 0 as treatment 1. The optimal treatment g^{opt} is the treatment regime that maximizes the expected potential outcome. In other words, the treatment the patient receives if the patient's characteristics are optimally used in the decision for a treatment alternative. The optimal treatment values are indicated by 0 and 1 in the leaves of the trees as can be seen in Figure 3 and 4.

Monte Carlo simulation design

Data sets are created based on a full factorial design, including the following factors:

- Sample size (*N*) has the values 150 and 300;
- Number of pre-treatment characteristics (*J*) has the values 5 and 20;
- Effect size (Cohen's *d*) has the values 0.5 and 1;
- Correlation between the pre-treatment characteristics (ρ) has values 0, 0.2 and 0.4;

- Type of scenario based on the true optimal treatment regime, g^{opt} (four types A, B, C, D, see **Data generation**).

Crossing all factors results in 2 (sample size) x 2 (number of pre-treatment characteristics) x 2 (effect size) x 3 (correlations between pre-treatment characteristics) x 4 (optimal treatment regime) = 96 combinations, with 100 Monte Carlo replications for each cell, resulting in 9,600 data sets. Each of these data sets will be analyzed by QUINT version 2.0 and MOB.

Analysis of simulated data sets

In this paragraph, the options and tuning parameters for the methods MOB and QUINT as specified in Sies and Van Mechelen (2016) will be described.

Model-based Recursive Partitioning. For estimating the tree-based treatment regimes with MOB, the R-package 'party' (version 1.3-0) was used (Zeileis et al., 2008) . As input argument, we used $Y = \beta_0 + \beta_1 A$ as the model, the possible split variables are the pre-treatment characteristics $X_1, ..., X_J$, and the tuning parameters do not deviate from the default settings with the exception of the minimum number of persons in a node to split: this number was set to 40. See Appendix C for example code.

Qualitative Interaction Trees. For estimating the tree-based treatment regimes with QUINT, the package 'quint' was used with an adaptation of the function quint() and the pruning function prune.quint(), based on our amelioration (see **Amelioration** and Appendices A and B). As mentioned in the paragraph **Amelioration**, the partitioning criterion used to grow a tree can be unstandardized (referred to as ‘difference in means’) or standardized (referred to as ‘effect size’). The default setting is effect size. The tuning parameters do not differ from the default settings, except for the minimum number of persons with $A = 0$ or $A = 1$ in a leaf: this was set at 10 per treatment[2]. See Appendix C for example code.

[2] There are 2 treatments and 2 daughter nodes per split, so 10 persons per treatment results in 2 x 2 x 10= 40 persons per split. This is the same amount required by MOB.

Evaluation criteria and analysis

The performances of QUINT and MOB were compared based on the following evaluation criteria:

- Criterion 1: Proportion of patients assigned to their best treatment alternative;
- Criterion 2: Type I error rate;
- Criterion 3: Type II error rate.

Criterion 1 is the proportion of patients assigned to their best treatment alternative and is estimated with Scenario (A), (B), (C) and (D). The amount of people that are assigned to the same treatment by the method under study as they should receive according to the true optimal treatment regime is divided by the total number of patients. The proportion of patients assigned to the best treatment alternative by QUINT depends on how patients assigned to class 3 are treated. Those patients benefit equally from both treatments and are thus never assigned wrongly. To deal with this situation, statistics are given with class 3 included in the proportion of patients good assigned and with class 3 excluded.

Criterion 2 is the Type I error rate: the probability that the null hypothesis is incorrectly rejected. The null hypothesis in this situation is that every patient should receive the same treatment alternative. Hence, in this specific situation a Type I error is present when a tree with two or more leaves is returned. To estimate this probability only Scenario (D) is used, since the other scenarios involve a qualitative interaction.

Criterion 3, lastly, is the Type II error rate: the probability that the null hypothesis is incorrectly accepted. A Type II error is therefore present when no tree is returned whereas a tree should have been returned. To estimate this probability Scenario (A), (B) and (C) are used, since these are the scenarios with a qualitative interaction.

QUINT (version 2.0) and MOB are compared to each other using a repeated measures analysis of variance (ANOVA) for each of the 3 outcome measures with method as a within-subjects variable and the design factors as the between-subjects variables. Since we are only interested in the performance of the methods rather than the overall performance of the design factors, we only report the within-subjects effects. In order to be reported, a main effect or interaction effect needs to be substantial, with substantial being defined as accounting for a certain percentage of the total within-subjects sum of squares. According to Cohen (1988), $\eta^2 = .06$ is a medium-sized effect size. Therefore, all effects of $\eta^2 \geq .06$ are reported. This means a substantial effect accounts for six percent or more of the total within-subjects sum of squares. Appendices D-G give an overview of the tests of the within-subjects effects as provided by SPSS (with η^2 being calculated separately).

Results

Proportion of patients assigned to the best treatment alternative. Overall, the mean proportion of patients assigned to the best treatment by QUINT is 0.738 with class 3 excluded and 0.803 with class 3 included. The mean proportion assigned to the best treatment by MOB is 0.793. According to the ANOVA, there is a main effect of method only when class 3 is excluded ($\eta^2 = .08$, see Appendix D). In contrast, with class 3 included, the proportion of patients assigned to the best treatment is influenced by the interaction between method and scenario ($\eta^2 = .13$, see Appendix E). In scenario A and C, QUINT (0.819 resp. 0.729) performs better than MOB (0.747 resp. 0.668), whereas MOB (0.874 resp. 0.883) outperforms QUINT (0.837 resp. 0.829) in scenario B and D (see Figure 6).

Type I error rate. The mean of the Type I error rate of QUINT is 0.323, whereas the Type I error rate of MOB is 0.589. According to the ANOVA, the error rate is influenced by

Table 1

Effects with $\eta^2 \geq .06$ resulting from the ANOVA on the Type I error rates

Effect	η^2
Method	.14
Method*effectsize	.11
Method*J	.09
Method*n	.06

the method used, the interaction between method and sample size, the interaction between method and the number of pre-treatment characteristics and the interaction between method and effect size (see Table 1). Besides the main effect, there are three interaction effects: (1) the Type I error rate is lower with a sample size of 150 than of 300, and this difference in error rate is much larger for MOB (0.399 resp. 0.778) than for QUINT (0.309 resp. 0.337, see Figure 7); (2) the Type I error rate is lower for QUINT with 5 pre-treatment characteristics (0.262) than with 20 pre-treatment characteristics (0.384). For MOB, however, it is the other way around (0.733 resp. 0.445, see Figure 8), and (3) QUINT and MOB differ little in Type I error rate when the effect size is .5 (0.569 resp. 0.607) and both have a lower Type I error rate when the effect size is 1 (0.077 resp. 0.570). The difference in Type I error rate, however, is much higher for MOB than it is for QUINT (see Figure 9). All Type I error rates are above the reference line, meaning none of the Type I error rate is acceptable.

Type II error rate. The mean of the Type II error rate of QUINT is 0.216, whereas the Type II error rate of MOB is 0.251. According to the ANOVA, this error rate is influenced by the interaction between method and sample size (η^2= .12). QUINT (0.206) performs better than MOB (0.443) when the sample size is small. MOB (0.059) outperforms QUINT (0.226) when the sample size is larger. This is due to a better performance of MOB (see Figure 10). The reference line shows that the Type II error rate of QUINT with a small sample size and of

MOB with a higher sample size are acceptable. There is no substantial difference between the Type II error rate of QUINT and MOB.

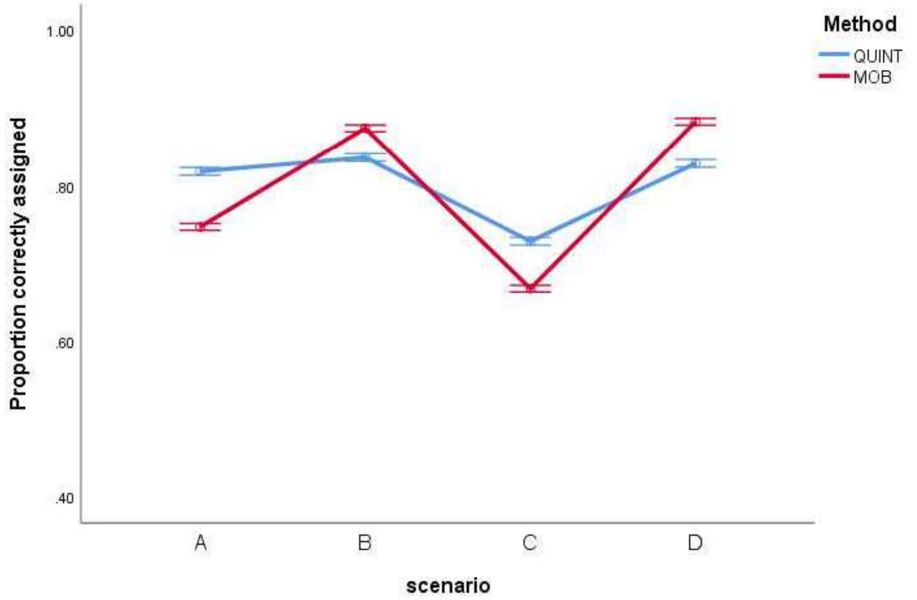


Figure 6. Proportion correctly assigned patients of QUINT and MOB depending on scenario. Error bars represent the 95% confidence intervals.

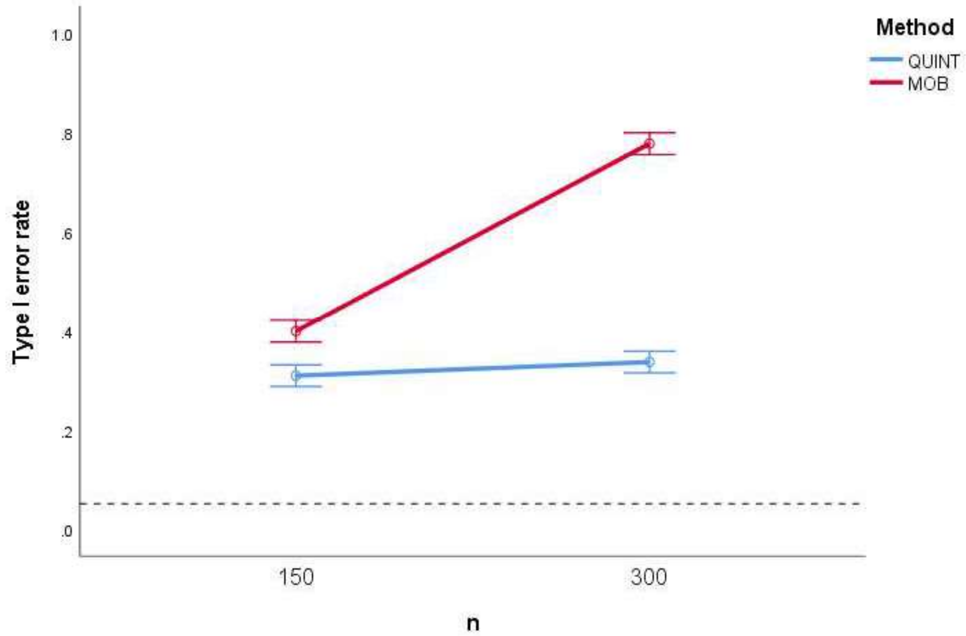


Figure 7. Type I error rate of QUINT and MOB depending on sample size. Error bars represent the 95% confidence intervals. A reference line with an acceptable Type I error rate of 0.05 is added.

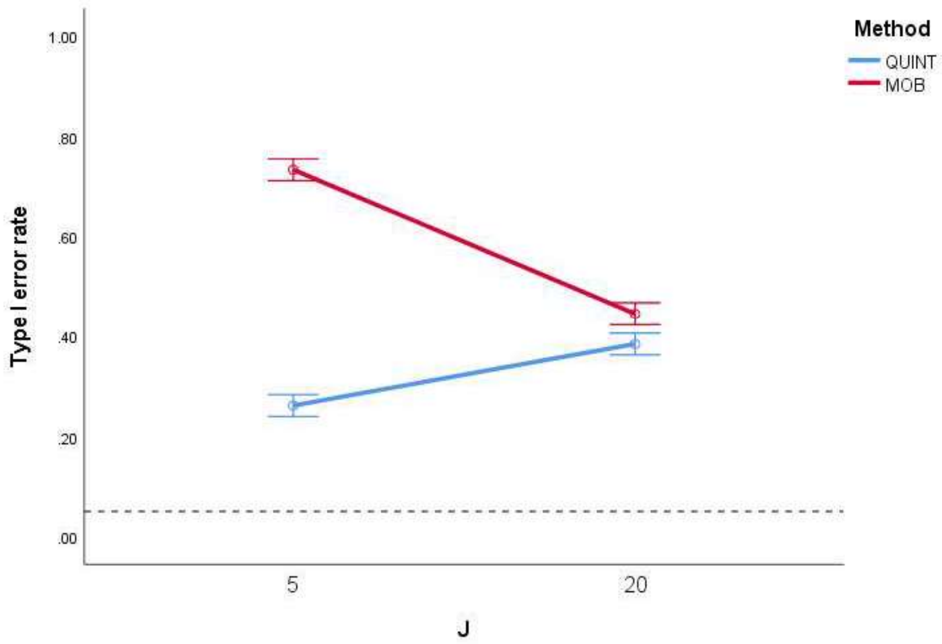


Figure 8. Type I error rate of QUINT and MOB depending on the number of pre-treatment characteristics. Error bars represent the 95% confidence intervals. A reference line with an acceptable Type I error rate of 0.05 is added.

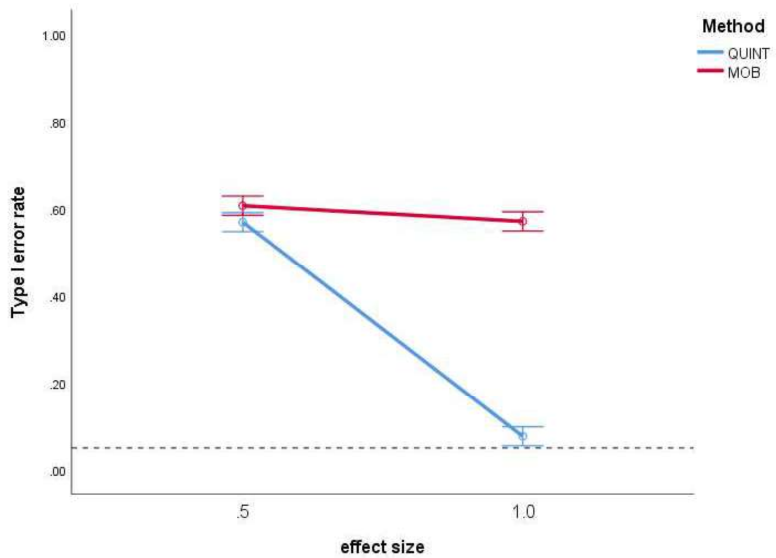


Figure 9. Type I error rate of QUINT and MOB depending on effect size. Error bars represen the 95% confidence intervals. A reference line with an acceptable Type I error rate of 0.05 is added.

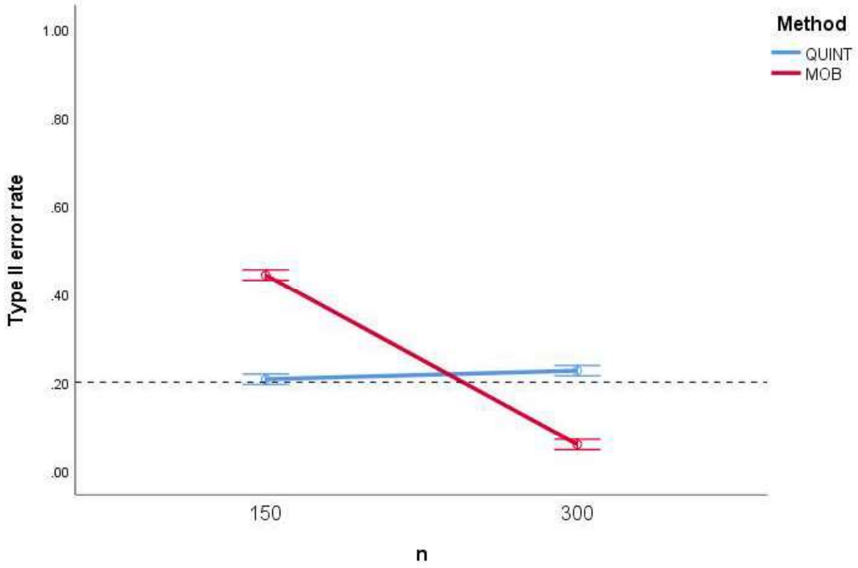


Figure 10. Type II error rate of QUINT and MOB depending on sample size. Error bars represent the 95% confidence intervals. A reference line with an acceptable Type II error rate of 0.20 is added.

Application

Introduction

The “Be Active & Relax” study consists of 329 office workers between the ages of 19 and 63 years (*M*= 42.10, *SD*= 9.95) of which their need for recovery from work (NFR) is measured. The office workers partake in a social environmental intervention, a physical environmental intervention, both or neither to see if it affected their NFR. Hence, there are four conditions. In order to analyse the data with QUINT, there need to be only two intervention groups. An evaluation of the data showed that there is no interaction effect between the social environmental intervention and the physical environmental intervention (Formanoy et al., 2016). The data is therefore suitable for analysing the unique contribution of both interventions. This means that the data can be used to investigate whether a subgroup of office workers is better (worse) off receiving a social environmental intervention instead of receiving no intervention and whether a subgroup of office workers is better (worse) off receiving a physical environmental intervention instead of receiving no intervention.

The social environmental intervention consists of four group sessions in which office workers of the same team are interviewed and motivated about physical activity and relaxation. The physical environmental intervention consists of applying changes to the work environment to increase physical activity and relaxation. This is done by adding table tennis tables, exercise balls, standing tables, footprints on stairs, posters, bar chairs, lounge chairs and noise reducing curtains.

Participants were measured at baseline and after 12 months. As evaluation criterion change in NFR was used. Neither the social environmental intervention nor the physical environmental intervention led to an overall decrease (increase) in NFR. Thus, neither one of the interventions made sure the office workers experienced on average less (more) work related fatigue. The same is true for participants receiving both interventions.

In this study, 25 baseline characteristics of the aforementioned study are used in search for a possible superior intervention per group of office workers. The baseline characteristics are NFR at baseline, age, sex, level of education, cohabiting, mother country, BMI, mental health, detachment at home, relaxation at home, physical activity, vitality, team commitment, organizational commitment, supervisor support, colleague support, job demands, decision authority, job insecurity, skill discretion, working overtime, detachment at work, relaxation at work, walking during lunch and active during lunch. Table 2 shows the baseline characteristics per condition. As can be seen in Table 2, not all participants were analysed. Out of the 329 study participants, 304 provided all the data. One moderator variable, general health, had 8 missing values and was not of any importance. Hence, this variable is removed and 312 office workers are used in the analyses.

The results of Formanoy et al. (2016) show that subgroups found by QUINT (version 1.2) depend on the partitioning criterion (either difference in means or effect size) used to grow the tree. Since the change in NFR can be viewed as ordinal as well as numeric, both

Table 2

Descriptive statistics for all variables involved in re-analyses of data from the "Be Active & Relax" study. The potential moderators were all measured at baseline (i.e., before receiving a physical activity and relaxation program). The statistics are given for both delivery modes: the social intervention and the physical environmental intervention (N = 312).

			Delivery mode: social environmental intervention				Delivery mode: physical environmental intervention			
			Yes $n = 149$		No $n = 163$		Yes $n = 132$		No $n = 180$	
Variable	Range		Mean	(*SD*)	Mean	(*SD*)	Mean	(*SD*)	Mean	(*SD*)
Outcome										
Improv. in need for recovery	−100.0	81.82	3.82	(25.32)	1.17	(21.98)	4.59	(23.38)	0.85	(23.75)
Potential moderators										
Need for recovery at baseline	0.00	100.00	30.42	(28.90)	30.36	(28.88)	32.18	(30.28)	29.08	(27.75)
Age (in years)	19.00	63.00	42.46	(10.05)	41.77	(9.89)	41.63	(10.39)	42.45	(9.63)
Sex (male vs. female)	0	1	0.62	(0.48)	0.63	(0.48)	0.62	(0.49)	0.63	(0.48)
Level of education	1.00	3.00	2.29	(0.86)	2.41	(0.77)	2.46	(0.76)	2.27	(0.84)
Cohabiting (yes vs. no)	0	1	0.75	(0.43)	0.77	(0.42)	0.8	(0.40)	0.74	(0.44)
Mother country (Neth. vs. other)	0	1	0.93	(0.26)	0.91	(0.29)	0.93	(0.25)	0.91	(0.29)
Body Mass Index	17.10	39.19	25.18	(4.35)	24.87	(3.74)	24.61	(3.56)	25.31	(4.34)
Mental Health	2.00	6.00	4.5	(0.72)	4.51	(0.73)	4.42	(0.69)	4.57	(0.74)
Detachment at home	1.00	7.00	4.76	(1.33)	4.9	(1.35)	4.8	(1.39)	4.86	(1.31)
Relaxation at home	2.00	7.00	5.16	(1.02)	5.25	(1.11)	5.05	(1.07)	5.33	(1.05)
Physical activity (in MET-min.)	375	29610	7527	(4234)	7521	(3937)	7066	(4018)	7860	(4095)
Vitality	2.00	7.00	5.00	(0.96)	5.06	(1.00)	4.92	(0.97)	5.11	(0.98)
Team commitment	1.00	5.00	4.07	(0.65)	4.14	(0.68)	3.99	(0.64)	4.19	(0.67)
Organizational commitment	2.57	5.00	4.00	(0.47)	4.08	(0.44)	3.97	(0.44)	4.09	(0.46)
Supervisor support	1.00	4.00	2.87	(0.51	2.89	(0.48)	2.86	(0.53)	2.89	(0.47)
Colleague support	2.00	4.00	3.09	(0.38)	3.09	(0.37)	3.05	(0.37)	3.12	(0.37)

Table 2 Continued

			Delivery mode: social environmental intervention				Delivery mode: physical environmental intervention			
			Yes $n = 149$		No $n = 163$		Yes $n = 132$		No $n = 180$	
Variable	Range		Mean	(*SD*)	Mean	(*SD*)	Mean	(*SD*)	Mean	(*SD*)
Potential Moderators										
Job demands	1.50	4.00	2.82	(0.49)	2.71	(0.40)	2.78	(0.46)	2.99	(0.44)
Decision authority	1.00	4.00	2.98	(0.53)	2.99	(0.54)	2.98	(0.56)	2.99	(0.52)
Skill discretion	1.83	4.00	3.03	(0.37)	3.09	(0.37)	3.1	(0.39)	3.03	(0.35)
Working overtime (in hrs. p. wk.)	0.00	40.00	2.85	(6.05)	3.19	(7.78)	2.74	(6.70)	3.25	(7.22)
Detachment at work	1.00	7.00	3.48	(1.39)	3.54	(1.34)	3.46	(1.30)	3.54	(1.41)
Relaxation at work	1.00	7.00	3.53	(1.25)	3.69	(1.31)	3.45	(1.19)	3.74	(1.33)
Walking during lunch	1	5	2.78	(1.45)	2.94	(1.47)	2.86	(1.39)	2.87	(1.52)
Active during lunch	1	4	1.92	(1.04)	1.91	(1.04)	1.83	(0.98)	1.97	(1.08)

Retrieved from Formanoy et al. (2016. p. 6).

criteria are used. When the criterion is effect size, the qualitative interaction tree for the social environmental intervention is a pruned tree with two leaves and “Age” as spitting variable. When using the same criterion for the physical environmental intervention, the result is a pruned tree with two leaves and “Working overtime” as spitting variable. When the partitioning criterion is difference in means, both interventions result in a pruned tree with four leaves and the above-mentioned splitting variables plus the splitting variables “Organizational commitment” and “Working overtime” respectively “Team commitment” and “Physical activity”.

The results of Formanoy et al. (2016) are retrieved with a d_{min} set at 0.299. With a higher d_{min} QUINT (version 1.2) is not able to retrieve a tree for the physical environmental intervention using effect size as the partitioning criterion. This value of d_{min} is problematic, since d_{min} should be based on a balance between the type I error and the type II error. Dusseldorp and Van Mechelen (2014) showed that with d_{min} set at 0.3, a good balance between the two can be obtained.

Analysis strategy

As mentioned earlier, in the present application we use both partitioning criteria in the analysis: effect size and difference in means. Doing the analysis with both criteria gives information about the stability of the trees found. As in Formanoy et al. (2016), first the analyses with the effect size criterion are reported and then the analyses with the difference in means criterion. Both series of analysis required 25 as the minimum sample size per intervention group per leaf, d_{min} was set at 0.3, the default value for maximum number of leaves (i.e. maxl= 10), and the default values of the weights of the partitioning criterion were used (i.e. w_1 is 1 / log(1 + $IQR(Y)$) if the difference in means criterion is used and 1 / log(1 + 3) if the effect size criterion is used and w_2 is 1 / log(0.50N)). To grow the tree, 1000 bootstrap samples were used, and in pruning the tree, the one-standard-error pruning rule is used. To test the difference in means of the two groups in each leaf of the pruned tree, independent t-tests were performed. Since the significance level of the t-tests are inflated, bias-corrected effect sizes in the leaves are given. These were estimated using a validation procedure for small data sets found in QUINT.

Results

Trees with criterium Effect size. The qualitative interaction tree for the social environmental intervention is a pruned tree with two leaves. The variable "Age" is the splitting variable with a split point of 46.5 years. Figure 11 displays the tree.

The qualitative interaction tree for the physical environmental intervention is a pruned tree with two leaves. The variable "Working overtime" is the splitting variable with a split point of 2.25 hours. Figure 12 displays the tree. Table 3 gives the descriptive statistics of the

Table 3

Descriptive statistics in the leaves of the results for QUINT (version 2.0) for the social environmental intervention (SEI; Figure 11) and the physical environmental intervention (PEI; Figure 12).

	n	Mean	*SD*	*n*	Mean	*SD*	Difference in means (95 % CI)	Bias-corrected effect size *d*
Fig. 4	SEI^+			SEI^-				
Leaf 1	90	8.29	22.27	107	-2.23	23.20	10.52 (4.12, 16.92)**	0.31
Leaf 2	59	-3.00	28.22	56	7.66	17.89	10.66 (-19.35, -1.96)*	-0.27
Fig. 5	PEI^+			PEI^-				
Leaf 1	103	6.15	23.90	128	-1.25	25.39	7.40 (0.99, 13.81)*	0.22
Leaf 2	29	-0.94	20.90	52	6.01	18.35	6.95 (-16.26, 2.36)	-0.05

The mean values and standard deviations on improvement in Need for Recovery (NFR) are displayed (with a higher score reflecting a larger reduction in NFR from baseline to 12-month follow-up), and the treatment outcome differences. CI: confidence interval; $**p < .01$; $*p < .05$, estimated with an independent t-test.

former trees. Comparing this table to Table 3 from Formanoy et al. (2016) shows that the results for trees with the difference in means criterium are the same for QUINT (version 2.0) as for QUINT (version 1.2).

Trees with criterium Difference in means. The qualitative interaction tree for social environmental intervention is a pruned tree with four leaves. The variable “Age” is the first splitting variable with a split point of 46.5 years, the variables “Organizational commitment” and “Working overtime” are the second and third splitting variables with split points of 3.94 and 0.75 hours respectively (see Figure 13).

The qualitative interaction tree for physical environmental intervention is a pruned tree with four leaves. The variable “Working overtime” is the first splitting variable with a split point of 2.25 hours, the variables “Team commitment” and “Physical activity” are the second and third splitting variables with split points of 3.83 and 7990 minutes respectively (see Figure 14). Figure 7 and 8 are the same as Figure 3 and 4 of Formanoy et al. (2016). QUINT (version 2.0) thus returns the same results as QUINT (version 1.2). Contrary to QUINT (version 1.2), QUINT (version 2.0) also returns a tree when a minimum effect size of 0.30 is used.

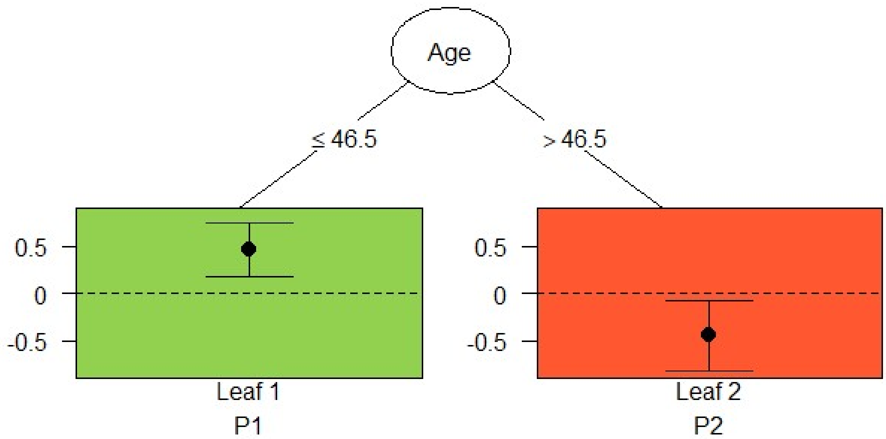


Figure 11. Pruned tree with splitting variable Age and a split point at 46.5 years. Office workers younger than 46.5 benefit from the social environmental intervention, but those older than 46.5 years are better off not receiving the intervention. The criterium used in this tree is the effect size.

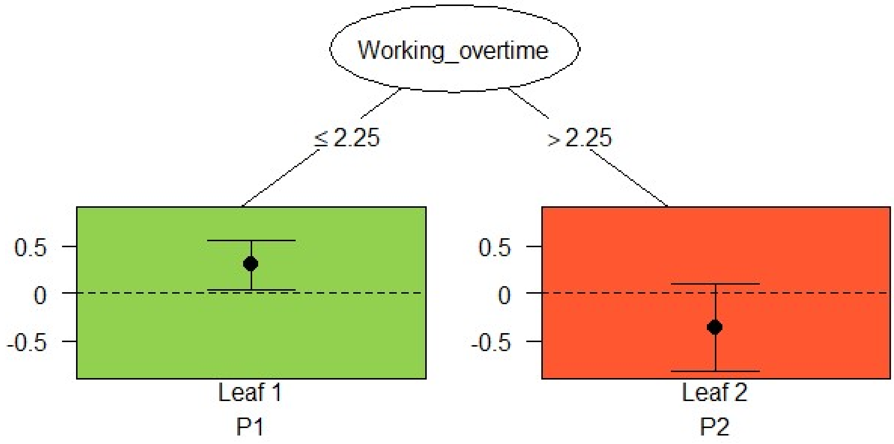


Figure 12. Pruned tree with splitting variable Working overtime and a split point at 2.25 hours. Office workers who work fewer hours overtime (≤ 2.25) have a better outcome with the physical environmental intervention than without the physical environmental intervention (Leaf 1) and those who work more hours overtime (> 2.25) have a worse outcome with the physical environmental intervention than without (Leaf 2). The criterium used in this tree is the effect size.

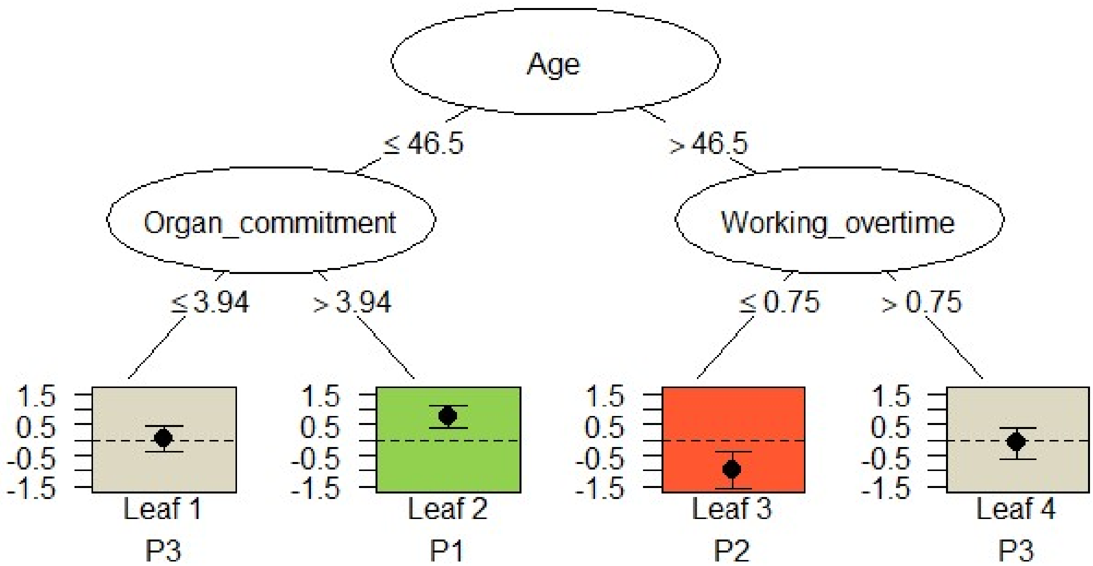


Figure 13. Pruned tree with splitting variables Age, Organizational commitment and Working overtime and split points at 46.5 years, 3.94 and 0.75 hours. Office workers younger than 46.5 and committed to the organization benefit from the social environmental intervention, but those older than 46.5 years and working few hours overtime are better off not receiving the intervention. The criterion used in this tree is difference in means. The measurement in the leaves, however, is the effect size.

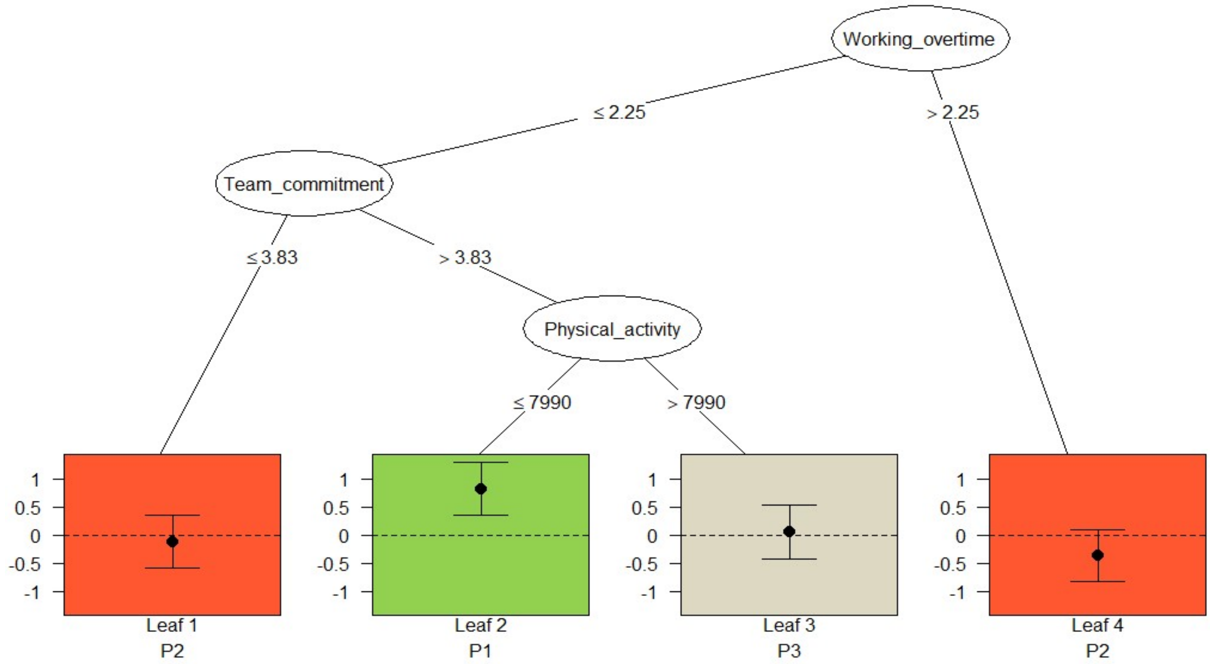


Figure 14. Pruned tree with splitting variables Working overtime, Team commitment and Physical activity and split points at 2.25 hours, 3.83 and 7990 minutes. Office workers who work few hours overtime, are committed to their team and are not that physical active have a better outcome with the physical environmental intervention than without. Those who work few hours overtime and are not that committed to their team or work more overtime have a worse outcome with the physical environmental intervention than without. The criterion used in this tree is difference in means. The measurement in the leaves, however, is the effect size.

Discussion

In this paper, QUINT is adapted with the aim to improve its Type II error rate. Subsequently, a simulation study is used to compare the new version, QUINT (version 2.0), to MOB on several criteria. The measures of evaluation are the proportion of patients correctly assigned, the Type I error rate and the Type II error rate. Ultimately, an application study is done to compare the subgroups that are found by QUINT (version 2.0) to the subgroups that are found by QUINT (version 1.2). The next paragraphs present the main findings of the simulation and the application study.

Findings

To provide an answer to the first research question, the proportion of patients assigned to the best treatment by QUINT (version 2.0) is compared to the proportion correctly assigned by MOB. As it turns out, whether there is a difference between the two methods depends upon the calculation of the proportion correctly assigned by QUINT. QUINT can assign patients to a subgroup that is indifferent to the assigned treatment alternative. One possibility is to consider this class incorrectly assigned. A second possibility is to consider this class correctly assigned. Using the first operationalization, MOB performs better than QUINT. This result can also be found in earlier studies that compared the methods to each other (Sies & Van Mechelen, 2016; Van der Geest, 2017). However, whereas other studies use this operationalization without second thought, it is not so straightforward how the proportion correctly assigned by QUINT should be calculated. While the first calculation takes into account that the worst treatment alternative is not ruled out as a possible treatment, the second calculation takes into account that the best treatment alternative is not ruled out as a possible treatment. Using the last operationalization, part of the difference between MOB and QUINT is accounted for by the interaction between method and scenario. This result can be found in earlier studies as well (Sies & Van Mechelen, 2016; Van der Geest, 2017). Either way the answer to the research question is not affected by the adaptation of QUINT. If we average the results of both operationalizations, QUINT and MOB do not differ in the proportion of patients assigned to the best treatment.

The second research question concerns the Type I error rate and Type II error rate of QUINT (version 2.0) and MOB. The Type I error rate of QUINT is lower than the Type I error rate of MOB. These error rates are influenced by interactions between method on one side and effect size, the number of pre-treatment characteristics and sample size on the other. These results were not found in the pilot study we performed (Van der Geest, 2017), but the first three results were found earlier (Sies & Van Mechelen, 2016). It should be noted that Sies and Van Mechelen (2016) used a higher cut-off value for the effect size. Using the same cut-off value in this study means the third and fourth result would not be present.

The Type I error rate of QUINT is much higher in the present study than in the pilot study (Van der Geest, 2017). The reverse is true for MOB. Since the Type I error rate is different for QUINT as well as for MOB it is highly likely that this is due to differences in the simulation design, i.e. smaller sample sizes and more iterations, rather than the adaptation of QUINT. Although the Type I error rate of QUINT is high, Dusseldorp and Van Mechelen (2014) show that this kind of error rate is to be expected with a medium- or large-sized effect size and a small sample size.

The Type II error rate is influenced by the interaction between method and sample size. This is in line with earlier research (Sies & Van Mechelen, 2016). There is no substantial difference between the Type II error rate of QUINT and the Type II error rate of MOB. This contrasts with findings from earlier research (Sies & Van Mechelen, 2016; Van der Geest, 2017). The Type II error rate (0.216) of QUINT (version 2.0) is clearly lower than the Type II error rate (0.776) of QUINT (version 1.2) as found in the pilot study. Since the sample sizes in the simulation study differ, direct comparison of the overall Type II error rate of QUINT is not appropriate, however. Both simulations do have Type II error rates for datasets consisting of 300 cases. With this sample size QUINT (version 2.0) still has a much lower Type II error rate than QUINT (version 1.2) (0.265 versus 0.717). The Type II error rate is changed for the better by the adaptation.

The third research question is answered by comparing the application of QUINT (version 2.0) to the application of QUINT (version 1.2) on data used in Formanoy et al. (2016). When the partitioning criterion is effect size, both versions of QUINT result in the same trees. When the partitioning criterion is difference in means, QUINT (version 1.2) fails to return a tree for the physical environmental intervention when in fact there is a qualitative interaction. QUINT (version 2.0) does return a tree in this situation. In this respect QUINT (version 2.0) is better. The application shows QUINT (version 2.0) is at least as good as QUINT (version 1.2).

Limitation

Although it seems like QUINT (version 2.0) is better than QUINT (version 1.2), the sample sizes currently used to study the effectiveness of QUINT (version 2.0) are rather small. Earlier studies have used sample sizes of 300 and 1000 (Sies & Van Mechelen, 2016; Van der Geest, 2017), whereas the present study uses sample sizes of 150 and 300. Using larger sample sizes could shed more light on the (acceptability of) the Type I error rate of QUINT (version 2.0).

Future research

Future research could expand the present study by adding an extra evaluation criterion. Sies and Van Mechelen (2016) used an evaluation criterion that takes into account the expected outcome that patients theoretically could have achieved when all patients receive their optimal treatment. To achieve this, the benefit of administering the treatments based on the decision trees over administering the overall best treatment is divided by the benefit of administrating each patient their optimal treatment over administering the overall best treatment. This criterion might be the most relevant criterion to the patient himself.

Another issue for future research is the method(s) used to compare QUINT to. MOB is a tree-based method, but not a method specifically designed to search for treatment-subgroup interactions. It would be appropriate to compare QUINT to another tree-based method looking for qualitative interactions, e.g., Interaction Trees (Su et al., 2009).

Conclusion

The simulation study shows that QUINT (version 2.0) has a lower Type II error rate than QUINT (version 1.2). The adaptation does not have a negative impact on the proportion good predicted and the Type I error rate. In addition, the application study shows that QUINT (version 2.0) is at least as competent as QUINT (version 1.2). Clearly, the adaptation of QUINT appears to be successful.

References

Albisser, A.M. (2000). The disease management equation. *IFAC Proceedings Volumes 33(3)*, 47-52.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* New York: Routledge.

Dusseldorp, E., Conversano, C., & Van Os, B. J. (2010). Combining an additive and tree-based regression model simultaneously: Stima. Journal of Computational and Graphical Statistics, 19(3), 514–530. doi:10.1198/jcgs.2010.06089

Dusseldorp, E., Doove, L., & Van Mechelen, I. (2015). Quint: An R package for identification of subgroups of clients who differ in which treatment alternative is best for them. *Behavior Research Methods*.

Dusseldorp, E., & Meulman, J. J. (2004). The regression trunk approach to discover treatment covariate interaction. Psychometrika, 69(3), 355–374.

Dusseldorp, E., & Van Mechelen, I. (2014). Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statistics in Medicine*, *33*(2), 219-237.

Epstein, R.S., & Sherwood, L.M. (1996). From outcomes research to disease management: a guide for the perplexed. *Ann Intern Med. 124(9)*, 832–837.

Formanoy, M. A., Dusseldorp, E., Coffeng, J. K., Van Mechelen, I., Boot, C. R., Hendriksen, I. J., & Tak, E. C. (2016). Physical activity and relaxation in the work setting to reduce the need for recovery: what works for whom? *BMC Public Health, 16(1)*, 866.

Foster, J., Taylor, J., & Ruberg, S. (2011). Subgroup identification from randomized clinical trial data. Statistics in Medicine, 30(24), 2867–2880.

Lipkovich, I., Dmitrienko, A., Denne, J., & Enas, G. (2011). Subgroup identification based on differential effect search–a recursive partitioning method for establishing response to treatment in patient subpopulations. Statistics in Medicine, 30(21), 2601–2621.

Peile, E. (2004). Reflections from medical practice: balancing evidence-based practice with practice-based evidence. In R. Pring & G. Thomas (Ed.), *Evidence-Based Practice in Education* (pp. 102-108). McGraw-Hill Education: London.

Sies, A., & Van Mechelen, I. (2016). Comparing four methods for estimating tree-based treatment regimes. Manuscript in preparation.

Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup analysis via recursive partitioning. The Journal of Machine Learning Research, 10, 141–158.

Van der Geest, M. (2017). *Simulation study QUINT vs. MOB*. Internal Leiden University Report: unpublished.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics, 17*, 492–514. doi: 10.1198/106186008X319331

Appendix A: R code quint2()

This appendix shows the code to search for a qualitative interaction tree. Red text is used to highlight code present in quint() but not in quint2() and green text is used to highlight code present in quint2() but not in quint().

```
quint2<- function(formula, data, control=NULL){
  #Dataformat without use of formula:
  #dat:data; first column in dataframe = the response variable
  #second column in dataframe = the dichotomous treatment vector
  #(coded with treatment A=1 and treatment B=2)
  #rest of the columns in dataframe are the predictors
  #maxl: maximum total number of leaves (terminal nodes) of the final tree
:
  #Lmax

  dat <- as.data.frame(data)
  if (missing(formula)) {
    y <- dat[, 1]
    tr <- dat[, 2]
    Xmat <- dat[, -c(1, 2)]
    dat <- na.omit(dat)
    if (length(levels(as.factor(tr))) != 2) {
      stop("Quint cannot be performed. The number of treatment conditions
            does not equal 2.")
    }
  } else {
    F1 <- Formula(formula)
    mf1 <- model.frame(F1, data = dat)
    y <- as.matrix(mf1[, 1])
    origtr <- as.factor(mf1[, 2])
    tr <- as.numeric(origtr)
    if (length(levels(origtr)) != 2) {
      stop("Quint cannot be performed. The number of treatment conditions
           does not equal 2.")
    }
    Xmat <- mf1[, 3:dim(mf1)[2]]
    dat <- cbind(y, tr, Xmat)
    dat <- na.omit(dat)
    cat("Treatment variable (T) equals 1 corresponds to",
        attr(F1, "rhs")[[1]], "=", levels(origtr)[1], "\n")
    cat("Treatment variable (T) equals 2 corresponds to",
        attr(F1, "rhs")[[1]], "=", levels(origtr)[2], "\n")
    names(dat)[1:2] <- names(mf1)[1:2]
  }
  cat("The sample size in the analysis is", dim(dat)[1], "\n")
```

```
  N<-length(y)
  if(is.null(control)) {
    control <- quint.control()  #Use default control parameters and criter
ion
  }

  #specify criterion , parameters a and b (parvec), weights and maximum
  #number of leaves:
  crit <- control$crit
  parvec <- control$parvec
  w <- control$w
  maxl <- control$maxl

  #if no control argument was specified ,use default parameter values
  #Default parameters a1 and a2 for treatment cardinality condition:
  if(is.null(parvec)){
    a1 <- round(sum(tr==1)/10)
    a2 <- round(sum(tr==2)/10)
    parvec <- c(a1, a2)
    control$parvec <- parvec
  }

  if(is.null(w)){
    #edif=expected mean difference between treatment and control; default
    #value for effect size criterion: edif = 3 (=Cohen's d),
    #and for difference in means criterion: edif= IQR(Y)
    edif <- ifelse(crit=="es", 3, IQR(y))
    w1 <- 1/log(1+edif)
    #bal= balance (ratio) between "difference in treatment outcomes
    #component" and "cardinality component"
    w2 <- 1/log(length(y)/2)
    w <- c(w1, w2)
    control$w <- w
  }

  ##Create matrix for results
  allresults <- matrix(0, nrow=maxl-1, ncol=6)
  splitpoints <- matrix(0, nrow=maxl-1, ncol=1)
  ## create a vector for true split points

  ##Start of the tree growing: all persons are in the rootnode. L=1;
  #Criterion value (cmax)=0
  root <- rep(1, length(y))
  cmax <- 0

  #Step 1
  #Generate design matrix D with admissable assignments after first split
  dmat1 <- matrix(c(1,2,2,1), nrow=2)

  #Select the optimal triplet for the first split: the triplet resulting i
n
  #the maximum value of the criterion (critmax1)
  #use the rootnode information: cardinality t=1, cardinality t=2, meant1,
```

```
  #meant0
  rootvec <- c(sum(tr==1), sum(tr==2), mean(y[tr==1])-mean(y[tr==2]))

  critmax1 <- bovar(y, Xmat, tr, gm=root, dmatsg=dmat1,
                    dmatsel=rep(1,nrow(dmat1)), parents=rootvec, parvec, w
,
                    nsplit=1, crit=crit)

  #Make the first split
  if(is.factor(Xmat[,critmax1[1]])==FALSE){
    Gmat <- makeGchmat(root, Xmat[,critmax1[1]], critmax1[2])  }
  if(is.factor(Xmat[,critmax1[1]])==TRUE){
    possibleSplits <- determineSplits(x=Xmat[,critmax1[1]], gm=root)
    assigMatrix <- makeCatmat(x=Xmat[,critmax1[1]], gm=root,
                              z=possibleSplits[[1]],
                              splits=possibleSplits[[2]])
    Gmat <- makeGchmatcat(gm=root, splitpoint=critmax1[2],
                          assigMatrix=assigMatrix)
    }

  cat("split 1","\n")
  cat("#leaves is 2","\n")

  ##Keep the child node numbers  nnum; #ncol(Gmat) is current number of
  ##leaves (=number of candidate parentnodes)=L;  #ncol(Gmat)+1 is total
  ##number of leaves after the split (Lafter)
  nnum <- c(2,3)
  L <- ncol(Gmat)

  ##Keep the results (split information, fit information, end node
  ##information) after the first split
  if(critmax1[4]!=0){
    allresults[1,] <- c(1,critmax1[-3])
    #Keep the splitpoints
    ifelse(is.factor(Xmat[,critmax1[1]])==F, splitpoints[1] <- critmax1[2]
,
           splitpoints[1] <- paste(as.vector(unique(
             sort(Xmat[Gmat[,1]==1, critmax1[1]]))), collapse=", "))
    dmatrow<-dmat1[critmax1[3],]
    cmax <- allresults[1, 4]
    endinf <- ctmat(Gmat,y,tr,crit=crit) ####changed
  } else { ##if there is no optimal triplet for the first split:
    Gmat <- Gmat*0
    dmatrow <- c(0,0)
    endinf <- matrix(0, ncol=8, nrow=2)
  }

  ##Check the qualitative interaction condition: Cohen's d in the leafs
  #after the first split >=dmin
  qualint <- "Present"
  if(abs(endinf[1,7])<control$dmin | abs(endinf[2,7])<control$dmin) {
    L <- maxl
    stop("The qualitative interaction condition is not satisfied: One or
```

```
        both of the effect sizes are lower than absolute value",
        control$dmin,". There is no clear qualitative interaction present
        in the data.","\n")
  }

  # Return an object of length 1 when C is 0
  if (cmax == 0) {
    object <- 1
    print("Quint cannot be performed. C is 0.")
    class(object) <- "quint"
    return(object)
  } else {

  ##Perform bias-corrected bootstrapping for the first split:
  if(control$Boot==TRUE&cmax!=0){
    #initiate bootstrap with stratification on treatment groups:
    indexboot <- Bootstrap(y, control$B, tr)
    critmax1boot <- matrix(0, ncol=6, nrow=control$B)

    #initialize matrices to keep results
    Gmattrain <- array(0, dim=c(N,maxl,control$B))
    Gmattest <- array(0, dim=c(N,maxl,control$B))
    allresultsboot <- array(0, dim=c(maxl-1,9,control$B))
    #find best first split for the k training sets
    for (b in 1:control$B) {
      cat("Bootstrap sample ",b,"\n")
      ##use the bootstrap data as training set
      critmax1boot[b,]<-
        bovar(y[indexboot[,b]],Xmat[indexboot[,b],],
              tr[indexboot[,b]],root,dmat1,rep(1,nrow(dmat1)),
              rootvec,parvec,w,1,crit=crit)
      if(is.factor(Xmat[,critmax1boot[b,1]])==FALSE){
        Gmattrain[,c(1:2),b]<-
          makeGchmat(gm=root, varx=Xmat[indexboot[,b],critmax1boot[b,1]],
                     splitpoint=critmax1boot[b,2])
        ##use the original data as testset
        Gmattest[,c(1:2),b]<-makeGchmat(gm=root,
                                        varx=Xmat[,critmax1boot[b,1]],
                                        splitpoint=critmax1boot[b,2])
      }

      if(is.factor(Xmat[,critmax1boot[b,1]])==TRUE){
        possibleSplits <-
          determineSplits(x=Xmat[indexboot[,b], critmax1boot[b,1]],
                          gm=root)
        assigMatrixTrain <-
          makeCatmat(x=Xmat[indexboot[,b], critmax1boot[b,1]], gm=root,
                     z=possibleSplits[[1]], splits=possibleSplits[[2]])
        Gmattrain[,c(1:2),b]<-
          makeGchmatcat(gm=root, splitpoint=critmax1boot[b,2],
                        assigMatrix=assigMatrixTrain)
        ##use the original data as testset
        assigMatrixTest <-
```

```
        makeCatmat(x=Xmat[,critmax1boot[b,1]], gm=root,
                   z=possibleSplits[[1]], splits=possibleSplits[[2]])
      Gmattest[,c(1:2),b]<-
        makeGchmatcat(gm=root, splitpoint=critmax1boot[b,2],
                      assigMatrix=assigMatrixTest)
    }

    End <- cpmat(Gmattest[,c(1:2),b], y, tr, crit=crit)
    #select the right row in the design matrix
    dmatsel <- t(dmat1[critmax1boot[b,3],])

    allresultsboot[1,c(1:8),b] <- c(1,critmax1boot[b,c(1:2)],
                                    computeCtest(End, dmatsel, w))
    allresultsboot[1,9,b] <- critmax1boot[b,4]-allresultsboot[1,4,b]
    if(critmax1boot[b,4]==0) {allresultsboot[1,,b]<-NA}
  }
}

#Repeat the tree growing procedure
stopc <- 0

while(L<maxl){
  cat("current value of C", cmax,"\n")
  cat("split", L, "\n")
  Lafter <- ncol(Gmat)+1
  cat("#leaves is", Lafter, "\n")
  ##make a designmatrix (dmat) for the admissible assignments of the
  #leaves after the split
  dmat <- makedmat(Lafter)
  dmatsg <- makedmats(dmat)
  #make parentnode information matrix, select best observed parent node
  #(with optimal triplet)
  parent <- cpmat(Gmat,y,tr,crit=crit)
  critmax <- bonode(Gmat,y,Xmat,tr,dmatrow,dmatsg,parent,parvec,w,L,
                    crit=crit)

  ##Perform the best split and keep results
  if(is.factor(Xmat[,critmax[2]])==FALSE){
    Gmatch <- makeGchmat(Gmat[,critmax[1]], Xmat[,critmax[2]],
                         critmax[3])
    }
  if(is.factor(Xmat[,critmax[2]])==TRUE){
    possibleSplits <- determineSplits(x=Xmat[,critmax[2]],
                                      gm=Gmat[,critmax[1]])
    assigMatrix <- makeCatmat(x=Xmat[,critmax[2]], gm=Gmat[,critmax[1]],
                              z=possibleSplits[[1]],
                              splits=possibleSplits[[2]])
    Gmatch <- makeGchmatcat(gm=Gmat[,critmax[1]], splitpoint=critmax[3],
                            assigMatrix=assigMatrix)
    }

  Gmatnew <- cbind(Gmat[,-critmax[1]], Gmatch)
  allresults[L,] <- c(nnum[critmax[1]], critmax[2:3], critmax[5:7])
```

```
    ifelse(is.factor(Xmat[,critmax[2]])==F,
           splitpoints[L] <- round(critmax[3], digits = 2),
           splitpoints[L] <-
             paste(as.vector(unique(sort(Xmat[Gmatch[,1]==1,critmax[2]])))
,
                   collapse=", "))
    dmatrownew <- dmatsg[critmax[4],]

    #check if cmax new is higher than current value
    if(allresults[L,4]<=cmax){
      cat("splitting process stopped after number of leaves equals",L,
          "because new value of C was not higher than current value of
          C","\n")
      stopc<-1
    }

    ##repeat this procedure for the bootstrap samples
    if(control$Boot==TRUE & stopc!=1){
      critmaxboot<-matrix(0,nrow=control$B,ncol=7)
      for (b in 1:control$B){
        cat("Bootstrap sample ",b,"\n")
        #make parentnode information matrix pmat
        parent <- cpmat(Gmattrain[,c(1:(Lafter-1)),b], y[indexboot[,b]],
                        tr[indexboot[,b]], crit=crit)
        critmaxboot[b,] <-
          bonode(Gmat=Gmattrain[,c(1:(Lafter-1)),b], y=y[indexboot[,b]],
                 Xmat=Xmat[indexboot[,b],], tr=tr[indexboot[,b]], dmatrow,
                 dmatsg, parent, parvec, w, nsplit=L, crit=crit)

        #best predictor and node of this split for the training samples
        if(is.factor(Xmat[,critmaxboot[b,2]])==FALSE){
          Gmattrainch <- makeGchmat(Gmattrain[, critmaxboot[b,1],b],
                                    Xmat[indexboot[,b], critmaxboot[b,2]],
                                    critmaxboot[b,3])
          Gmattestch <- makeGchmat(Gmattest[,critmaxboot[b,1],b],
                                   Xmat[, critmaxboot[b,2]],
                                   critmaxboot[b,3])
        }
        if(is.factor(Xmat[,critmaxboot[b,2]])==TRUE){
          possibleSplits <-
            determineSplits(x=Xmat[indexboot[,b], critmaxboot[b,2]],
                            gm=Gmattrain[,critmaxboot[b,1],b])
          assigMatrixTrain <-
            makeCatmat(x=Xmat[indexboot[,b], critmaxboot[b,2]],
                       gm=Gmattrain[,critmaxboot[b,1],b],
                       z=possibleSplits[[1]], splits=possibleSplits[[2]])
          Gmattrainch <- makeGchmatcat(gm=Gmattrain[,critmaxboot[b,1],b],
                                       splitpoint=critmaxboot[b,3],
                                       assigMatrix=assigMatrixTrain)

          assigMatrixTest <-
            makeCatmat(x=Xmat[,critmaxboot[b,2]],
                       gm=Gmattest[,critmaxboot[b,1],b],
```

```
                     z=possibleSplits[[1]],
                     splits=possibleSplits[[2]])
        Gmattestch <- makeGchmatcat(gm=Gmattest[,critmaxboot[b,1],b],
                                    splitpoint=critmaxboot[b,3],
                                    assigMatrix=assigMatrixTest)
      }

      Gmattrain[,c(1:Lafter),b] <-
        cbind(Gmattrain[,c(1:(Lafter-1))[-critmaxboot[b,1]],b],
              Gmattrainch)
      Gmattest[,c(1:Lafter),b] <-
        cbind(Gmattest[,c(1:(Lafter-1))[-critmaxboot[b,1]],b],
              Gmattestch)

      ##compute criterion value for the test sets
      End <- cpmat(Gmattest[,c(1:Lafter),b],y,tr,crit=crit)
      #select the right row in the design matrix
      if(critmaxboot[b,5]!=0){
        dmatsel<-t(dmatsg[critmaxboot[b,4],])
        allresultsboot[L,c(1:8),b] <-
          c(nnum[critmaxboot[b,1]],critmaxboot[b,2],critmaxboot[b,3],
            computeCtest(End, dmatsel, w))
        allresultsboot[L,9,b]<-critmaxboot[b,5]-allresultsboot[L,4,b]
      }
      if(critmaxboot[b,5]==0){
        allresultsboot[L,,b] <-NA
      }
    }

    if(sum(is.na(allresultsboot[L,9,]))/control$B > .10 ){
      warning("After split ",L,", the partitioning criterion cannot be
              computed in more than 10 percent of the bootstrap samples.
              The split is unstable." )
    }
  }

  #update the parameters after the split:
  if(stopc==0) {
    Gmat <- Gmatnew
    dmatrow <- dmatrownew
    cmax <- allresults[L,4]
    L <- ncol(Gmat)
    nnum <- c(nnum[-critmax[1]],nnum[critmax[1]]*2,nnum[critmax[1]]*2+1)
  } else {L <- maxl}

  #end of while loop
}

Lfinal <- ncol(Gmat)  #Lfinal=final number of leaves of the tree

#create endnode information of the tree
endinf <- matrix(0,nrow=length(nnum),ncol=10)
if(cmax!=0){
```

```
    endinf[,c(2:9)] <- ctmat(Gmat,y,tr,crit=crit)} ####changed
  endinf <- data.frame(endinf)
  endinf[,10] <- dmatrow
  endinf[,1] <- nnum
  index <- leafnum(nnum)
  endinf <- endinf[index,]
  rownames(endinf) <- paste("Leaf ",1:Lfinal,sep="")
  if(crit == 'es'){ ### this was added/changed
    colnames(endinf) <- c("node","#(T=1)", "meanY|T=1", "SD|T=1","#(T=2)",
                          "meanY|T=2","SD|T=2","d", "se", "class")}
  if(crit == 'dm'){ ### this was added
    colnames(endinf) <- c("node","#(T=1)", "meanY|T=1", "SD|T=1","#(T=2)",
                          "meanY|T=2","SD|T=2","diff", "se", "class")}
  if(Lfinal==2){allresults <- c(2,allresults[1,])}
  if(Lfinal>2){
    allresults <- cbind(2:Lfinal, allresults[1:(Lfinal-1),])
  }

  #compute final estimate of optimism and standard error:
  if(control$Boot==TRUE){
    #raw mean and sd:
    opt <- sapply(1:(Lfinal-1),
                  function(kk, allresultsboot){mean(allresultsboot[kk,9,],
                                                    na.rm=TRUE)},
                  allresultsboot=allresultsboot)
    se_opt <- sapply(1:(Lfinal-1),
                     function(kk,allresultsboot){sd(allresultsboot[kk,9,],
                                                    na.rm=TRUE) /
                         sqrt(sum(!is.na(allresultsboot[kk,9,])))},
                     allresultsboot=allresultsboot)

    if(Lfinal==2){allresults <- c(allresults[1:5], allresults[5]-opt,opt,
                                  se_opt, allresults[6:7])
    allresults <- data.frame(t(allresults))
    }
    if(Lfinal>2){
      allresults <- cbind(allresults[,1:5], allresults[,5]-opt,opt, se_opt
,
                          allresults[,6:7])
      allresults <- data.frame(allresults)
    }
    allresults[,3] <- colnames(Xmat)[allresults[,3]]
    splitnr <- 1:(Lfinal-1)
    allresults <- cbind(splitnr, allresults)
    colnames(allresults) <- c("split", "#leaves", "parentnode",
                              "splittingvar", "splitpoint", "apparent",
                              "biascorrected", "opt", "se","compdif",
                              "compcard")
  }

  if(control$Boot==FALSE){
    if(Lfinal>2){
      allresults <- data.frame( allresults)
```

```
    }
    if(Lfinal==2){
      allresults <- data.frame(t(allresults))
    }
    allresults[,3] <- colnames(Xmat)[allresults[,3]]
    splitnr <- 1:(Lfinal-1)
    allresults <- cbind(splitnr, allresults)
    colnames(allresults) <- c("split", "#leaves", "parentnode",
                              "splittingvar", "splitpoint",  "apparent",
                              "compdif","compcard")
  }
  colnames(Gmat) <- nnum

  ##split information (si): also include childnode numbers
  si <- allresults[,3:5]
  cn <- paste(si[,1]*2, si[,1]*2+1, sep=",")
  si <- cbind(parentnode=si[,1], childnodes=cn, si[,2:3],
              truesplitpoint=splitpoints[1:nrow(si)])
  rownames(si) <- paste("Split ", 1:(Lfinal-1), sep="")

  if(control$Boot==FALSE){
    object <- list(call=match.call(), crit=crit, control=control,
                   data=dat, si=si, fi=allresults[,c(1:2,6:8)], li=endinf,
                   nind=Gmat[,index])
  }
  if(control$Boot==TRUE){
    nam <- c("parentnode", "splittingvar", "splitpoint",
             "C_boot", "C_compdif", "checkdif", "C_compcard",
             "checkcard", "opt")
    dimnames(allresultsboot) <- list(NULL, nam, NULL)
    object <- list(call = match.call(), crit = crit, control = control,
                   indexboot = indexboot, data = dat, si = si,
                   fi = allresults[, c(1:2, 6:11)], li = endinf,
                   nind = Gmat[, index], siboot = allresultsboot)
  }
  class(object) <- "quint"
  return(object)
}
}
```

Appendix B: R code prune.quint2()

This appendix shows the code to prune the qualitative interaction tree. Green text is used to highlight code present in prune.quint2() but not in prune.quint().

```
prune.quint2 <- function(tree, pp=1,...){
  object <- tree
  if(length(object) == 1) {
    besttree <- 1
    class(besttree) <- "quint"
    return(besttree)
  } else {

    #pp=pruning parameter
    if(names(object$fi[4])=="Difcomponent"){
      stop("Pruning is not possible; The quint object lacks estimates of t
he
           biascorrected criterion. Grow again a large tree using the
           bootstrap procedure." )}

    object$fi[is.na(object$fi[,4]),4]<-0
    object$fi[is.na(object$fi[,5]),5]<-0
    maxrow<-which(object$fi[,4]==max(object$fi[,4]))[1]
    if(is.na(object$fi[maxrow,6])) maxrow <- maxrow - 1
    bestrow<-min( which(object$fi[,4]>=
                          (object$fi[maxrow,4]-pp*object$fi[maxrow,6]) ) )
    con<-object$control
    con$Boot<-FALSE
    con$maxl <- bestrow + 1
    besttree <- quint2(data = object$data, control = con)
    besttree$fi <- object$fi[1:bestrow, ]
    objboot <- list(siboot = object$siboot[1:bestrow, , ])
    besttree <- c(besttree, objboot)
    besttree$control$Boot <- object$control$Boot

    # Check whether there is a qualitative interaction
    if(colnames(besttree$li)[8]=="d"){  # criterium is es
      if((any(abs(subset(besttree$li, class == 1, d)) >= con$dmin) &
          any(abs(subset(besttree$li, class == 2, d)) >= con$dmin)) ==
         FALSE) {
        besttree <- 1
      }
    } else {  # criterium is dm
      if((any(abs(subset(besttree$li, class == 1, diff) /
                  sqrt(((besttree$li[besttree$li[,10]==1, 2] - 1) *
                        besttree$li[besttree$li[,10]==1, 4] ^ 2 +
                        (besttree$li[besttree$li[,10]==1, 5] - 1) *
                        besttree$li[besttree$li[,10]==1, 7] ^ 2) /
```

```
                        (sum(besttree$li[besttree$li[,10]==1, c(2, 5)]) -
                        2))) >= con$dmin) &
          any(abs(subset(besttree$li, class == 2, diff) /
                  sqrt(((besttree$li[besttree$li[,10]==2, 2] - 1) *
                        besttree$li[besttree$li[,10]==2, 4] ^ 2 +
                        (besttree$li[besttree$li[,10]==2, 5] - 1) *
                        besttree$li[besttree$li[,10]==2, 7] ^ 2) /
                       (sum(besttree$li[besttree$li[,10]==2, c(2, 5)]) -
                        2))) >= con$dmin)) == FALSE) {
        besttree <- 1
      }
    }

    class(besttree) <- "quint"
    return(besttree)
  }
  }
```

Appendix C: Example R code for MOB and QUINT

Example code MOB with Pima Indians diabetes data

```
# Load MOB
library(party)

# Load Pima Indians diabetes data
data(PimaIndiansDiabetes2, package = "mlbench")
PimaIndiansDiabetes <- na.omit(PimaIndiansDiabetes2[,-c(4, 5)])  # remove missing values

# Create formula with diabetes as outcome variable
fmPID <- mob(diabetes ~ glucose | pregnant + pressure + mass + pedigree + age,
        data = PimaIndiansDiabetes, model = glinearModel, family = binomial())

# Visualize the model
plot(fmPID)

# Show coefficients and corresponding odds ratios
coef(fmPID)
exp(coef(fmPID)[,2])
```

Example code QUINT with BCRP data

```
# Load QUINT
library(quint)

# Read data into memory
data(bcrp)
ex_data <- subset(bcrp, cond < 3)  # exclude the control condition

# Create formula with the change score in depression as outcome variable
formula1 <- I(cesdt1 - cesdt3) ~ cond | cesdt1 + negsoct1 + uncomt1 +
  disopt1 + comorbid + age + wcht1 + nationality + marital + trext
```

```
# Fix the seed
set.seed(47)

# Analysis with change score in depression as outcome variable
quint1 <- quint(formula1, data = ex_data)

# Give a summary of the analysis
summary(quint1)
quint1$fi
quint1$si
quint1$li

# Prune tree to avoid overfitting
quint1pr <- prune(quint1)

# Plot the pruned tree
plot(quint1pr)

# Round the leaf information of the pruned tree at two decimals
round(quint1pr$li, digits = 2)
```

Appendix D: Repeated measures ANOVA Proportion good predicted (excl. class 3)

This appendix shows the SPSS table of the within-subjects effects with the proportion good predicted as the dependent variable. Class 3 is considered as predicted incorrectly.

Table D1

Repeated Measures Analysis of Variance of Proportion good predicted with class 3 excluded (Within-Subjects Effects)

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
Method	Sphericity Assumed	14.417	1	14.417	982.795	.000	.094	.081
	Greenhouse-Geisser	14.417	1.000	14.417	982.795	.000	.094	.081
	Huynh-Feldt	14.417	1.000	14.417	982.795	.000	.094	.081
	Lower-bound	14.417	1.000	14.417	982.795	.000	.094	.081
Method * n	Sphericity Assumed	.621	1	.621	42.357	.000	.004	.004
	Greenhouse-Geisser	.621	1.000	.621	42.357	.000	.004	.004
	Huynh-Feldt	.621	1.000	.621	42.357	.000	.004	.004
	Lower-bound	.621	1.000	.621	42.357	.000	.004	.004
Method * J	Sphericity Assumed	3.700	1	3.700	252.201	.000	.026	.021
	Greenhouse-Geisser	3.700	1.000	3.700	252.201	.000	.026	.021
	Huynh-Feldt	3.700	1.000	3.700	252.201	.000	.026	.021
	Lower-bound	3.700	1.000	3.700	252.201	.000	.026	.021
Method * effect.size	Sphericity Assumed	4.356	1	4.356	296.938	.000	.030	.025
	Greenhouse-Geisser	4.356	1.000	4.356	296.938	.000	.030	.025
	Huynh-Feldt	4.356	1.000	4.356	296.938	.000	.030	.025
	Lower-bound	4.356	1.000	4.356	296.938	.000	.030	.025

Table D1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
Method * rho	Sphericity Assumed	.265	2	.133	9.049	.000	.002	.001
	Greenhouse-Geisser	.265	2.000	.133	9.049	.000	.002	.001
	Huynh-Feldt	.265	2.000	.133	9.049	.000	.002	.001
	Lower-bound	.265	2.000	.133	9.049	.000	.002	.001
Method * scenario	Sphericity Assumed	2.180	3	.727	49.527	.000	.015	.012
	Greenhouse-Geisser	2.180	3.000	.727	49.527	.000	.015	.012
	Huynh-Feldt	2.180	3.000	.727	49.527	.000	.015	.012
	Lower-bound	2.180	3.000	.727	49.527	.000	.015	.012
Method * n * J	Sphericity Assumed	.013	1	.013	.861	.353	.000	.000
	Greenhouse-Geisser	.013	1.000	.013	.861	.353	.000	.000
	Huynh-Feldt	.013	1.000	.013	.861	.353	.000	.000
	Lower-bound	.013	1.000	.013	.861	.353	.000	.000
Method * n * effect.size	Sphericity Assumed	.263	1	.263	17.957	.000	.002	.001
	Greenhouse-Geisser	.263	1.000	.263	17.957	.000	.002	.001
	Huynh-Feldt	.263	1.000	.263	17.957	.000	.002	.001
	Lower-bound	.263	1.000	.263	17.957	.000	.002	.001
Method * n * rho	Sphericity Assumed	.010	2	.005	.346	.707	.000	.000
	Greenhouse-Geisser	.010	2.000	.005	.346	.707	.000	.000
	Huynh-Feldt	.010	2.000	.005	.346	.707	.000	.000
	Lower-bound	.010	2.000	.005	.346	.707	.000	.000
Method * n * scenario	Sphericity Assumed	6.771	3	2.257	153.853	.000	.046	.038
	Greenhouse-Geisser	6.771	3.000	2.257	153.853	.000	.046	.038
	Huynh-Feldt	6.771	3.000	2.257	153.853	.000	.046	.038
	Lower-bound	6.771	3.000	2.257	153.853	.000	.046	.038

Table D1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
Method * J * effect.size	Sphericity Assumed	.675	1	.675	46.006	.000	.005	.004
	Greenhouse-Geisser	.675	1.000	.675	46.006	.000	.005	.004
	Huynh-Feldt	.675	1.000	.675	46.006	.000	.005	.004
	Lower-bound	.675	1.000	.675	46.006	.000	.005	.004
Method * J * rho	Sphericity Assumed	.178	2	.089	6.084	.002	.001	.001
	Greenhouse-Geisser	.178	2.000	.089	6.084	.002	.001	.001
	Huynh-Feldt	.178	2.000	.089	6.084	.002	.001	.001
	Lower-bound	.178	2.000	.089	6.084	.002	.001	.001
Method * J * scenario	Sphericity Assumed	.042	3	.014	.954	.413	.000	.000
	Greenhouse-Geisser	.042	3.000	.014	.954	.413	.000	.000
	Huynh-Feldt	.042	3.000	.014	.954	.413	.000	.000
	Lower-bound	.042	3.000	.014	.954	.413	.000	.000
Method * effect.size * rho	Sphericity Assumed	.061	2	.031	2.089	.124	.000	.000
	Greenhouse-Geisser	.061	2.000	.031	2.089	.124	.000	.000
	Huynh-Feldt	.061	2.000	.031	2.089	.124	.000	.000
	Lower-bound	.061	2.000	.031	2.089	.124	.000	.000
Method * effect.size * scenario	Sphericity Assumed	2.006	3	.669	45.573	.000	.014	.011
	Greenhouse-Geisser	2.006	3.000	.669	45.573	.000	.014	.011
	Huynh-Feldt	2.006	3.000	.669	45.573	.000	.014	.011
	Lower-bound	2.006	3.000	.669	45.573	.000	.014	.011
Method * rho * scenario	Sphericity Assumed	.337	6	.056	3.829	.001	.002	.002
	Greenhouse-Geisser	.337	6.000	.056	3.829	.001	.002	.002
	Huynh-Feldt	.337	6.000	.056	3.829	.001	.002	.002
	Lower-bound	.337	6.000	.056	3.829	.001	.002	.002

Table D1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
Method * n * J * effect.size	Sphericity Assumed	.090	1	.090	6.112	.013	.001	.001
	Greenhouse-Geisser	.090	1.000	.090	6.112	.013	.001	.001
	Huynh-Feldt	.090	1.000	.090	6.112	.013	.001	.001
	Lower-bound	.090	1.000	.090	6.112	.013	.001	.001
Method * n * J * rho	Sphericity Assumed	.031	2	.015	1.056	.348	.000	.000
	Greenhouse-Geisser	.031	2.000	.015	1.056	.348	.000	.000
	Huynh-Feldt	.031	2.000	.015	1.056	.348	.000	.000
	Lower-bound	.031	2.000	.015	1.056	.348	.000	.000
Method * n * J * scenario	Sphericity Assumed	.549	3	.183	12.476	.000	.004	.003
	Greenhouse-Geisser	.549	3.000	.183	12.476	.000	.004	.003
	Huynh-Feldt	.549	3.000	.183	12.476	.000	.004	.003
	Lower-bound	.549	3.000	.183	12.476	.000	.004	.003
Method * n * effect.size * rho	Sphericity Assumed	.020	2	.010	.671	.511	.000	.000
	Greenhouse-Geisser	.020	2.000	.010	.671	.511	.000	.000
	Huynh-Feldt	.020	2.000	.010	.671	.511	.000	.000
	Lower-bound	.020	2.000	.010	.671	.511	.000	.000
Method * n * effect.size * scenario	Sphericity Assumed	.471	3	.157	10.695	.000	.003	.003
	Greenhouse-Geisser	.471	3.000	.157	10.695	.000	.003	.003
	Huynh-Feldt	.471	3.000	.157	10.695	.000	.003	.003
	Lower-bound	.471	3.000	.157	10.695	.000	.003	.003
Method * n * rho * scenario	Sphericity Assumed	.110	6	.018	1.246	.279	.001	.001
	Greenhouse-Geisser	.110	6.000	.018	1.246	.279	.001	.001
	Huynh-Feldt	.110	6.000	.018	1.246	.279	.001	.001
	Lower-bound	.110	6.000	.018	1.246	.279	.001	.001

Table D1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
Method * J * effect.size * rho	Sphericity Assumed	.018	2	.009	.600	.549	.000	.000
	Greenhouse-Geisser	.018	2.000	.009	.600	.549	.000	.000
	Huynh-Feldt	.018	2.000	.009	.600	.549	.000	.000
	Lower-bound	.018	2.000	.009	.600	.549	.000	.000
Method * J * effect.size * scenario	Sphericity Assumed	.091	3	.030	2.067	.102	.001	.001
	Greenhouse-Geisser	.091	3.000	.030	2.067	.102	.001	.001
	Huynh-Feldt	.091	3.000	.030	2.067	.102	.001	.001
	Lower-bound	.091	3.000	.030	2.067	.102	.001	.001
Method * J * rho * scenario	Sphericity Assumed	.022	6	.004	.255	.958	.000	.000
	Greenhouse-Geisser	.022	6.000	.004	.255	.958	.000	.000
	Huynh-Feldt	.022	6.000	.004	.255	.958	.000	.000
	Lower-bound	.022	6.000	.004	.255	.958	.000	.000
Method * effect.size * rho * scenario	Sphericity Assumed	.141	6	.024	1.602	.142	.001	.001
	Greenhouse-Geisser	.141	6.000	.024	1.602	.142	.001	.001
	Huynh-Feldt	.141	6.000	.024	1.602	.142	.001	.001
	Lower-bound	.141	6.000	.024	1.602	.142	.001	.001
Method * n * J * effect.size * rho	Sphericity Assumed	.172	2	.086	5.866	.003	.001	.001
	Greenhouse-Geisser	.172	2.000	.086	5.866	.003	.001	.001
	Huynh-Feldt	.172	2.000	.086	5.866	.003	.001	.001
	Lower-bound	.172	2.000	.086	5.866	.003	.001	.001
Method * n * J * effect.size * scenario	Sphericity Assumed	.057	3	.019	1.286	.277	.000	.000
	Greenhouse-Geisser	.057	3.000	.019	1.286	.277	.000	.000
	Huynh-Feldt	.057	3.000	.019	1.286	.277	.000	.000
	Lower-bound	.057	3.000	.019	1.286	.277	.000	.000

Table D1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
Method * n * J * rho * scenario	Sphericity Assumed	.030	6	.005	.336	.918	.000	.000
	Greenhouse-Geisser	.030	6.000	.005	.336	.918	.000	.000
	Huynh-Feldt	.030	6.000	.005	.336	.918	.000	.000
	Lower-bound	.030	6.000	.005	.336	.918	.000	.000
Method * n * effect.size * rho * scenario	Sphericity Assumed	.160	6	.027	1.821	.091	.001	.001
	Greenhouse-Geisser	.160	6.000	.027	1.821	.091	.001	.001
	Huynh-Feldt	.160	6.000	.027	1.821	.091	.001	.001
	Lower-bound	.160	6.000	.027	1.821	.091	.001	.001
Method * J * effect.size * rho * scenario	Sphericity Assumed	.193	6	.032	2.192	.041	.001	.001
	Greenhouse-Geisser	.193	6.000	.032	2.192	.041	.001	.001
	Huynh-Feldt	.193	6.000	.032	2.192	.041	.001	.001
	Lower-bound	.193	6.000	.032	2.192	.041	.001	.001
Method * n * J * effect.size * rho * scenario	Sphericity Assumed	.031	6	.005	.353	.909	.000	.000
	Greenhouse-Geisser	.031	6.000	.005	.353	.909	.000	.000
	Huynh-Feldt	.031	6.000	.005	.353	.909	.000	.000
	Lower-bound	.031	6.000	.005	.353	.909	.000	.000
Error(Method)	Sphericity Assumed	139.414	9504	.015				.785
	Greenhouse-Geisser	139.414	9504.000	.015				.785
	Huynh-Feldt	139.414	9504.000	.015				.785
	Lower-bound	139.414	9504.000	.015				.785

Appendix E: Repeated measures ANOVA Proportion good predicted (incl. class 3)

This appendix shows the SPSS table of the within-subjects effects with the proportion good predicted as the dependent variable. Class 3 is considered as predicted correctly.

Table E1

Repeated Measures Analysis of Variance of Proportion good predicted with class 3 included (Within-Subjects Effects)

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
Method	Sphericity Assumed	.491	1	.491	53.899	.000	.006	.004
	Greenhouse-Geisser	.491	1.000	.491	53.899	.000	.006	.004
	Huynh-Feldt	.491	1.000	.491	53.899	.000	.006	.004
	Lower-bound	.491	1.000	.491	53.899	.000	.006	.004
Method * n	Sphericity Assumed	.813	1	.813	89.125	.000	.009	.007
	Greenhouse-Geisser	.813	1.000	.813	89.125	.000	.009	.007
	Huynh-Feldt	.813	1.000	.813	89.125	.000	.009	.007
	Lower-bound	.813	1.000	.813	89.125	.000	.009	.007
Method * J	Sphericity Assumed	.334	1	.334	36.587	.000	.004	.003
	Greenhouse-Geisser	.334	1.000	.334	36.587	.000	.004	.003
	Huynh-Feldt	.334	1.000	.334	36.587	.000	.004	.003
	Lower-bound	.334	1.000	.334	36.587	.000	.004	.003
Method * effect.size	Sphericity Assumed	2.013	1	2.013	220.774	.000	.023	.017
	Greenhouse-Geisser	2.013	1.000	2.013	220.774	.000	.023	.017
	Huynh-Feldt	2.013	1.000	2.013	220.774	.000	.023	.017
	Lower-bound	2.013	1.000	2.013	220.774	.000	.023	.017

Table E1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
Method * rho	Sphericity Assumed	.052	2	.026	2.842	.058	.001	.000
	Greenhouse-Geisser	.052	2.000	.026	2.842	.058	.001	.000
	Huynh-Feldt	.052	2.000	.026	2.842	.058	.001	.000
	Lower-bound	.052	2.000	.026	2.842	.058	.001	.000
Method * scenario	Sphericity Assumed	15.214	3	5.071	556.162	.000	.149	.130
	Greenhouse-Geisser	15.214	3.000	5.071	556.162	.000	.149	.130
	Huynh-Feldt	15.214	3.000	5.071	556.162	.000	.149	.130
	Lower-bound	15.214	3.000	5.071	556.162	.000	.149	.130
Method * n * J	Sphericity Assumed	.007	1	.007	.802	.371	.000	.000
	Greenhouse-Geisser	.007	1.000	.007	.802	.371	.000	.000
	Huynh-Feldt	.007	1.000	.007	.802	.371	.000	.000
	Lower-bound	.007	1.000	.007	.802	.371	.000	.000
Method * n * effect.size	Sphericity Assumed	.367	1	.367	40.239	.000	.004	.003
	Greenhouse-Geisser	.367	1.000	.367	40.239	.000	.004	.003
	Huynh-Feldt	.367	1.000	.367	40.239	.000	.004	.003
	Lower-bound	.367	1.000	.367	40.239	.000	.004	.003
Method * n * rho	Sphericity Assumed	.124	2	.062	6.821	.001	.001	.001
	Greenhouse-Geisser	.124	2.000	.062	6.821	.001	.001	.001
	Huynh-Feldt	.124	2.000	.062	6.821	.001	.001	.001
	Lower-bound	.124	2.000	.062	6.821	.001	.001	.001
Method * n * scenario	Sphericity Assumed	6.496	3	2.165	237.456	.000	.070	.056
	Greenhouse-Geisser	6.496	3.000	2.165	237.456	.000	.070	.056
	Huynh-Feldt	6.496	3.000	2.165	237.456	.000	.070	.056
	Lower-bound	6.496	3.000	2.165	237.456	.000	.070	.056

Table E1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
Method * J * effect.size	Sphericity Assumed	.046	1	.046	5.023	.025	.001	.000
	Greenhouse-Geisser	.046	1.000	.046	5.023	.025	.001	.000
	Huynh-Feldt	.046	1.000	.046	5.023	.025	.001	.000
	Lower-bound	.046	1.000	.046	5.023	.025	.001	.000
Method * J * rho	Sphericity Assumed	.085	2	.042	4.657	.010	.001	.001
	Greenhouse-Geisser	.085	2.000	.042	4.657	.010	.001	.001
	Huynh-Feldt	.085	2.000	.042	4.657	.010	.001	.001
	Lower-bound	.085	2.000	.042	4.657	.010	.001	.001
Method * J * scenario	Sphericity Assumed	.424	3	.141	15.506	.000	.005	.004
	Greenhouse-Geisser	.424	3.000	.141	15.506	.000	.005	.004
	Huynh-Feldt	.424	3.000	.141	15.506	.000	.005	.004
	Lower-bound	.424	3.000	.141	15.506	.000	.005	.004
Method * effect.size * rho	Sphericity Assumed	.022	2	.011	1.229	.293	.000	.000
	Greenhouse-Geisser	.022	2.000	.011	1.229	.293	.000	.000
	Huynh-Feldt	.022	2.000	.011	1.229	.293	.000	.000
	Lower-bound	.022	2.000	.011	1.229	.293	.000	.000
Method * effect.size * scenario	Sphericity Assumed	1.326	3	.442	48.465	.000	.015	.011
	Greenhouse-Geisser	1.326	3.000	.442	48.465	.000	.015	.011
	Huynh-Feldt	1.326	3.000	.442	48.465	.000	.015	.011
	Lower-bound	1.326	3.000	.442	48.465	.000	.015	.011
Method * rho * scenario	Sphericity Assumed	.684	6	.114	12.499	.000	.008	.006
	Greenhouse-Geisser	.684	6.000	.114	12.499	.000	.008	.006
	Huynh-Feldt	.684	6.000	.114	12.499	.000	.008	.006
	Lower-bound	.684	6.000	.114	12.499	.000	.008	.006

Table E1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
Method * n * J * effect.size	Sphericity Assumed	.191	1	.191	20.893	.000	.002	.001
	Greenhouse-Geisser	.191	1.000	.191	20.893	.000	.002	.001
	Huynh-Feldt	.191	1.000	.191	20.893	.000	.002	.001
	Lower-bound	.191	1.000	.191	20.893	.000	.002	.001
Method * n * J * rho	Sphericity Assumed	.005	2	.002	.248	.780	.000	.000
	Greenhouse-Geisser	.005	2.000	.002	.248	.780	.000	.000
	Huynh-Feldt	.005	2.000	.002	.248	.780	.000	.000
	Lower-bound	.005	2.000	.002	.248	.780	.000	.000
Method * n * J * scenario	Sphericity Assumed	.468	3	.156	17.105	.000	.005	.004
	Greenhouse-Geisser	.468	3.000	.156	17.105	.000	.005	.004
	Huynh-Feldt	.468	3.000	.156	17.105	.000	.005	.004
	Lower-bound	.468	3.000	.156	17.105	.000	.005	.004
Method * n * effect.size * rho	Sphericity Assumed	.060	2	.030	3.293	.037	.001	.001
	Greenhouse-Geisser	.060	2.000	.030	3.293	.037	.001	.001
	Huynh-Feldt	.060	2.000	.030	3.293	.037	.001	.001
	Lower-bound	.060	2.000	.030	3.293	.037	.001	.001
Method * n * effect.size * scenario	Sphericity Assumed	.139	3	.046	5.089	.002	.002	.001
	Greenhouse-Geisser	.139	3.000	.046	5.089	.002	.002	.001
	Huynh-Feldt	.139	3.000	.046	5.089	.002	.002	.001
	Lower-bound	.139	3.000	.046	5.089	.002	.002	.001
Method * n * rho * scenario	Sphericity Assumed	.101	6	.017	1.854	.085	.001	.001
	Greenhouse-Geisser	.101	6.000	.017	1.854	.085	.001	.001
	Huynh-Feldt	.101	6.000	.017	1.854	.085	.001	.001
	Lower-bound	.101	6.000	.017	1.854	.085	.001	.001

Table E1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
Method * J * effect.size * rho	Sphericity Assumed	.026	2	.013	1.403	.246	.000	.000
	Greenhouse-Geisser	.026	2.000	.013	1.403	.246	.000	.000
	Huynh-Feldt	.026	2.000	.013	1.403	.246	.000	.000
	Lower-bound	.026	2.000	.013	1.403	.246	.000	.000
Method * J * effect.size * scenario	Sphericity Assumed	.123	3	.041	4.481	.004	.001	.001
	Greenhouse-Geisser	.123	3.000	.041	4.481	.004	.001	.001
	Huynh-Feldt	.123	3.000	.041	4.481	.004	.001	.001
	Lower-bound	.123	3.000	.041	4.481	.004	.001	.001
Method * J * rho * scenario	Sphericity Assumed	.062	6	.010	1.137	.338	.001	.001
	Greenhouse-Geisser	.062	6.000	.010	1.137	.338	.001	.001
	Huynh-Feldt	.062	6.000	.010	1.137	.338	.001	.001
	Lower-bound	.062	6.000	.010	1.137	.338	.001	.001
Method * effect.size * rho * scenario	Sphericity Assumed	.050	6	.008	.908	.488	.001	.000
	Greenhouse-Geisser	.050	6.000	.008	.908	.488	.001	.000
	Huynh-Feldt	.050	6.000	.008	.908	.488	.001	.000
	Lower-bound	.050	6.000	.008	.908	.488	.001	.000
Method * n * J * effect.size * rho	Sphericity Assumed	.044	2	.022	2.418	.089	.001	.000
	Greenhouse-Geisser	.044	2.000	.022	2.418	.089	.001	.000
	Huynh-Feldt	.044	2.000	.022	2.418	.089	.001	.000
	Lower-bound	.044	2.000	.022	2.418	.089	.001	.000
Method * n * J * effect.size * scenario	Sphericity Assumed	.049	3	.016	1.796	.146	.001	.000
	Greenhouse-Geisser	.049	3.000	.016	1.796	.146	.001	.000
	Huynh-Feldt	.049	3.000	.016	1.796	.146	.001	.000
	Lower-bound	.049	3.000	.016	1.796	.146	.001	.000

Table E1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
Method * n * J * rho * scenario	Sphericity Assumed	.041	6	.007	.752	.607	.000	.000
	Greenhouse-Geisser	.041	6.000	.007	.752	.607	.000	.000
	Huynh-Feldt	.041	6.000	.007	.752	.607	.000	.000
	Lower-bound	.041	6.000	.007	.752	.607	.000	.000
Method * n * effect.size * rho * scenario	Sphericity Assumed	.048	6	.008	.879	.509	.001	.000
	Greenhouse-Geisser	.048	6.000	.008	.879	.509	.001	.000
	Huynh-Feldt	.048	6.000	.008	.879	.509	.001	.000
	Lower-bound	.048	6.000	.008	.879	.509	.001	.000
Method * J * effect.size * rho * scenario	Sphericity Assumed	.129	6	.022	2.365	.028	.001	.001
	Greenhouse-Geisser	.129	6.000	.022	2.365	.028	.001	.001
	Huynh-Feldt	.129	6.000	.022	2.365	.028	.001	.001
	Lower-bound	.129	6.000	.022	2.365	.028	.001	.001
Method * n * J * effect.size * rho * scenario	Sphericity Assumed	.046	6	.008	.841	.538	.001	.000
	Greenhouse-Geisser	.046	6.000	.008	.841	.538	.001	.000
	Huynh-Feldt	.046	6.000	.008	.841	.538	.001	.000
	Lower-bound	.046	6.000	.008	.841	.538	.001	.000
Error(Method)	Sphericity Assumed	86.661	9504	.009				.742
	Greenhouse-Geisser	86.661	9504.000	.009				.742
	Huynh-Feldt	86.661	9504.000	.009				.742
	Lower-bound	86.661	9504.000	.009				.742

Appendix F: Repeated measures ANOVA Type I error rate

This appendix shows the SPSS table of the within-subjects effects with the Type I error rate as the dependent variable.

Table F1

Repeated Measures Analysis of Variance of Type I error rate (Within-Subjects Effects)

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
tree_returned	Sphericity Assumed	84.801	1	84.801	602.622	.000	.202	.143
	Greenhouse-Geisser	84.801	1.000	84.801	602.622	.000	.202	.143
	Huynh-Feldt	84.801	1.000	84.801	602.622	.000	.202	.143
	Lower-bound	84.801	1.000	84.801	602.622	.000	.202	.143
tree_returned * n	Sphericity Assumed	37.101	1	37.101	263.651	.000	.100	.063
	Greenhouse-Geisser	37.101	1.000	37.101	263.651	.000	.100	.063
	Huynh-Feldt	37.101	1.000	37.101	263.651	.000	.100	.063
	Lower-bound	37.101	1.000	37.101	263.651	.000	.100	.063
tree_returned * J	Sphericity Assumed	50.430	1	50.430	358.372	.000	.131	.085
	Greenhouse-Geisser	50.430	1.000	50.430	358.372	.000	.131	085
	Huynh-Feldt	50.430	1.000	50.430	358.372	.000	.131	085
	Lower-bound	50.430	1.000	50.430	358.372	.000	.131	085
tree_returned * effect.size	Sphericity Assumed	62.563	1	62.563	444.595	.000	.158	.106
	Greenhouse-Geisser	62.563	1.000	62.563	444.595	.000	.158	.106
	Huynh-Feldt	62.563	1.000	62.563	444.595	.000	.158	.106
	Lower-bound	62.563	1.000	62.563	444.595	.000	.158	.106

Table F1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
tree_returned * rho	Sphericity Assumed	7.152	2	3.576	25.411	.000	.021	.012
	Greenhouse-Geisser	7.152	2.000	3.576	25.411	.000	.021	.012
	Huynh-Feldt	7.152	2.000	3.576	25.411	.000	.021	.012
	Lower-bound	7.152	2.000	3.576	25.411	.000	.021	.012
tree_returned * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000
tree_returned * n * J	Sphericity Assumed	.480	1	.480	3.411	.065	.001	.001
	Greenhouse-Geisser	.480	1.000	.480	3.411	.065	.001	.001
	Huynh-Feldt	.480	1.000	.480	3.411	.065	.001	.001
	Lower-bound	.480	1.000	.480	3.411	.065	.001	.001
tree_returned * n * effect.size	Sphericity Assumed	.120	1	.120	.853	.356	.000	.000
	Greenhouse-Geisser	.120	1.000	.120	.853	.356	.000	.000
	Huynh-Feldt	.120	1.000	.120	.853	.356	.000	.000
	Lower-bound	.120	1.000	.120	.853	.356	.000	.000
tree_returned * n * rho	Sphericity Assumed	2.252	2	1.126	8.001	.000	.007	.004
	Greenhouse-Geisser	2.252	2.000	1.126	8.001	.000	.007	.004
	Huynh-Feldt	2.252	2.000	1.126	8.001	.000	.007	.004
	Lower-bound	2.252	2.000	1.126	8.001	.000	.007	.004
tree_returned * n * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000

Table F1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
tree_returned * J * effect.size	Sphericity Assumed	.101	1	.101	.717	.397	.000	.000
	Greenhouse-Geisser	.101	1.000	.101	.717	.397	.000	.000
	Huynh-Feldt	.101	1.000	.101	.717	.397	.000	.000
	Lower-bound	.101	1.000	.101	.717	.397	.000	.000
tree_returned * J * rho	Sphericity Assumed	3.705	2	1.852	13.164	.000	.011	.006
	Greenhouse-Geisser	3.705	2.000	1.852	13.164	.000	.011	.006
	Huynh-Feldt	3.705	2.000	1.852	13.164	.000	.011	.006
	Lower-bound	3.705	2.000	1.852	13.164	.000	.011	.006
tree_returned * J * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000
tree_returned * effect.size * rho	Sphericity Assumed	3.307	2	1.653	11.749	.000	.010	.006
	Greenhouse-Geisser	3.307	2.000	1.653	11.749	.000	.010	.006
	Huynh-Feldt	3.307	2.000	1.653	11.749	.000	.010	.006
	Lower-bound	3.307	2.000	1.653	11.749	.000	.010	.006
tree_returned * effect.size * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000
tree_returned * rho * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000

Table F1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
tree_returned * n * J * effect.size	Sphericity Assumed	.521	1	.521	3.701	.054	.002	.001
	Greenhouse-Geisser	.521	1.000	.521	3.701	.054	.002	.001
	Huynh-Feldt	.521	1.000	.521	3.701	.054	.002	.001
	Lower-bound	.521	1.000	.521	3.701	.054	.002	.001
tree_returned * n * J * rho	Sphericity Assumed	.945	2	.472	3.358	.035	.003	.002
	Greenhouse-Geisser	.945	2.000	.472	3.358	.035	.003	.002
	Huynh-Feldt	.945	2.000	.472	3.358	.035	.003	.002
	Lower-bound	.945	2.000	.472	3.358	.035	.003	.002
tree_returned * n * J * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000
tree_returned * n * effect.size * rho	Sphericity Assumed	1.820	2	.910	6.467	.002	.005	.003
	Greenhouse-Geisser	1.820	2.000	.910	6.467	.002	.005	.003
	Huynh-Feldt	1.820	2.000	.910	6.467	.002	.005	.003
	Lower-bound	1.820	2.000	.910	6.467	.002	.005	.003
tree_returned * n * effect.size * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000
tree_returned * n * rho * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000

Table F1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
tree_returned * J * effect.size * rho	Sphericity Assumed	1.307	2	.653	4.643	.010	.004	.002
	Greenhouse-Geisser	1.307	2.000	.653	4.643	.010	.004	.002
	Huynh-Feldt	1.307	2.000	.653	4.643	.010	.004	.002
	Lower-bound	1.307	2.000	.653	4.643	.010	.004	.002
tree_returned * J * effect.size * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000
tree_returned * J * rho * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000
tree_returned * effect.size * rho * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000
tree_returned * n * J * effect.size * rho	Sphericity Assumed	1.047	2	.523	3.719	.024	.003	.002
	Greenhouse-Geisser	1.047	2.000	.523	3.719	.024	.003	.002
	Huynh-Feldt	1.047	2.000	.523	3.719	.024	.003	.002
	Lower-bound	1.047	2.000	.523	3.719	.024	.003	.002
tree_returned * n * J * effect.size * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000

Table F1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squared
tree_returned * n * J * rho * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000
tree_returned * n * effect.size * rho * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000
tree_returned * J * effect.size * rho * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000
tree_returned * n * J * effect.size * rho * scenario	Sphericity Assumed	.000	0	.	.	.	.000	.000
	Greenhouse-Geisser	.000	.000	.	.	.	.000	.000
	Huynh-Feldt	.000	.000	.	.	.	.000	.000
	Lower-bound	.000	.000	.	.	.	.000	.000
Error(tree_returned)	Sphericity Assumed	334.350	2376	.141				.565
	Greenhouse-Geisser	334.350	2376.000	.141				.565
	Huynh-Feldt	334.350	2376.000	.141				.565
	Lower-bound	334.350	2376.000	.141				.565

Appendix G: Repeated measures ANOVA Type II error rate

This appendix shows the SPSS table of the within-subjects effects with the Type II error rate as the dependent variable.

Table G1

Repeated Measures Analysis of Variance of Type II error rate (Within-Subjects Effects)

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squar-ed
typeII_error	Sphericity Assumed	4.340	1	4.340	33.868	.000	.005	.004
	Greenhouse-Geisser	4.340	1.000	4.340	33.868	.000	.005	.004
	Huynh-Feldt	4.340	1.000	4.340	33.868	.000	.005	.004
	Lower-bound	4.340	1.000	4.340	33.868	.000	.005	.004
typeII_error * n	Sphericity Assumed	146.814	1	146.814	1145.618	.000	.138	.122
	Greenhouse-Geisser	146.814	1.000	146.814	1145.618	.000	.138	.122
	Huynh-Feldt	146.814	1.000	146.814	1145.618	.000	.138	.122
	Lower-bound	146.814	1.000	146.814	1145.618	.000	.138	.122
typeII_error * J	Sphericity Assumed	26.694	1	26.694	208.302	.000	.028	.022
	Greenhouse-Geisser	26.694	1.000	26.694	208.302	.000	.028	.022
	Huynh-Feldt	26.694	1.000	26.694	208.302	.000	.028	.022
	Lower-bound	26.694	1.000	26.694	208.302	.000	.028	.022
typeII_error * effect.size	Sphericity Assumed	8.703	1	8.703	67.907	.000	.009	.007
	Greenhouse-Geisser	8.703	1.000	8.703	67.907	.000	.009	.007
	Huynh-Feldt	8.703	1.000	8.703	67.907	.000	.009	.007
	Lower-bound	8.703	1.000	8.703	67.907	.000	.009	.007

Table G1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squar -ed
typeII_error * rho	Sphericity Assumed	.205	2	.103	.800	.449	.000	.000
	Greenhouse-Geisser	.205	2.000	.103	.800	.449	.000	.000
	Huynh-Feldt	.205	2.000	.103	.800	.449	.000	.000
	Lower-bound	.205	2.000	.103	.800	.449	.000	.000
typeII_error * scenario	Sphericity Assumed	19.543	2	9.772	76.251	.000	.021	.016
	Greenhouse-Geisser	19.543	2.000	9.772	76.251	.000	.021	.016
	Huynh-Feldt	19.543	2.000	9.772	76.251	.000	.021	.016
	Lower-bound	19.543	2.000	9.772	76.251	.000	.021	.016
typeII_error * n * J	Sphericity Assumed	6.084	1	6.084	47.478	.000	.007	.005
	Greenhouse-Geisser	6.084	1.000	6.084	47.478	.000	.007	.005
	Huynh-Feldt	6.084	1.000	6.084	47.478	.000	.007	.005
	Lower-bound	6.084	1.000	6.084	47.478	.000	.007	.005
typeII_error * n * effect.size	Sphericity Assumed	20.702	1	20.702	161.546	.000	.022	.017
	Greenhouse-Geisser	20.702	1.000	20.702	161.546	.000	.022	.017
	Huynh-Feldt	20.702	1.000	20.702	161.546	.000	.022	.017
	Lower-bound	20.702	1.000	20.702	161.546	.000	.022	.017
typeII_error * n * rho	Sphericity Assumed	1.503	2	.751	5.863	.003	.002	.001
	Greenhouse-Geisser	1.503	2.000	.751	5.863	.003	.002	.001
	Huynh-Feldt	1.503	2.000	.751	5.863	.003	.002	.001
	Lower-bound	1.503	2.000	.751	5.863	.003	.002	.001
typeII_error * n * scenario	Sphericity Assumed	3.810	2	1.905	14.866	.000	.004	.003
	Greenhouse-Geisser	3.810	2.000	1.905	14.866	.000	.004	.003
	Huynh-Feldt	3.810	2.000	1.905	14.866	.000	.004	.003
	Lower-bound	3.810	2.000	1.905	14.866	.000	.004	.003

Table G1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squar-ed
typeII_error * J * effect.size	Sphericity Assumed	.444	1	.444	3.468	.063	.000	.000
	Greenhouse-Geisser	.444	1.000	.444	3.468	.063	.000	.000
	Huynh-Feldt	.444	1.000	.444	3.468	.063	.000	.000
	Lower-bound	.444	1.000	.444	3.468	.063	.000	.000
typeII_error * J * rho	Sphericity Assumed	.676	2	.338	2.637	.072	.001	.001
	Greenhouse-Geisser	.676	2.000	.338	2.637	.072	.001	.001
	Huynh-Feldt	.676	2.000	.338	2.637	.072	.001	.001
	Lower-bound	.676	2.000	.338	2.637	.072	.001	.001
typeII_error * J * scenario	Sphericity Assumed	6.393	2	3.196	24.942	.000	.007	.005
	Greenhouse-Geisser	6.393	2.000	3.196	24.942	.000	.007	.005
	Huynh-Feldt	6.393	2.000	3.196	24.942	.000	.007	.005
	Lower-bound	6.393	2.000	3.196	24.942	.000	.007	.005
typeII_error * effect.size * rho	Sphericity Assumed	5.483	2	2.741	21.392	.000	.006	.005
	Greenhouse-Geisser	5.483	2.000	2.741	21.392	.000	.006	.005
	Huynh-Feldt	5.483	2.000	2.741	21.392	.000	.006	.005
	Lower-bound	5.483	2.000	2.741	21.392	.000	.006	.005
typeII_error * effect.size * scenario	Sphericity Assumed	22.130	2	11.065	86.344	.000	.024	.018
	Greenhouse-Geisser	22.130	2.000	11.065	86.344	.000	.024	.018
	Huynh-Feldt	22.130	2.000	11.065	86.344	.000	.024	.018
	Lower-bound	22.130	2.000	11.065	86.344	.000	.024	.018
typeII_error * rho * scenario	Sphericity Assumed	6.374	4	1.593	12.434	.000	.007	.005
	Greenhouse-Geisser	6.374	4.000	1.593	12.434	.000	.007	.005
	Huynh-Feldt	6.374	4.000	1.593	12.434	.000	.007	.005
	Lower-bound	6.374	4.000	1.593	12.434	.000	.007	.005

Table G1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squar -ed
typeII_error * n * J * effect.size	Sphericity Assumed	.284	1	.284	2.220	.136	.000	.000
	Greenhouse-Geisser	.284	1.000	.284	2.220	.136	.000	.000
	Huynh-Feldt	.284	1.000	.284	2.220	.136	.000	.000
	Lower-bound	.284	1.000	.284	2.220	.136	.000	.000
typeII_error * n * J * rho	Sphericity Assumed	2.247	2	1.123	8.766	.000	.002	.002
	Greenhouse-Geisser	2.247	2.000	1.123	8.766	.000	.002	.002
	Huynh-Feldt	2.247	2.000	1.123	8.766	.000	.002	.002
	Lower-bound	2.247	2.000	1.123	8.766	.000	.002	.002
typeII_error * n * J * scenario	Sphericity Assumed	.618	2	.309	2.410	.090	.001	.001
	Greenhouse-Geisser	.618	2.000	.309	2.410	.090	.001	.001
	Huynh-Feldt	.618	2.000	.309	2.410	.090	.001	.001
	Lower-bound	.618	2.000	.309	2.410	.090	.001	.001
typeII_error * n * effect.size * rho	Sphericity Assumed	.075	2	.038	.294	.745	.000	.000
	Greenhouse-Geisser	.075	2.000	.038	.294	.745	.000	.000
	Huynh-Feldt	.075	2.000	.038	.294	.745	.000	.000
	Lower-bound	.075	2.000	.038	.294	.745	.000	.000
typeII_error * n * effect.size * scenario	Sphericity Assumed	2.930	2	1.465	11.433	.000	.003	.002
	Greenhouse-Geisser	2.930	2.000	1.465	11.433	.000	.003	.002
	Huynh-Feldt	2.930	2.000	1.465	11.433	.000	.003	.002
	Lower-bound	2.930	2.000	1.465	11.433	.000	.003	.002
typeII_error * n * rho * scenario	Sphericity Assumed	.176	4	.044	.344	.849	.000	.000
	Greenhouse-Geisser	.176	4.000	.044	.344	.849	.000	.000
	Huynh-Feldt	.176	4.000	.044	.344	.849	.000	.000
	Lower-bound	.176	4.000	.044	.344	.849	.000	.000

Table G1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squar -ed
typeII_error * J * effect.size * rho	Sphericity Assumed	.651	2	.325	2.540	.079	.001	.001
	Greenhouse-Geisser	.651	2.000	.325	2.540	.079	.001	.001
	Huynh-Feldt	.651	2.000	.325	2.540	.079	.001	.001
	Lower-bound	.651	2.000	.325	2.540	.079	.001	.001
typeII_error * J * effect.size * scenario	Sphericity Assumed	.340	2	.170	1.327	.265	.000	.000
	Greenhouse-Geisser	.340	2.000	.170	1.327	.265	.000	.000
	Huynh-Feldt	.340	2.000	.170	1.327	.265	.000	.000
	Lower-bound	.340	2.000	.170	1.327	.265	.000	.000
typeII_error * J * rho * scenario	Sphericity Assumed	.207	4	.052	.404	.806	.000	.000
	Greenhouse-Geisser	.207	4.000	.052	.404	.806	.000	.000
	Huynh-Feldt	.207	4.000	.052	.404	.806	.000	.000
	Lower-bound	.207	4.000	.052	.404	.806	.000	.000
typeII_error * effect.size * rho * scenario	Sphericity Assumed	2.147	4	.537	4.188	.002	.002	.002
	Greenhouse-Geisser	2.147	4.000	.537	4.188	.002	.002	.002
	Huynh-Feldt	2.147	4.000	.537	4.188	.002	.002	.002
	Lower-bound	2.147	4.000	.537	4.188	.002	.002	.002
typeII_error * n * J * effect.size * rho	Sphericity Assumed	.263	2	.132	1.028	.358	.000	.000
	Greenhouse-Geisser	.263	2.000	.132	1.028	.358	.000	.000
	Huynh-Feldt	.263	2.000	.132	1.028	.358	.000	.000
	Lower-bound	.263	2.000	.132	1.028	.358	.000	.000
typeII_error * n * J * effect.size * scenario	Sphericity Assumed	.268	2	.134	1.047	.351	.000	.000
	Greenhouse-Geisser	.268	2.000	.134	1.047	.351	.000	.000
	Huynh-Feldt	.268	2.000	.134	1.047	.351	.000	.000
	Lower-bound	.268	2.000	.134	1.047	.351	.000	.000

Table G1 Continued

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Eta Squar-ed
typeII_error * n * J * rho * scenario	Sphericity Assumed	.866	4	.217	1.690	.149	.001	.001
	Greenhouse-Geisser	.866	4.000	.217	1.690	.149	.001	.001
	Huynh-Feldt	.866	4.000	.217	1.690	.149	.001	.001
	Lower-bound	.866	4.000	.217	1.690	.149	.001	.001
typeII_error * n * effect.size * rho * scenario	Sphericity Assumed	1.134	4	.284	2.213	.065	.001	.001
	Greenhouse-Geisser	1.134	4.000	.284	2.213	.065	.001	.001
	Huynh-Feldt	1.134	4.000	.284	2.213	.065	.001	.001
	Lower-bound	1.134	4.000	.284	2.213	.065	.001	.001
typeII_error * J * effect.size * rho * scenario	Sphericity Assumed	.144	4	.036	.282	.890	.000	.000
	Greenhouse-Geisser	.144	4.000	.036	.282	.890	.000	.000
	Huynh-Feldt	.144	4.000	.036	.282	.890	.000	.000
	Lower-bound	.144	4.000	.036	.282	.890	.000	.000
typeII_error * n * J * effect.size * rho * scenario	Sphericity Assumed	.279	4	.070	.544	.704	.000	.000
	Greenhouse-Geisser	.279	4.000	.070	.544	.704	.000	.000
	Huynh-Feldt	.279	4.000	.070	.544	.704	.000	.000
	Lower-bound	.279	4.000	.070	.544	.704	.000	.000
Error(typeII_error)	Sphericity Assumed	913.470	7128	.128				.757
	Greenhouse-Geisser	913.470	7128.000	.128				.757
	Huynh-Feldt	913.470	7128.000	.128				.757
	Lower-bound	913.470	7128.000	.128				.757