The Clinical Neuropsychologist, 2013
Vol. 27, No. 6, 1019–1042, http://dx.doi.org/10.1080/13854046.2013.806677

Diagnostic Accuracy of a Bayesian Latent Group Analysis for the Detection of Malingering-Related Poor Effort

Alonso Ortega[1,2], Stephan Labrenz[1,3], Hans J. Markowitsch[1,3] and Martina Piefke[1,3,4]

[1]Physiological Psychology, Bielefeld University, Bielefeld, Germany
[2]Escuela de Psicología, Facultad de Medicina, Universidad de Valparaíso, Valparaíso, Chile
[3]Center of Excellence "Cognitive Interaction Technology" (CITEC), Bielefeld University, Bielefeld, Germany
[4]Neurobiology and Genetics of Behavior, Department of Psychology and Psychotherapy, Witten/Herdecke University, Witten, Germany

In the last decade, different statistical techniques have been introduced to improve assessment of malingering-related poor effort. In this context, we have recently shown preliminary evidence that a Bayesian latent group model may help to optimize classification accuracy using a simulation research design. In the present study, we conducted two analyses. Firstly, we evaluated how accurately this Bayesian approach can distinguish between participants answering in an honest way (honest response group) and participants feigning cognitive impairment (experimental malingering group). Secondly, we tested the accuracy of our model in the differentiation between patients who had real cognitive deficits (cognitively impaired group) and participants who belonged to the experimental malingering group. All Bayesian analyses were conducted using the raw scores of a visual recognition forced-choice task (2AFC), the Test of Memory Malingering (TOMM, Trial 2), and the Word Memory Test (WMT, primary effort subtests). The first analysis showed 100% accuracy for the Bayesian model in distinguishing participants of both groups with all effort measures. The second analysis showed outstanding overall accuracy of the Bayesian model when estimates were obtained from the 2AFC and the TOMM raw scores. Diagnostic accuracy of the Bayesian model diminished when using the WMT total raw scores. Despite, overall diagnostic accuracy can still be considered excellent. The most plausible explanation for this decrement is the low performance in verbal recognition and fluency tasks of some patients of the cognitively impaired group. Additionally, the Bayesian model provides individual estimates, $p(z_i|\mathrm{D})$, of examinees' effort levels. In conclusion, both high classification accuracy levels and Bayesian individual estimates of effort may be very useful for clinicians when assessing for effort in medico-legal settings.

We thank our colleagues in the Physiological Psychology Department and the CITEC Research Groups at Bielefeld University, for their technical support and helpful advice. We thank also Dr. Eric-Jan Wagenmakers for sharing his knowledge about Bayesian methods with our research group. We would like to give special thanks to all participants, patients and to the staff from the Zentrum für Ambulante Rehabilitation, Bielefeld; the Median Rehabilitationsklinik, Bad Oeynhausen; and the Median Klinik, Bad Salzuflen, for their help and assistance when carrying out our clinical study. Finally we acknowledge the important contribution of our interns Ronja Boege, Jasmin Faber, Shaline Kockmann, Amelie Nikstat, and Verena Wittenberg during the research process. This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) (EC 277; MP, HJM), and by the Becas Chile Scholarship, CONICYT (AO).

Address correspondence to: Alonso Ortega, Universidad de Valparaíso, Facultad de Medicina, Escuela de Psicología, Av. Brasil 2140, Valparaíso, 2362854 Chile. Email: alonso.ortega@uv.cl

Accepted for publication May 15, 2013 First published online: June 12, 2013

Keywords: Malingering; Poor effort; Bayesian analysis; Symptom Validity Tests; Simulation Research Design; TOMM; WMT.

INTRODUCTION

Neuropsychological tests aim to measure a person's optimum performance in a variety of cognitive domains such as learning, memory, and attention. Therefore, neuropsychological assessment requires the examinees' best effort during testing in order to obtain valid results (Denney, 2008). Consequently, the validity of results of neuropsychological tests is susceptible to the influence of poor effort, exaggeration, and feigning. The effect of effort on test results is often acknowledged among malingering researchers (e.g., Iverson, 2007; Larrabee, 2007; Larrabee, Greiffenstein, Grewe, & Bianchini, 2007). Furthermore, Denney (2008) stated that the influence of effort has a much greater effect on neuropsychological test scores than brain injury or neurological conditions. Professional neuropsychological societies, such as the National Academy of Neuropsychology (NAN) and the American Academy of Clinical Neuropsychology (AACN), recommend the inclusion of effort measures in each neuropsychological testing battery in order to increase our confidence on the validity of testing results (Bush et al., 2005; Heilbronner et al., 2007; Heilbronner, Sweet, Morgan, Larrabee, & Millis, 2009).

Despite the inclusion of standard procedures to detect poor effort in neuropsychological testing situations, the risk of misclassification is still present. In practice, the occurrence of both false positive and false negative errors may lead to incorrect legal decisions or delay public support for people who are truly in need (Franzen, Iverson, & McCracken, 1990; Mossman, Wygant, & Gervais, 2012). To prevent potentially negative consequences of misclassification, malingering researchers have proposed and refined different malingering diagnostic criteria (Boone, 2007; Larrabee et al., 2007; Slick, Sherman, & Iverson, 1999). Similarly, they have also developed diverse strategies for the detection of poor effort (see Rogers, 2008).

Over the last two decades, malingering researchers have used embedded measures (i.e., effort indices obtained from standard cognitive tests) as well as other strategies that have been developed specifically for the detection of malingering. Among the latter, symptom validity tests (SVT) have been broadly used to distinguish poor effort from real cognitive impairment in neuropsychological settings (Grote & Hook, 2007). SVTs are based on a forced-choice paradigm that was initially implemented by Grosz and Zimmerman (1965) in an experimental analysis of hysterical blindness and used a decade later by Pankratz, Fausti, and Peed (1975) in the context of malingering detection. SVTs include a defined number of forced-choice trials. The probability of obtaining a predetermined amount of correct answers by chance alone is estimated using the normal approximation to the binomial distribution. Thus, "scores that fall below the chance level are considered as evidence of an intentional attempt to perform poorly on the test by the active avoidance of the correct answer" (Grote & Hook, 2007, p. 45). SVTs have been widely used in malingering research and practice. However, some authors have criticized the use of the below-chance criterion as a single decision rule to determine poor effort because of the low to moderate sensitivity levels (Beetar & Williams, 1995; Haines & Norris, 1995; Rogers, 2008; Slick et al., 2003). To increase sensitivity levels, test developers have derived cutoff scores from samples of

cognitively impaired patients (Iverson & Binder, 2000) instead of relying solely on the below-chance criterion (Grote & Hook, 2007). Currently, most SVTs do not use the below-chance criterion as the primary decision rule to determine poor effort (Frederick & Speed, 2007). Along with improvements made to SVTs, other interesting techniques have been introduced to improve the classification accuracy and sensitivity of effort measures.

In recent years, statistical methods used to detect poor effort in neuropsychological practice and research included odds and likelihood ratios (Bieliauskas, Fastenau, Lacy, & Roper, 1997; Weinborn, Orr, Woods, Conover, & Feix, 2003), the aggregation across multiple indicators (Larrabee et al., 2007; Larrabee, Millis, & Meyers, 2008), Bayesian average modeling (Larrabee et al., 2008; Millis & Volinsky, 2001; Wolfe et al., 2010), and Bayesian latent class modeling (Mossman et al., 2012). In this context, we recently tested a Bayesian latent group analysis as a method that may help to improve classification accuracy of effort testing in various neuropsychological settings (Ortega, Wagenmakers, Lee, Markowitsch, & Piefke, 2012). Our data show initial evidence that this Bayesian method may help to improve classification accuracy and test sensitivity both in clinical samples and in healthy controls who received different malingering instructions. The Bayesian latent group model also accurately differentiated between stroke patients with moderate cognitive impairment and participants who were trained to feign cognitive impairment. Data further suggest that the proposed method is resistant to coaching. However, additional research is required to corroborate the efficacy of this Bayesian approach.

In the present study, we aimed at evaluating the diagnostic accuracy of this Bayesian latent group model in two different experimental settings. Firstly, we intended to differentiate between healthy participants who gave their best effort when being tested (honest response group) from healthy participants who feigned cognitive impairment (experimental malingering group). This first analysis used a simulation research design. Secondly, we used a modified simulation research design in which the control group consisted of neurological patients with real cognitive deficits (cognitively impaired group). Patients were asked to give their best effort during testing.

To validate the Bayesian model, we evaluated all participants using three different screening measures of effort: a visual recognition two-alternative forced choice task (2AFC), the Test of Memory Malingering (TOMM; Tombaugh, 1996), and the Word Memory Test primary effort subtests (i.e., immediate recognition, IR; delayed recognition, DR; and consistency, CNS; Green, 2005). Raw scores from each effort measure were used as input in the Bayesian latent group model. Later, we estimated diagnostic accuracy indices for the Bayesian model using the results obtained in each effort measure.

We hypothesized that classification accuracy (e.g., sensitivity, specificity, etc.) of the Bayesian latent group model would be high in both analyses, irrespective of the effort measure. A special feature of this Bayesian latent group model is that it provides individual posterior classification estimates that represent the degree of certainty in the classification of each participant. These probabilistic estimates can be also viewed as the level of effort displayed by each participant during testing. We therefore propose that these probabilistic classification estimates may help practitioners and clinicians to make more informed decisions when assessing effort.

METHOD

Participants

Our sample consisted of 40 healthy participants and a clinical sample of 20 neurological patients with moderate cognitive impairment. Healthy participants were recruited in the city of Bielefeld, Germany. They were randomly assigned to either an honest response group (HR; $N = 20$; 5 males; 15 females; mean age = 27.10 ± 7.74 years) or an experimental malingering group (EM; $N = 20$; 12 males; 8 females; mean age = 26.75 ± 5.05 years). To be included in the study, participants had to be native German speakers between 18 and 60 years old. Exclusion criteria were current or lifetime neurological or psychiatric disorders and treatment with medication affecting the central nervous system.

The cognitively impaired group (CI) consisted of 20 patients (14 males; 6 females; mean age = 57.50 ± 15.19 years) with different neurological diseases (see later). Patients were recruited in three rehabilitation clinics: an outpatient rehabilitation center (Zentrum für Ambulante Rehabilitation [ZAR], Bielefeld) and two in-patient rehabilitation clinics (Median Rehabilitationsklinik, Bad Oeynhausen; Median Klinik, Bad Salzuflen) in the region of North Rhine-Westphalia, Germany. Clinicians at these institutions had previously determined the type and degree of the patients' cognitive impairment. In addition, this information was corroborated by a comprehensive neuropsychological evaluation of the patients' cognitive state (see subsection Measures). Most patients had chronic cognitive impairment after stroke ($N = 16$; mean duration after the event = 6.4 ± 4.3 months). Other neurological conditions included multiple sclerosis ($N = 1$), alcoholic polyneuropathy ($N = 1$), cerebral arteriovenous malformation ($N = 1$), and meningitis ($N = 1$).

Exclusion criteria for patients included severe sensorimotor deficits (e.g., hemiparesis or hemiplegia), severe visual and auditory deficits (e.g., hemianopia), severe anterograde amnesia, global aphasia, and large bilateral cerebral damage. Additionally, we decided to exclude patients who were involved in any kind of litigation process associated with economic compensation from health insurance companies. This decision was made in light of Mittenberg, Patton, Canyock, and Condit's (2002) findings, which reported that in worker's compensation settings malingering base rates tend to be higher than in other medical or psychiatric contexts. Further, the decision to exclude these patients helped prevent the inclusion of potential "true malingerers" in the study. Within all groups (i.e., HR, EM, CI), inclusion and exclusion criteria were tested by comprehensive medical and demographic anamnesis.

The study was accomplished in compliance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Written informed consent was obtained from all participants prior to participation. The ethics committee of the German Society of Psychology (Deutsche Gesellschaft für Psychologie, DGPs) approved the study.

Procedure

Prior to random assignment of patients into the HR and the CI groups, a comprehensive neuropsychological battery was applied. This was done to evaluate patients' cognitive status before receiving the instructions associated to the experimental phase of the study.

In the experimental phase, before being evaluated with three different screening measures of effort (see later), both the HR and the EM groups were instructed to imagine the following scenario: The participant was involved in a car accident and had a short loss of consciousness, or "black-out", immediately after the accident (<30 minutes) without long-term effects on their cognitive functioning. Afterwards, these two groups received specific role instructions: participants of the HR group were asked to give their best during the effort assessment. Participants of the EM group were instructed to feign some degree of cognitive impairment in the most credible way in order to receive benefits from a health insurance company (e.g., economic compensation or medical leave). They were also reminded that, if they were unable to be convincing, no benefits would be obtained. The scenario and the role instructions given to the healthy participants are consistent with methods used in previous studies on malingering (e.g., Jelicic, Merckelbach, Candel, & Geraerts, 2007; Powell, Gfeller, Hendricks, & Sharland, 2004). It was not necessary to present any kind of scenario to patients of the CI group because they suffered from real cognitive impairment.

Furthermore, a financial incentive was used to ensure that all participants would follow the role instructions. Each participant received €10 reimbursement for participation in the study. Additionally, we used a cover story in which participants were told that it was possible to get an extra incentive of €100 if they followed the role instructions as best as possible. According to Rogers and Cruise (1998), this sort of external incentive may enhance the external validity of simulation designs. After each participant had accomplished the experiment, the cover story was disclosed and the extra incentive was raffled among all 60 participants regardless of their performance.

Immediately after the role instructions were given, all participants were evaluated using the above-mentioned screening measures of effort: (i) a visual recognition 2AFC task, (ii) the TOMM, and (iii) the WMT primary effort subtests (for details see subsection Measures). Effort measures were administered in pseudo randomized order to control for any order effects. Thereafter, a post-test interview was applied to examine the participants' response strategies and to confirm whether participants understood and followed the role instructions (see Appendix 1). Two participants of the EM group reported that they did not completely follow the role instructions. Consequently, they were excluded from data analyses. After the experiment was finished, the intended purpose of the study was disclosed.

Measures

Visual recognition two-alternative forced-choice task. We used a visual recognition task based on a standard forced-choice paradigm (see Gutiérrez & Gur, 2012). The two-alternative forced-choice task (2AFC) task was previously developed in our research group for preliminary studies on the application of Bayesian inference for the assessment of malingering. We decided to include this task because our previous work (Ortega et al., 2012) included the same forced-choice paradigm. In this way, we were able to evaluate our previous findings. All items were psychometrically validated using the content validity ratio technique (see Lawshe, 1975). This ensured that all pictorial stimuli included in the task were adequate to measure visual recognition memory. The visual recognition 2AFC task consists of a learning and a recognition phase. During the learning phase, 50 simple colored drawings are presented. Half of them are

living things, the other half are non-living things. We distinguished between living and non-living things since patients with brain damage often show selective impairments in the processing of one of these categories while the other is preserved (double dissociation; see Gaffan & Heywood, 1993; Warrington & Shallice, 1984). Then, both living and non-living things were further subdivided into two sub-categories (i.e., animals and plants for the living things; vehicles and furniture for the non-living things). All drawings consisted of complete objects (instead of fractions or details) since patients with some neurological and psychiatric conditions may selectively fail to integrate disparate parts of visual objects (see Behrmann & Williams, 2007; Grailet, Seron, Bruyer, Coyette, & Frederix, 1990). These clinical neuropsychological considerations are in line with the recent recommendation of Bigler (2012) of taking into account neuropsychological findings when developing and interpreting SVTs.

During the recognition phase of the 2AFC, 50 pairs of drawings are presented. Each pair includes one presented item (i.e., target) and one novel item (i.e., distractor). The examinee presses a button (A or B) to select which item of each pair was presented in the learning phase. There is no time limit for answering each recognition trial. Following the recommendation of Hiscock and Hiscock (1989), trial-by-trial feedback was provided during the recognition phase. This feedback is also given in the TOMM and the WMT (see later). Examinees receive one point for each successfully recognized target. Thus, a maximum raw score of 50 can be obtained in the task. The 2AFC is a computer-based task. It was programmed and applied using the DirectRT™ software (Jarvis, 2008; New York: Empirisoft Corporation; www.empirisoft.com). The Bayesian latent group model used these 2AFC raw scores to estimate each examinee's probabilities of displaying poor effort.

TOMM. The TOMM (Tombaugh, 1996) is a visual memory recognition SVT that has been widely used as an effort measure in medico-legal or forensic settings (Constantinou & McCaffrey, 2003; Delain, Stafford, & Ben-Porath, 2003; Duncan, 2005; Iverson, Le Page, Koehler, Shojania, & Badii, 2007; MacAllister, Nakhutina, Bender, Karantzoulis, & Carlson, 2009). The test includes three parts: Trial 1, Trial 2, and a Retention Trial. Both Trials 1 and 2 consist of a learning phase followed by a recognition phase. The Retention Trial consists of a single recognition phase that should be administered approximately 15 minutes after the recognition phase of Trial 2. In each learning phase, 50 simple black and white line drawings are presented with a stimulus onset time of 3 seconds for each item. In the following recognition phase, every presented stimulus (i.e., target) is paired with a non-presented stimulus (i.e., distractor). The participants' task is to identify the drawings that were previously presented during the learning phases. There is no time limit for answering. Some authors argue that for efficiency reasons the Retention Trial can be left out (see Bauer, O'Bryant, Lynch, McCaffrey, & Fisher, 2007; Booksh, Aubert, & Andrews, 2007). A maximum score of 50 can be obtained in each section of the TOMM. The Bayesian latent group model used the TOMM Trial 2 raw scores to estimate the examinee's probabilities of displaying poor effort.

WMT. The WMT (Green, 2005) is also a widely used SVT. In contrast to the visual recognition 2AFC and the TOMM, the WMT is a verbal test. The WMT consists of both "effort" and "memory ability" subtests (Green, 2005). The WMT's primary effort subtests consist of two learning phases and two recognition phases.

Within the recognition phases, a word that was previously shown in the learning phase (i.e., target) is presented together with a word that was not presented previously (i.e., distractor). Participants have to decide which word they have seen before in the learning phases. An immediate recognition test (IR) is presented directly after each of the two learning trials. A delayed recognition test (DR) is then presented after a time interval of 30 minutes. A consistency score (CNS) is calculated by comparing the pattern of answers between IR and DR. The WMT includes also four memory ability subtests. These subtests are further divided into two "relatively easy memory subtests" and two "most difficult memory subtests" (Green, 2005). The former include a multiple choice task (MC) and a paired associates task (PA). The latter include a free recall task (FR) and long delayed free recall task (LDFR).

In this study we only used the WMT primary effort subtests. This decision relates to several well-designed studies that also used only the WMT primary effort subtests as a screening effort measure (e.g., Batt, Shores, & Chekaluk, 2008; Greiffenstein, Greve, Bianchini, & Baker, 2008; Mossman et al., 2012). The rationale behind this methodological decision was to ensure that all effort measures share a comparable theoretical construct targeting memory recognition. It is reasonable to assume that equivalency across tests diminishes possible testing biases that may affect diagnostic accuracy estimations of the Bayesian model. As Green (2005) states, "the recognition subtests (i.e., IR, DR) were designed to avoid confusing actual impairment with deliberate exaggeration" (p. 6). In our view, WMT subtests, which do not measure memory recognition (e.g., cued recall, free recall, long delayed free recall) would not have constituted adequate measures which can be used to estimate the diagnostic accuracy of our Bayesian model. According to Batt et al. (2008), the WMT primary effort subtests aim at measuring "effort" rather than cognitive "abilities". This idea is consistent with our decision of using only the WMT primary effort subtests as measure of effort.

In addition to the previous argument, Larrabee and Berry (2007) stated that aggregated indicators are more powerful predictors than any single indicator. That is, using two or more measures of effort (e.g., tests or subtests) will always lead to better predictions about the presence of poor effort than using a single measure. Consequently, accuracy estimations based on single measures may be underestimated when compared to estimations based on multiple measures. For this reason, all accuracy estimates were obtained using a single indicator from each effort measure in all analyses. Following the procedure used by Mossman et al. (2012) the WMT recognition subtests were transformed into a single effort indicator (i.e., WMT total score). The WMT IR, DR, and CNS percentages were first converted into raw scores, and then averaged to obtain a WMT total score. The Bayesian latent group model used this WMT total score to estimate the examinee's probabilities of displaying poor effort.

Neuropsychological assessment. As previously mentioned, all participants of the study were assessed with a standard neuropsychological testing battery prior to the experiment. Neuropsychological assessment aimed at: (i) describing the cognitive profile of each participant, (ii) comparing all three groups with respect to standard measures of cognitive performance, and (iii) excluding the presence of cognitive deficits that may affect the interpretation of the effort measures. Neuropsychological assessment included the following measures: (a) attention (d2-test; Brickenkamp, 2002), (b) visuospatial abilities (Rey–Osterrieth Complex Figure; Osterrieth, 1944; Rey, 1941), (c) word

recognition and fluency (Leistungsprüfungsystem, LPS; Horn, 1983), (d) analytic thinking (Leistungsprüfungsystem, LPS; Horn, 1983), and (e) personality traits (Freiburger Persönlichkeitsinventar, FPI; Fahrenberg, Hampel, & Selg, 2001). Additionally, subjects were screened for anxiety disorders (Beck Anxiety Scale, BAS; Margraf & Ehlers, 2007), and depression (Beck Depression Inventory, BDI; Beck, Steer, & Brown, 2006). No participants were excluded because of the presence of cognitive deficits.

Data analysis

Descriptive and inferential statistics. We analyzed the raw scores of the 2AFC and the TOMM, as well as the WMT total scores to describe each group's average task performance. To compare performance between groups we used the Kruskal–Wallis tests and *post-hoc* multiple comparisons with the Mann–Whitney *U*-test when assumptions for parametric testing were not met.

Bayesian latent group analysis. To introduce our Bayesian model, we firstly provide a short description of how Bayesian inference works. A standard Bayesian analysis usually includes three sources of information: (i) a model that indicates how latent parameters generate data, (ii) a prior probability distribution that represents previous knowledge about the parameters to be estimated (e.g., malingering base rates), and (iii) the observed data (e.g., effort measures raw scores). By combining these three elements, we obtain a posterior probability distribution, which represents knowledge about the parameters of the model after the data has been observed. This is the core of almost every Bayesian analysis.

In the following, we specify the main features and assumptions of the Bayesian latent group model used in this study. Firstly, our Bayesian model purports to identify participants who are displaying poor effort when tested. For this purpose, the model provides probabilistic estimates of the level of effort displayed by each examinee, as well as the degree of confidence with which each participant is classified. Secondly, the model's assumptions should be specified. The present model assumes the existence of two latent groups: (i) one group who will answer the effort tests in an honest way (i.e., HR or CI) and (ii) one group who will feign some kind of cognitive impairment when completing the same tasks (i.e., EM). The model also assumes that both HR and CI groups will obtain higher success rates than the EM group. Together, these two assumptions imply that both HR and CI groups will perform—at least—at chance level or above (i.e., $\geqslant 50\%$), whereas the EM group will perform always worse than both honest answering groups (i.e., HR, CI).

The use of a forced-choice paradigm implies the existence of only two possible outcomes per answer, which in this case are success or failure. Considering this, we can assume that data will be binomially distributed within each group. Therefore, each group will have an unknown underlying base rate θ (i.e., prevalence) that will be determined by the observed amount of successes and failures. Since θ is unknown, we need to determine a prior value for this parameter before introducing the collected data. This prior information, θ, is of particular relevance in Bayesian inference, and especially in our model. The more specific the prior information, the more accurate are the posterior estimates. Hence, using appropriate prior information on malingering base rates, a given test result will lead to more precise posterior estimates of effort. In our

view, this constitutes an important advantage of the Bayesian model that will provide clinicians with more precise information that assists them to reach more informed decisions when assessing for effort. More importantly, when information about base rates is unknown or not available, our Bayesian model can estimate them from the data. The clinical utility of considering prior information about base rates is further elaborated in the discussion section.

In the present study we assumed a prior value of 40% as malingering base rate (Larrabee, 2003, 2007; Larrabee, Millis, & Meyers, 2009) for both analyses. This value represents our initial knowledge about θ before seeing the data. Our model also assumes that the base rate θ is "beta-binomially distributed", with parameters "α" and "β". A beta-binomial distribution is a family of discrete probability distributions. The advantage of using a beta-binomial distribution is that parameters α and β can be thought of as counts for "prior successes" and "prior failures" (see Lee & Wagenmakers, 2010). For example, the expression $\theta \sim$Beta (60,40) implies that the base rate θ is beta-binomially distributed (i.e., data are discrete) with parameters α and β equivalent to 60 successful answers and 40 failed answers. Our prior knowledge on θ thus expects that failure rates will be distributed around 40%. This prior value on θ is congruent with the assumption of our Bayesian model, which expects success rates $\geqslant$50% for both the HR and CI groups. Additionally, success and failure rates are governed by a group-level beta distribution ("Beta-binomial hierarchical model"; see Lee & Wagenmakers, 2010). Using a hierarchical model implies that, in this case, group membership will determine success and failure rates. In this way, high failure rates are expected for the EM group and low failure rates are expected for the HR and the CI groups.

Finally, we obtain a posterior probability distribution, $p(\theta \mid \mathrm{D})$, which represents the updated knowledge about θ after data have been observed. Individual posterior classification probabilities, $p(\mathrm{z}_i \mid \mathrm{D})$, are estimated for each participant. The z_i classification variable can take values from 0 to 1. A $p(\mathrm{z}_i \mid \mathrm{D})$ mean value close to 0 indicates that a participant has a low probability of being classified as giving poor effort. On the contrary, a $p(\mathrm{z}_i \mid \mathrm{D})$ mean value close to 1 indicates a high probability of being classified as giving poor effort. The parameter $p(\mathrm{z}_i \mid \mathrm{D})$ therefore indicates the theoretical position of each participant's performance along a continuum of effort (see Iverson, 2010, p. 110). This feature of our Bayesian latent group model is consistent with recent views that consider effort and malingering as continua rather than taxonomies (see Iverson, 2010; Walters et al., 2008; Walters, Berry, Lanyon, & Murphy, 2009). We believe that providing individual posterior classification estimates of effort constitutes another advantage of the proposed model. To our knowledge, these $p(\mathrm{z}_i \mid \mathrm{D})$ estimates are not provided by any other currently available method for effort assessment.

All Bayesian analyses were implemented using the Markov-Chain Monte Carlo sampling method (MCMC; e.g., Gamerman & Lopes, 2006; Gilks, Richardson, & Spiegelhalter, 1996) in the WinBUGS software program (Lunn, Spiegelhalter, Thomas, & Best, 2009; Lunn, Thomas, Best, & Spiegelhalter, 2000). For a more detailed description of our Bayesian model see Ortega et al. (2012). General information about Bayesian inference can be found in O'Hagan and Forster (2004), Lee and Wagenmakers (2010), and Dienes (2011).

Analysis of diagnostic accuracy. The utility of any assessment method must be clearly stated in terms of diagnostic accuracy statistics (i.e., classification accuracy

indices; Aronoff et al., 2007; Etherton, Bianchini, Ciota, Heinly, & Greve, 2006). For this reason we calculated classification accuracy indices for our Bayesian approach. We defined a $p(z_i \mid D)$ mean value greater than 90% as a cutoff to classify each participant as suspicious for poor effort. Note that the only purpose of dichotomizing our Bayesian estimates was to allow the calculation of diagnostic accuracy indices. As we recently stated, a main advantage of the Bayesian model is that it offers probabilistic effort estimates (as a continuum), and therefore cutoffs are not necessary.

We report sensitivity, specificity, and predictive values because they are the most relevant classification accuracy indices in malingering research (Etherton et al., 2006). Additionally, we report values for the estimated area under the receiver operating characteristic curve (ROC). The area under the curve (AUC) provides an overall discrimination value for any diagnostic method. The higher the AUC value, the better the discrimination capacity of the test. In this study, we considered the Hosmer and Lemeshow (2000) criteria. These suggest that AUC values equal to 0.5 represent "no discrimination", AUC values between 0.70 and 0.80 represent "acceptable discrimination", AUC values between 0.81 and 0.90 represent "excellent discrimination", and AUC values greater than 0.90 represent "outstanding discrimination".

Complementary analyses. We conducted complementary statistical analyses of the neuropsychological testing battery. All groups were compared for analytic thinking, word recognition and fluency, concentration, personality scales. For this purpose, between-group ANOVAs as well as Tukey *post-hoc* comparisons were performed on all aforementioned neuropsychological variables. All these complementary analyses were conducted using SPSS version 18.

RESULTS

Descriptive and inferential analyses

Descriptive analyses showed that the HR group obtained the best results in all effort measures, followed by the CI group and the EM group (Table 1). Standard

Table 1. Means and standard deviation of the honest response, cognitively impaired, and experimental malingering groups in the effort measures

		Effort measures					
		Visual recognition 2AFC		Test of Memory Malingering		Word Memory Test	
		Raw scores		Raw scores (Trial 2)		Total raw scores[a]	
Group	*n*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Honest response	20	49.30	1.08	50.00	0.00	39.00	1.42
Cognitively impaired	20	47.40	3.42	48.80	3.05	35.00	4.74
Experimental malingering	20	31.65	4.67	33.05	5.91	26.00	4.03

[a]Word Memory Test total raw scores correspond to the converted and averaged immediate recognition (ID), delayed recognition (DR), and consistency (CNS) primary effort subscales.

deviations (*SD*) for the HR group are smaller than those for the CI group and the EM group on all measures.

Kruskall–Wallis inferential analyses and *post-hoc* multiple comparisons showed that the HR group significantly over performed both the CI and the EM groups in all effort measures ($p < .01$). Moreover, performance of the CI group was significantly better than the EM group in all effort measures ($p < .01$).

Honest response group vs. experimental malingering group

All individual classification posterior probabilities $p(z_i \mid D)$ were close to 0 within the HR group and close to 1 within the EM group, across all measures (i.e., 2AFC, TOMM, WMT). The Bayesian latent analysis thus clearly distinguished participants from each group. In terms of classification accuracy, the Bayesian model achieved a 100% sensitivity and specificity, independently of the used effort measure.

Cognitive impairment group vs. experimental malingering group

The Bayesian model showed high sensitivity levels, irrespective of the effort measure used. Only one false negative was found when using the 2AFC (i.e., P_{31}), the WMT (i.e., P_{31}), and the TOMM (i.e., P_{27}) raw scores to estimate the examinee's probabilities of displaying poor effort. High specificity levels were observed for the Bayesian model when estimates were obtained from the 2AFC and TOMM raw scores.

When using the 2AFC and the TOMM Trial 2 raw scores, only one patient from the CI group (i.e., P_{20}) was found to be a false positive. In contrast, when using the WMT total raw scores, the posterior individual classification probabilities $p(z_i \mid D)$ did not allow for a correct classification of all participants. In particular, five false positive were found in the CI group (Table 2). Consequently, specificity levels decreased.

The overall classification accuracy for the Bayesian model was outstanding when using the 2AFC and the TOMM raw scores, following Hosmer and Lemeshow (2000) criteria. All accuracy indices were identical when using the raw scores of these two effort measures as input for the Bayesian model. When the WMT total raw scores served as input for the Bayesian model the classification accuracy diminished. Despite this decrement, classification accuracy of the Bayesian model was still excellent according to Hosmer and Lemeshow (2000) criteria. Variations on the classification accuracy indices for the Bayesian model are summarized in Table 3.

Best overall classification accuracy estimates were obtained when using the 2AFC and the TOMM raw scores. The positive Likelihood Ratio (LR+) and the Positive Predictive Value (PPV) showed the most important differences when estimates were based on the 2AFC and the TOMM. Classification accuracy was lower when using the WMT total scores. The negative predictive value (NPV) was almost similar, independently of the used effort measure. However, the positive predictive power (PPV) diminished when estimations were obtained from the WMT total scores.

Neuropsychological assessment

Here we present the most relevant neuropsychological results. A report of the main neuropsychological findings is provided in the Appendix Table 2a. Appendix

Table 2. Results of the Bayesian individual posterior classification probabilities $p(z_i \mid D)$ for the 2AFC visual recognition task, the Test of Memory Malingering (Trial 2), and the Word Memory Test (primary effort subscales)

		Screening measures of effort											
		2AFC Visual recognition task				TOMM (Trial 2)				WMT Primary effort subtests			
		Individual probability $p(z_i \mid D)$[a]				Individual probability $p(z_i \mid D)$[a]				Individual probability $p(z_i \mid D)$[a]			
Group	Participant$_i$	Raw score	Mean	*SD*	Decision	Raw score	Mean	*SD*	Decision	Total raw score	Mean	*SD*	Decision
Cognitively impaired	P_1	47	.002	.046	pass	50	.000	.000	pass	34	.567	.496	pass
	P_2	49	.000	.000	pass	50	.000	.000	pass	37	.032	.177	pass
	P_3	45	.039	.194	pass	50	.000	.000	pass	38	.007	.086	pass
	P_4	49	.000	.006	pass	50	.000	.000	pass	40	.000	.015	pass
	P_5	46	.010	.100	pass	48	.002	.047	pass	35	.304	.459	pass
	P_6	50	.000	.000	pass	50	.000	.000	pass	39	.001	.034	pass
	P_7	44	.138	.345	pass	46	.085	.278	pass	26	.999	.011	fail*
	P_8	50	.000	.000	pass	50	.000	.000	pass	39	.002	.039	pass
	P_9	46	.009	.096	pass	50	.000	.000	pass	29	.997	.055	fail*
	P_{10}	48	.000	.023	pass	49	.000	.014	pass	39	.001	.036	pass
	P_{11}	50	.000	.000	pass	50	.000	.000	pass	40	.000	.012	pass
	P_{12}	49	.000	.000	pass	50	.000	.000	pass	28	.999	.037	fail*
	P_{13}	49	.000	.000	pass	46	.087	.282	pass	37	.033	.178	pass
	P_{14}	50	.000	.006	pass	50	.000	.000	pass	39	.000	.030	pass
	P_{15}	50	.000	.000	pass	50	.000	.000	pass	39	.001	.034	pass
	P_{16}	50	.000	.000	pass	50	.000	.000	pass	35	.297	.457	pass
	P_{17}	44	.138	.345	pass	50	.000	.006	pass	30	.996	.065	fail*
	P_{18}	46	.011	.105	pass	50	.000	.000	pass	39	.001	.036	pass
	P_{19}	50	.000	.000	pass	50	.000	.000	pass	40	.001	.036	pass
	P_{20}	36	.999	.024	fail*	37	.999	.011	fail*	28	.998	.043	fail*

(*Continued*)

Table 2. (*Continued*).

		Screening measures of effort											
		2AFC Visual recognition task				TOMM (Trial 2)				WMT Primary effort subtests			
		Individual probability $p(z_i \mid D)$[a]				Individual probability $p(z_i \mid D)$[a]				Individual probability $p(z_i \mid D)$[a]			
Group	Participant$_i$	Raw score	Mean	*SD*	Decision	Raw score	Mean	*SD*	Decision	Total raw score	Mean	*SD*	Decision
Experimental malingering	P_{21}	27	1	.000	fail	32	1	.000	fail	23	1	.000	fail
	P_{22}	35	.999	.011	fail	39	.999	.027	fail	27	.999	.022	fail
	P_{23}	28	1	.000	fail	39	.999	.027	fail	32	.915	.279	fail
	P_{24}	34	.999	.011	fail	34	1	.000	fail	27	.999	.023	fail
	P_{25}	35	.999	.023	fail	33	1	.000	fail	25	1	.006	fail
	P_{26}	32	1	.006	fail	34	1	.000	fail	27	.999	.027	fail
	P_{27}	39	.979	.141	fail	45	.285	.451	pass**	32	.913	.282	fail
	P_{28}	29	1	.000	fail	29	1	.000	fail	23	1	.000	fail
	P_{29}	29	1	.000	fail	26	1	.000	fail	23	1	.000	fail
	P_{30}	30	1	.000	fail	30	1	.000	fail	25	.999	.009	fail
	P_{31}	41	.825	.380	pass**	40	.996	.063	fail	33	.790	.409	pass**
	P_{32}	38	.994	.075	fail	40	.997	.052	fail	32	.915	.278	fail
	P_{33}	31	1	.000	fail	33	1	.000	fail	24	1	.000	fail
	P_{34}	25	1	.000	fail	24	1	.000	fail	24	1	.006	fail
	P_{35}	27	1	.000	fail	28	1	.000	fail	24	1	.000	fail
	P_{36}	36	.999	.030	fail	38	.999	.015	fail	30	.989	.104	fail
	P_{37}	32	1	.006	fail	35	1	.000	fail	30	.988	.108	fail
	P_{38}	25	1	.000	fail	23	1	.000	fail	20	1	.000	fail
	P_{39}	33	1	.006	fail	32	1	.000	fail	21	1	.000	fail
	P_{40}	27	1	.000	fail	27	1	.000	fail	21	1	.000	fail

Note: Samples from the posterior distribution = 27,000. *False Positives. **False Negatives.

[a]All individual classification probabilities $p(z_i \mid D)$ were estimated considering a prior on malingering base rate $\theta \sim dbeta(6,4) \approx 40\%$.

Table 3. Classification accuracy of the Bayesian Model using the 2AFC visual recognition task, the Test of Memory Malingering (Trial 2), and the Word Memory Test (primary effort subscales)

	Classification accuracy indices					
	2AFC (raw scores)		TOMM Trial 2 (raw scores)		WMT primary effort subscales (total raw scores)	
Index	Value	95% CI	Value	95% CI	Value	95% CI
AUC[a]	.95	(.87–1)	.95	(.87–1)	.85	(.72–.97)
Sensitivity	.95	(.75–.99)	.95	(.75–.99)	.95	(.75–.99)
Specificity	.95	(.75–.99)	.95	(.75–.99)	.75	(.51–.91)
LR (+)	19	(2.81–128.7)	19	(2.81–128.7)	3.80	(1.77–8.17)
LR (–)	0.05	(0.01–0.36)	0.05	(0.01–0.36)	0.07	(0.01–0.46)
PPV[b]	.93	(.90–.94)	.93	(.90–.94)	.72	(.68–.75)
NPV[b]	.97	(.95–.98)	.97	(.95–.98)	.96	(.94–.97)
TP	19		19		19	
TN	19		19		15	
FP	1		1		5	
FN	1		1		1	
N	40		40		40	

Note: AUC = estimated Area Under the Curve; LR(+) = Likelihood Ratio for a positive test result; LR(–) = Likelihood Ratio for a negative test result; PPV = Positive Predictive Value; NPV = Negative Predictive Value; TP = True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives; N = Total Sample Size.

[a]Hosmer and Lemeshow (2000) guidelines were considered to interpret the AUC value.

[b]PPV and NPV were estimated considering a malingering base rate (i.e., prevalence) of 40% (Larrabee, 2003, 2007; Larrabee et al., 2009).

Table 2b provides all neuropsychological test results for personality and clinical measures.

No significant differences were observed between all three groups in measures of personality traits (FPI) and long-term memory (Rey–Osterrieth Complex Figure), depression (BDI) and anxiety (BAS). Secondly, the HR and EM group evidenced significant differences in a measure of analytic thinking (LPS). However, no significant differences were observed between the aforementioned groups on measures of concentration (d2-test) and word recognition and fluency (LPS).

Participants of the HR and EM groups had normal average cognitive performance, whereas patients of the CI group showed moderate levels of cognitive impairment. The CI group scored significantly worse than both HR and EM groups on concentration, and word recognition and fluency ($p < .01$) and had lower scores on analytic thinking than HR (see Appendix Table 2a).

CI patients who failed in the WMT (i.e., false positives) performed significantly worse in word recognition and fluency (LPS; Horn, 1983) than those CI patients who passed the WMT, $t(18) = 2.99$, $MSE = 2.50$, $p < .01$.

DISCUSSION

The present study included two analyses that aimed at evaluating the diagnostic accuracy of a Bayesian latent group model to detect malingering-related poor effort.

The first analysis used a simulation research design to distinguish between participants assigned to an honest response group (HR) and those assigned to an experimental malingering group (EM). The second analysis aimed at differentiating patients with genuine cognitive impairment (CI) from participants assigned to the EM group.

The first analysis replicated our previous findings (Ortega et al., 2012), corroborating that a Bayesian approach is highly accurate when differentiating HR from EM participants. The accuracy level of the Bayesian model was 100%, regardless of the effort measure. Based on our previous study, these results were partially expected because HR and EM participants constitute extreme groups (i.e., their average performance show "high" or "low" scores in all effort measures), and therefore they are relatively easy to identify with this Bayesian analysis. For this reason, the second analysis included a clinical group where patients with real cognitive deficits (CI) had to be differentiated from participants assigned to the EM group. We proposed that including the CI group would allow for a better extrapolation of the findings to real life settings.

Descriptive analyses showed that the CI group scored significantly better than the EM group on all effort measures. This finding is consistent with the view of Denney (2008) who suggested that examinees' effort has more impact on test results than their clinical condition. The pattern of answers was more homogeneous within the HR group. In contrast, within both the CI and EM groups, the answering patterns showed greater levels of variability. Regarding the CI group, our interpretation is that variability may reflect both differences in the type and degree of cognitive impairment as well as the strategies used to cope with cognitive disabilities. Regarding the observed variability in the EM group, we hypothesize that it may reflect interindividual differences in "feigning" strategies used to follow the role instructions both within and between measures of effort. This hypothesis is in agreement with Rogers (2008), who proposed that malingerers commonly change and combine their response strategies both between and within different effort measures.

Following the Hosmer and Lemeshow (2000) criteria, the overall diagnostic accuracy of the Bayesian model can be classified as outstanding when estimates are obtained from the 2AFC and the TOMM raw scores. However, when using the WMT total scores specificity, positive likelihood ratio (LR+), and positive predictive value (PPV) decreased markedly. Regarding predictive values, it must be considered that they are affected by changes in the base rates (i.e., prevalence rates). For instance, as the base rate drops so does the PPV, whereas the proportion of false positives increases (Streiner, 2003). However, the lower the base rate, the higher the NPV. In this study, which considered a malingering base rate of 40%, NPVs were high for all effort measures. This means that a negative test result gives us great confidence to discard the presence of poor effort, at least interpreting the results of these particular tests. Nevertheless, if we consider a different base rate, these values and their interpretations may vary.

When using the WMT total scores, the false positive rate for the Bayesian model was 25%. This finding may at least in part depend on the fact that we did not apply all WMT subtests as suggested by Green (2005). To this respect, we also acknowledge that the WMT classification results might have been indeed different (e.g., lower false positive rate) if applying all WMT subtests and analyzing the GMIP profile. Nevertheless, Mossman et al. (2012) recently obtained high overall diagnostic accuracy (i.e., AUC = .929 ± .020) using the same WMT total raw scores as we did under a comparable Bayesian framework. Therefore, the reason for the high false positive rate

observed in our study needs to be further investigated. Performance of the CI group in verbal fluency and recognition tasks may provide an alternative explanation for the false positives obtained when using WMT total raw scores. In particular, we observed that those patients who failed in the WMT effort subtests performed significantly worse in verbal fluency and recognition in contrast to patients that passed the WMT effort subtests. Considering the verbal nature of the WMT, it is reasonable to assume that low scores in verbal abilities may have influenced these patients' WMT performance. Taken together, these data and the findings of Mossman et al. (2012) suggest that low performance of some patients in verbal fluency and recognition represent another highly plausible interpretation for the WMT false positive rates in the context of the present study.

The Genuine Memory Impairment Profile (GMIP; see Martins & Martins, 2010) may have helped to diminish the WMT false positive rates, but this would have required using the entire WMT test. However, the use of the WMT ability subtests would not have allowed for comparable measures in the 2AFC, the TOMM, and the WMT. Despite the low specificity observed when using the WMT total scores, the overall classification accuracy of the Bayesian model can still be considered excellent according to Hosmer and Lemeshow's (2000) criteria. Moreover, diagnostic accuracy of the Bayesian model increases when estimates are obtained from the 2AFC and the TOMM. In summary, the overall classification accuracy of the Bayesian latent group analysis is well supported by our data.

As Iverson (2007) stated, "Over the past several years, researchers have been encouraging the use of Bayesian methods for effort testing. Unfortunately, Bayesian methods and other interesting statistical methodologies … are rarely used in mainstream clinical practice" (p. 97). Therefore, we can highlight here the clinical and practical relevance of our Bayesian model. This Bayesian latent group analysis allows for a probabilistic estimation of the examinees' level of effort. The model considers effort as a continuum that may vary in magnitude instead of a dichotomous variable. Recently, Bigler (2012) published an exhaustive review of the SVT method. In his work, Bigler (2012) specifically addressed issues concerning the use of cutoff scores to determine the presence or absence of effort. In Bigler's (2012) opinion, cutoff scores impose artificial "pass/fail" dichotomies that should not be used as a dichotomous defining point for presence or absence of a deficit. Furthermore, he argued that scores which are "above chance" but "below the cutoff score" (i.e., near-pass scores) may confront clinicians with the risk of type I and type II errors (i.e., false positives, false negatives). Considering Bigler's (2012) criticisms of SVT cutoff scores, we propose that Bayesian classification probabilities, $p(z_i \mid D)$, provide complementary information to interpret an examinee's performance and thus help clinicians to avoid committing type I and type II errors. In our data we can observe different cases in which the use of $p(z_i \mid D)$ estimates may be of high utility. For instance, patients P_7 and P_{17} of the cognitively impaired group scored 44/50 in the 2AFC. This score may be considered as a near pass score, since is slightly below 90% of correct answers (i.e., <45). However, the estimated $p(z_i \mid D)$ mean value for these patients is 13.8%, which represents a relatively low probability of giving poor effort. Considering this probabilistic estimate, the clinician may suggest additional testing instead of determining the possible presence of poor effort based solely on the mentioned cutoff score. This example emphasizes how Bayesian individual probabilities may help clinicians to

avoid type I errors in presence of a "near-pass" score. However, failing on effort measures is not the unique criterion to determine the presence of malingering. Therefore, these results should be interpreted properly following the recommended guidelines (see Boone, 2007; Larrabee et al., 2007; Slick et al., 1999).

In addition, the use of prior information related to malingering base rates may also support the clinical utility of this Bayesian analysis. As broadly known, a cutoff score of 45/50 represents a "pass" in the TOMM. However, in our view, this information cannot be properly interpreted without considering information about malingering base rates. Several studies report different malingering base rates, which show variations according to the setting in which they were obtained (e.g., Ardolf, Denney, & Houston, 2007; Chafetz, 2008; Duff et al., 2011; Larrabee, 2007; Mittenberg et al., 2002). According to these studies, lower base rates are found in clinical settings where patients are not seeking for compensation. On the contrary, in medico-legal settings higher base rates have been reported. From our perspective, this information should not be obviated. For example, assuming malingering base rate equal to 10%, the $p(z_i \mid D)$ mean value is 97% for a patient who scored exactly 45 in the TOMM (i.e., "pass"). This additional information may be sufficient to suggest additional testing. The same analysis, but considering a malingering base rate of 60%, lead to a $p(z_i \mid D)$ mean value of 29%. In this case, this information may suggest the clinician discard the presence of poor effort. This example illustrates how prior information about base rates may be used clinically. In 2000, Rosenfeld, Sands, and Van Gorp (2000) already emphasized the impact of malingering base rates on the accuracy of predictive models. The availability of such base rates may encourage clinicians to try Bayesian approaches in applied settings (e.g., clinical, forensic). The recent development of specialized software (e.g., WinBUGS; Lunn et al., 2000; Lunn et al., 2009) and modern sampling methods (e.g., MCMC; Gamerman & Lopes, 2006) could facilitate the application of the proposed model. In our previous work (Ortega et al., 2012) we provided a supplementary file that includes all technical tools and elements required to conduct a Bayesian latent group analysis for effort assessment.

One limitation of our study is that the use of rather small sample sizes might have affected the estimation of diagnostic accuracy statistics. The relevant indices (i.e., sensitivity, specificity, likelihood ratios, and predictive values) may decrease when they are estimated from larger samples. It should be emphasized, however, that small sample sizes do not affect individual posterior classification probabilities $p(z_i \mid D)$ since Bayesian estimations show little variations irrespective of the sample size (see Jaynes, 2003; Jeffreys, 1961). From this view, this can also be seen as an advantage in clinical studies with highly specific inclusion criteria where sample sizes are small. A second limitation is related to the use of a simulation design, which tends to minimize the external validity of any study. Nonetheless, the implementation of more sophisticated designs (e.g., known-groups comparison; see Rogers, 2008) does not ensure the external validity of the results. In this respect Millis (2009) stated, "defining the reference sample in diagnostic studies of malingering tests can be challenging because there is no universally accepted 'gold standard' for malingering" (p. 24). We emphasize that this lack of a gold standard is a common situation in medicine, epidemiology and not limited to neuropsychology (Joseph, Gyorkos, & Coupal, 1995). The investigation of clinical patients who were not engaged in financial compensation seeking constitutes another limitation of our study. To provide a reliable generalization of the present

findings, studies using similar Bayesian approaches need to be carried out in different populations and settings. It would be of high scientific interest and practical relevance to evaluate the utility of Bayesian approaches in at-risk malingering patients (e.g., cognitive disability claimants) or in forensic contexts. Finally, using only the WMT primary effort subtests may constitute another limitation, but our experimental design did not allow for using the entire WMT. In future studies, the application of the entire WMT within a Bayesian framework will be definitely considered.

In sum, our data suggest that a Bayesian latent group analysis provides relevant information that may help practitioners to reach more informed decisions regarding the level of effort displayed by an examinee during neuropsychological assessment. To our knowledge, neither SVTs, nor other currently available approaches provide these individual probabilistic estimates of effort. However, we emphasize that we do not promote the use of Bayesian models as an exclusive method for the detection of poor effort. Rather, Bayesian statistics should be applied to complement information provided by the available techniques of effort assessment. Taken together, the current results and our previous study (Ortega et al., 2012) highlight the utility of Bayesian probabilistic approaches in the detection of malingering-related poor effort. Future research needs to be conducted in order to further explore and specify the benefits of Bayesian modeling in malingering assessment.

REFERENCES

Ardolf, B. R., Denney, R. L., & Houston, C. M. (2007). Base rates of negative response bias and malingered neurocognitive dysfunction among criminal defendants referred for neuropsychological evaluation. *The Clinical Neuropsychologist, 21*, 899–916.

Aronoff, G., Mandel, S., Genovese, E., Maitz, E., Dorto, A., Klimek, E., & Staats, T. E. (2007). Evaluating malingering in contested injury or illness. *Pain Practice, 7*, 178–204.

Bauer, L., O'Bryant, S. E., Lynch, J. K., McCaffrey, R. J., & Fisher, J. M. (2007). Examining the Test of Memory Malingering trial 1 and word memory test immediate recognition as screening tools for insufficient effort. *Assessment, 14*, 215–222.

Batt, K., Shores, E. A., & Chekaluk, E. (2008). The effect of distraction on the Word Memory Test and Test of Memory Malingering performance in patients with a severe brain injury. *Journal of the International Neuropsychological Society, 14*, 1074–1080.

Beck, A. T., Steer, R. A., & Brown, G. L. (2006). *Beck Depressionsinventar* (2nd ed.). Frankfurt/Main, Germany: Harcourt Test Services GmbH.

Beetar, J. T., & Williams, J. M. (1995). Malingering response styles on the memory assessment scales and symptom validity tests. *Archives of Clinical Neuropsychology, 10*, 57–72.

Behrmann, M., & Williams, P. (2007). Impairments in part–whole representations of objects in two cases of integrative visual agnosia. *Cognitive Neuropsychology, 24*, 701–730.

Bieliauskas, L. A., Fastenau, P. S., Lacy, M. A., & Roper, B. L. (1997). Use of the odds ratio to translate neuropsychological test scores into real-world outcomes: From statistical significance to clinical significance. *Journal of Clinical and Experimental Neuropsychology, 19*, 889–896.

Bigler, E. D. (2012). Symptom validity testing, effort, and neuropsychological assessment. *Journal of the International Neuropsychological Society, 18*, 632–642.

Booksh, R. L., Aubert, M. J., & Andrews, S. R. (2007). Should the retention trial of the Test of Memory Malingering be optional? A reply. *Archives of Clinical Neuropsychology, 22*, 87–89.

Boone, K. B. (2007). A reconsideration of the Slick et al. criteria for malingered neurocognitive dysfunction. In K. B. Boone (Ed.), *Assessment of feigned cognitive impairment: A neuropsychological perspective* (pp. 29–49). New York, NY: Guilford Press.

Brickenkamp, R., & Zillmer, E. (1998). *The d2 test of attention*. Seattle, WA: Hogrefe & Huber.

Bush, S. S., Ruff, R. M., Troster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., … Silver, C. H. (2005). Symptom validity assessment: Practice issues and medical necessity – NAN policy & planning committee. *Archives of Clinical Neuropsychology, 20*, 419–426.

Chafetz, M. D. (2008). Malingering on the social security disability consultative exam: Predictors and base rates. *The Clinical Neuropsychologist, 22*, 529–546.

Constantinou, M., & McCaffrey, R. J. (2003). Using the TOMM for evaluating children's effort to perform optimally on neuropsychological measures. *Child Neuropsychology, 9*, 81–90.

Delain, S. L., Stafford, K. P., & Ben-Porath, Y. S. (2003). Use of the TOMM in a criminal court forensic assessment setting. *Assessment, 10*, 370–381.

Denney, R. L. (2008). Negative response bias and malingering during neuropsychological assessment in criminal forensic settings. In R. L. Denney & J. P. Sullivan (Eds.), *Clinical neuropsychology in the criminal forensic setting* (pp. 91–134). New York, NY: Guilford Press.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274–290.

Duff, K., Spering, C. C., O'Bryant, S. E., Beglinger, L. J., Moser, D. J., Bayless, J. D., … Scott, J. G. (2011). The RBANS Effort Index: Base rates in geriatric samples. *Applied Neuropsychology, 18*, 11–17.

Duncan, A. (2005). The impact of cognitive and psychiatric impairment of psychotic disorders on the Test of Memory Malingering (TOMM). *Assessment, 12*, 123–129.

Etherton, J. L., Bianchini, K. J., Ciota, M. A., Heinly, M. T., & Greve, K. W. (2006). Pain, malingering and the WAIS-III working memory index. *Spine Journal, 6*, 61–71.

Fahrenberg, J., Hampel, R., & Selg, H. (2001). *Freiburger Persönlichkeitsinventar* (7th ed.). Göttingen, Germany: Hogrefe Verlag.

Franzen, M., Iverson, G. L., & McCracken, L. (1990). The detection of malingering in neuropsychological assessment. *Neuropsychology Review, 1*, 247–279.

Frederick, R. I., & Speed, F. M. (2007). On the interpretation of below-chance responding in forced-choice tests. *Assessment, 14*, 3–11.

Gaffan, D., & Heywood, C. A. (1993). A spurious category-specific visual agnosia for living things in normal humans and nonhuman primates. *Journal of Cognitive Neuroscience, 5*, 118–128.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference* (2nd ed., Texts in Statistical Science Series). New York, NY: Chapman & Hall/CRC.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice* (Interdisciplinary Statistics). London, UK: Chapman & Hall.

Grailet, J. M., Seron, X., Bruyer, R., Coyette, F., & Frederix, M. (1990). Case report of a visual integrative agnosia. *Cognitive Neuropsychology, 7*, 275–309.

Green, P. (2005). *Word Memory Test for Windows: Test manual* (revised June 2005). Edmonton, Alberta, Canada: Green's Publishing.

Greiffenstein, M. F., Greve, K. W., Bianchini, K. J., & Baker, W. J. (2008). Test of Memory Malingering and Word Memory Test: A new comparison of failure concordance rates. *Archives of Clinical Neuropsychology, 23*, 801–807.

Grosz, H. J., & Zimmerman, J. (1965). Experimental analysis of hysterical blindness: A follow-up report and new experimental data. *Archives of General Psychiatry, 13*, 255–260.

Grote, C. L., & Hook, J. N. (2007). Forced-Choice recognition tests of malingering. In G. J. Larrabee (Ed.), *Assessment of malingered neuropsychological deficits* (pp. 44–79). Oxford, New York: Oxford University Press.

Gutiérrez, J. M., & Gur, R. C. (2012). Detection of malingering using forced-choice techniques. In C. R. Reynolds & A. M. Horton Jr (Eds.), *Detection of malingering during head injury litigation* (2nd ed., pp. 151–167). New York, NY: Springer.

Haines, M. E., & Norris, M. P. (1995). Detecting the malingering of cognitive deficits: An update. *Neuropsychology Review, 5*, 125–148.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., & Millis, S. R. (2009). American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 23*, 1093–1129.

Heilbronner, R. L.,Taylor, H. G., Wills, K., Boone, K., Bigler, E., Fortuny, L. A. I., ... Schmidt, M. (2007). American Academy of Clinical Neuropsychology (AACN) practice guidelines for neuropsychological assessment and consultation. *The Clinical Neuropsychologist*, 21, 209–231.

Hiscock, M., & Hiscock, C. K. (1989). Refining the forced-choice method for the detection of malingering. *Journal of Clinical and Experimental Neuropsychology, 11*, 967–974.

Horn, W. (1983). *Leistungsprüfsystem* (2nd ed.). Göttingen, Germany: Hogrefe Verlag.

Hosmer, D., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: Wiley Interscience.

Iverson, G. L. (2007). Identifying exaggeration and malingering. *Pain Practice, 7*, 94–102.

Iverson, G. L. (2010). Detecting exaggeration, poor effort, and malingering in neuropsychology. In A. M. HortonJr & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology* (pp. 91–135). New York, NY: Springer.

Iverson, G. L., Le Page, J., Koehler, B., Shojania, K., & Badii, M. (2007). Test of Memory Malingering (TOMM) scores are not affected by chronic pain or depression in patients with fibromyalgia. *The Clinical Neuropsychologist, 21*, 532–546.

Jarvis, B. G. (2008). *DirectRT (Version v2008.1.13) [Computer Software]*. New York, NY: Empirisoft Corporation.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Clarendon Press.

Jelicic, M., Merckelbach, H., Candel, I., & Geraerts, E. (2007). Detection of feigned cognitive dysfunction using special malingering tests: A simulation study in naïve and coached malingerers. *International Journal of Neuroscience, 117*, 1185–1192.

Joseph, L., Gyorkos, T. W., & Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology, 141*, 263–272.

Larrabee, G. J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist, 17*, 410–425.

Larrabee, G. J. (2007). Malingering, research designs, and base rates. In G. J. Larrabee (Ed.), *Assessment of malingered neuropsychological deficits* (pp. 3–14). Oxford, New York: Oxford University Press.

Larrabee, G. J. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *The Clinical Neuropsychologist, 22*, 666–679.

Larrabee, G. J., & Berry, D. (2007). Diagnostic classification studies and diagnostic validity. In G. J. Larrabee (Ed.), *Assessment of malingered neuropsychological deficits* (pp. 14–26). Oxford, New York: Oxford University Press.

Larrabee, G. J., Greiffenstein, M., Grewe, K., & Bianchini, K. (2007). Refining diagnostic criteria for malingering. In G. J. Larrabee (Ed.), *Assessment of malingered neuropsychological deficits* (pp. 334–372). Oxford, New York: Oxford University Press.

Larrabee, G. J., Millis, S. R., & Meyers, J. E. (2008). Sensitivity to brain dysfunction of the Halstead-Reitan vs an ability-focused neuropsychological battery. *The Clinical Neuropsychologist, 22*, 813–825.

Larrabee, G. J., Millis, S. R., & Meyers, J. E. (2009). 40 plus or minus 10, a new magical number: reply to Russell. *The Clinical Neuropsychologist, 23*, 841–849.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575.

Lee, M. D., & Wagenmakers, E. J. (2010). A course in Bayesian graphical modeling for cognitive science: Unpublished course materials. Retrieved from http://www.ejwagenmakers.com/BayesCourse/BayesBookWeb.pdf (accessed December 14, 2012).

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*, 325–337.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine, 28*, 3049–3067.

MacAllister, W. S., Nakhutina, L., Bender, H. A., Karantzoulis, S., & Carlson, C. (2009). Assessing effort during neuropsychological evaluation with the TOMM in children and adolescents with epilepsy. *Child Neuropsychology, 15*, 521–531.

Margraf, J., & Ehlers, A. (2007). *Beck Angst-Inventar - BAI [Manual und 25 Testbögen]*. Frankfurt, Germany: Harcourt Test Services.

Martins, M., & Martins, I. P. (2010). Memory malingering: Evaluating WMT criteria. *Applied Neuropsychology, 17*, 177–182.

Millis, S. R. (2009). What clinicians really need to know about symptom exaggeration, insufficient effort, and malingering: Statistical and measurement matters. In J. Morgan & J. Sweet (Eds.), *Neuropsychology of malingering casebook* (pp. 21–37). New York, NY: Psychology Press.

Millis, S. R., & Volinsky, C. T. (2001). Assessment of response bias in mild head injury: Beyond malingering tests. *Journal of Clinical and Experimental Neuropsychology, 23*, 809–828.

Mittenberg, W., Patton, C., Canyock, E., & Condit, D. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 24*, 1094–1102.

Mossman, D., Wygant, D. B., & Gervais, R. O. (2012). Estimating the accuracy of neurocognitive effort measures in the absence of a "gold standard". *Psychological Assessment, 24*, 815–822.

O'Hagan, A., & Forster, J. J. (2004). *Kendall's advanced theory of statistics* (Vol. 2B: Bayesian inference (2nd ed.). London, UK: Arnold.

Ortega, A., Wagenmakers, E.-J., Lee, M. D., Markowitsch, H. J., & Piefke, M. (2012). A Bayesian latent group analysis for detecting poor effort in the assessment of malingering. *Archives of Clinical Neuropsychology, 27*, 453–465.

Osterrieth, P. A. (1944). Le test du copie d'une figure complexe. *Archives of Psychology (Chicago), 30*, 206–356.

Pankratz, L., Fausti, S. A., & Peed, S. (1975). Forced-Choice technique to evaluate deafness in hysterical or malingering patient. *Journal of Consulting and Clinical Psychology, 43*, 421–422.

Powell, M. R., Gfeller, J. D., Hendricks, B. L., & Sharland, M. (2004). Detecting symptom- and test-coached simulators with the Test of Memory Malingering. *Archives of Clinical Neuropsychology, 19*, 693–702.

Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathic traumatique. *Archives of Psychology (Chicago), 28*, 286–340.

Rogers, R. (2008). Detection strategies for malingering and defensiveness. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (3rd ed., pp. 14–35). New York, NY: Guilford Press.

Rogers, R., & Cruise, C. R. (1998). Assessment of malingering with simulation designs: Threats to external validity. *Law and Human Behavior, 22*, 273–285.

Rosenfeld, B., Sands, S. A., & Van Gorp, W. G. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Archives of Clinical Neuropsychology, 15*, 349–359.

Slick, D. J., Sherman, E. M. S., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist, 13*, 545–561.

Slick, D. J., Tan, J. E., Strauss, E., Mateer, C. A., Harnadek, M., & Sherman, E. M. (2003). Victoria Symptom Validity Test scores of patients with profound memory impairment: Nonlitigants case studies. *The Clinical Neuropsychologist, 17*, 390–394.

Streiner, D. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment, 81*, 209–219.

Tombaugh, T. N. (1996). *Test of Memory Malingering (TOMM)*. New York, NY: Multi Health Systems.

Walters, G. D., Berry, D. T. R., Lanyon, R. I., & Murphy, M. P. (2009). Are exaggerated health complaints continuous or categorical? A taxometric analysis of the health problem overstatement scale. *Psychological Assessment, 21*, 219–226.

Walters, G. D., Rogers, R., Berry, D. T. R., Miller, H. A., Duncan, S. A., McCusker, P. J., … Granacher, R. P. (2008). Malingering as a categorical or dimensional construct: The latent structure of feigned psychopathology as measured by the SIRS and MMPI-2. *Psychological Assessment, 20*, 238–247.

Warrington, E. K., & Shallice, T. (1984). Category specific impairments. *Brain, 107*, 829–854.

Weinborn, M., Orr, T., Woods, S. P., Conover, E., & Feix, J. (2003). A validation of the Test of Memory Malingering in a forensic psychiatric setting. *Journal of Clinical and Experimental Neuropsychology, 25*, 979–990.

Wolfe, P. L., Millis, S. R., Hanks, R., Fichtenberg, N., Larrabee, G. J., & Sweet, J. J. (2010). Effort indicators within the California Verbal Learning Test-II (CVLT-II). *The Clinical Neuropsychologist, 24*, 153–168.

Appendix 1: post test interview questionnaire

Questions which were presented to all participants

- What do you thing is the purpose of this study?
- Did you notice anything about the tests which you think is worthy to note?
- Do you have any conclusive remarks?

Questions which were presented to participants of the bone fide and clinical group

- Did you give your best while taking the tests?
- How difficult was it for you to accomplish the tests successfully?
- Did you think you achieved good results in the tests?

Questions which were presented to participants of the malingering group

- Did you follow the instructions to simulate cognitive deficits while taking the tests?
- What strategies did you use in order to fake cognitive impairment?
- Do you think these strategies were successful?
- Did you have any problems while applying those strategies?
- Did you have any other problems while feigning impairment e.g., moral doubts?
- Despite the role instruction you had, did you knew the right answers? How would you rate the level of difficulty of these tests?

APPENDIX 2: NEUROPSYCHOLOGICAL ASSESSMENT

Table 2A. Neuropsychological test results and comparisons between groups for cognitive measures

	D2			LPS Speech			LPS Analytic			Rey Figure		
Groups	Mean	*SD*	*p*-Value	Mean	*SD*	*p*-Value	Mean	*SD*	*p*-Value	Mean	*SD*	*p*-Value
Honest response vs.	102.55	13.31	*ns*	56.17	5.73	*ns*	50.50	8.87	<.01	52.35	7.55	*ns*
Malingering	107.45	16.43		57.08	4.18		62.50	4.44		53.45	6.77	
Honest response vs.	102.55	13.31	<.01	56.17	5.73	<.01	50.75	8.93	*ns*	52.35	7.55	*ns*
Clinical	87.15	13.67		48.92	6.36		50.50	8.87		51.45	8.22	
Clinical vs.	87.15	13.67	<.01	48.92	6.36	<.01	50.75	8.93	<.01	51.45	8.22	*ns*
Malingering	107.45	16.43		57.08	4.18		62.50	4.44		53.45	6.77	

Note: *ns* = non significant; D2 = D2 test of attention "concentration score" (sw-scale); LPS Speech = Leistungsprüfsystem, mean score of the subscales 1, 2, 5, and 12 (t-scale); LPS Analytic = Leistungsprüfsystem, subtest 4 (t-scale); Rey Figure = Rey–Osterrieth Figure, CQM score (t-scale). *p*-Values are derived from ANOVA *post-hoc* analysis of the Tukey HSD test after revealing significant overall group differences in the D2 and the LPS.

Table 2b. Neuropsychological test results for personality and clinical measures

Group	Descriptive Statistics	FPI Satisfaction	FPI Social	FPI Achievement	FPI Shyness	FPI Irritability	FPI Aggressiveness	FPI Demandedness	FPI Physical	FPI Health	FPI Openness	FPI Extraversion	FPI Emotions	BDI	BAS
Honest response	Mean	4.85	6.90	4.40	5.10	4.90	3.70	4.65	5.10	4.45	5.20	4.80	5.10	5.10	5.90
	SD	1.76	1.68	1.57	2.22	2.10	1.75	1.63	1.65	1.50	1.40	1.88	1.62	4.85	4.29
Clinical	Mean	4.60	5.30	5.45	5.10	4.85	4.55	5.30	5.35	4.10	4.70	4.80	5.15	9.40	7.90
	SD	2.09	1.69	1.90	1.71	1.63	2.14	1.56	1.79	2.00	2.00	1.44	1.76	7.73	7.57
Malingering	Mean	4.70	7.00	5.50	4.85	4.95	4.25	5.10	4.60	4.45	5.70	5.50	5.30	7.15	8.55
	SD	2.36	1.62	1.54	1.81	2.16	1.59	1.77	1.98	2.19	1.87	1.85	1.53	9.45	7.25

Note: FPI = Freiburger Persönlichkeitsinventar (stanine-scale); BDI = Beck Depression Inventory (raw scores); BAS = Beck Anxiety Scale (raw scores). ANOVA mean comparisons between groups were not significant for any measure and hence no inferential statistics are presented.