Journal of Applied Research in Memory and Cognition 2 (2013) 149–157

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Target article

Simple rules for detecting depression☆

Mirjam A. Jenny[a,*], Thorsten Pachur[a], S. Lloyd Williams[b], Eni Becker[c], Jürgen Margraf[b]

[a] Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany
[b] Ruhr University Bochum, Department of Clinical Psychology and Psychotherapy, Universitätsstrasse 150, 44780 Bochum, Germany
[c] Radboud University Nijmegen, Faculty of Social Sciences, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 6 December 2012
Received in revised form 5 June 2013
Accepted 16 June 2013
Available online 24 June 2013

Keywords:
Depressed mood
Fast and frugal trees
Medical decision making
Screening
Beck Depression Inventory

ABSTRACT

Depressive disorders are major public health issues worldwide. We tested the capacity of a simple lexicographic and noncompensatory *fast and frugal tree* (FFT) and a simple compensatory unit-weight model to detect depressed mood relative to a complex compensatory logistic regression and a naïve maximization model. The FFT and the two compensatory models were fitted to the Beck Depression Inventory (BDI) score of a representative sample of 1382 young women and cross validated on the women's BDI score approximately 18 months later. Although the FFT on average inspected only approximately one cue, it outperformed the naïve maximization model and performed comparably to the compensatory models. The heavier false alarms were weighted relative to misses, the better the FFT and the unit-weight model performed. We conclude that simple decision tools—which have received relatively little attention in mental health settings so far—might offer a competitive alternative to complex weighted assessment models in this domain.

1. Introduction

Clinical depression, which is characterized by sadness and loss of interest, affects approximately 1.9–12% of the world's population (Andrade et al., 2003; Kessler et al., 2003; World Health Organization, 2001). Depression can lead to a reduced quality of life and productivity (World Health Organization, 2001), harm the immune system (Herbert & Cohen, 1993), and increase stroke and suicide mortality (Everson, Roberts, Goldberg, & Kaplan, 1998; Inskip, Harris, & Barraclough, 1998).

Given the risks associated with untreated depressive disorders, it is important to have tools available to detect them early and reliably, based on the available indicators or cues (e.g., crying, feeling hopeless). Such detection tools should be simple to keep the extent and cost of the assessment procedure to a minimum. They also should be user friendly, allowing professionals without specific training—such as general practitioners, who are often the first point of contact for people with symptoms of depression, and non-experts (e.g., school or military officials)—to screen certain populations for depression.

In this article, we examine the capacity of simple decision models to detect depressed mood as assessed by the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), a commonly used screening tool. Specifically, we compare a fast and frugal tree (FFT)—which often limits information search—with two compensatory models, namely a unit-weight model and a logistic regression model, as well as with a simple baseline model. While compensatory models integrate all of the available information to make a categorization, noncompensatory models such as FFTs can decide on the basis of a single piece of information.

2. The role of FFTs in assessing health in medical settings: is more always better?

In medical decision making, errors in diagnosing a patient's health status can have severe and possibly lethal consequences. This may explain why doctors tend to gather more rather than less information when making decisions. But is more information always better? Green and Mehr (1997) suggest that this may not always be the case. They compared a simple decision tree for deciding whether to send a patient suffering from chest pain to the

☆ We thank Jan Woike and Michael Lee for discussing methodological aspects of our analyses, Noortje Vriends for her input on the clinical practice of assessing clinical depression, Eva-Lotta Brakemeier and Laura Martignon for comments on an earlier version of this paper, and Laura Wiles and Susannah Goss for editing the manuscript. Mirjam Jenny was supported by Swiss National Science Foundation Grant 100014_138174/1 granted to Jörg Rieskamp, and Thorsten Pachur was supported by Grant HE 2768/7-1 from the German Research Foundation (DFG) as part of the priority program on New Frameworks of Rationality (SPP 1516).

* Corresponding author.
E-mail addresses: jenny@mpib-berlin.mpg.de (M.A. Jenny), pachur@mpib-berlin.mpg.de (T. Pachur), s.lloyd.williams@rub.de (S. Lloyd Williams), e.becker@psych.ru.nl (E. Becker), juergen.margraf@ruhr-uni-bochum.de (J. Margraf).

2211-3681/$ – see front matter © 2013 Society for Applied Research in Memory and Cognition. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jarmac.2013.06.001

coronary care unit with a more complex statistical method and found that both methods were equally able to differentiate between patients with and without a heart attack. Relatedly, Fischer et al. (2002) found that a simple decision tree was able to compete with a complex regression-based method for assessing children's risk of pneumonia.

Does the potential of simple decision models extend to the mental health domain? Mental health is usually assessed using extensive procedures, such as lengthy structured interviews (Margraf, Schneider, Soeder, Neumer, & Becker, 1996). One frequently used tool to screen for depressed mood is the BDI (Steer, Cavalieri, Leonard, & Beck, 1999), which encompasses 21 questions. Because practitioners have difficulty remembering all criteria of major depression, there have been calls for shortened manuals (Bowers, Jorm, Henderson, & Harris, 1992; Krupinski & Tiller, 2001). The performance of such shortened procedures (e.g., Margraf, 1994) sometimes converges with that of more extensive procedures (Zimmerman et al., 2010). Whooley, Avins, Miranda, and Browner (1997) compared a simple two-question instrument with more complex approaches, and found that the simple instrument had similar (or even higher) discriminability in detecting depression.

Developing simple and robust decision methods has also been a key endeavor in decision science. In his seminal work, Dawes (1979; Dawes & Corrigan, 1974; see also Einhorn & Hogarth, 1975) showed that simple unit-weight models (which consider only the sign of a cue and weight all cues equally) often outperform regression models (which have differential weights) in prediction. More recently, Gigerenzer and colleagues (Gigerenzer & Goldstein, 1996; Gigerenzer, Hertwig, & Pachur, 2011; Gigerenzer, Todd, & the ABC Research Group, 1999; Katsikopoulos, 2010; Pachur, 2010; Pachur, Hertwig, & Rieskamp, in press) demonstrated that robustness in prediction can also be achieved by lexicographic and noncompensatory mechanisms with simple search, stopping, and decision rules. Moreover, simple lexicographic strategies are often used in professional decision making (Garcia-Retamero & Dhami, 2009; Pachur & Marinello, 2013). In lexicographic mechanisms, cues are inspected following a specific hierarchy, usually defined by cue validity or "diagnosticity." Search is stopped and a decision is made as soon as the inspected cue has a particular value. These mechanisms are *noncompensatory* because once information search is stopped, cues lower in the cue hierarchy cannot compensate for cue information further up in the hierarchy. Cue values are neither weighted nor added.

We examined the capacity of FFTs—simple categorization mechanisms that have received considerable attention in decision research (Gigerenzer & Selten, 2001; Luan, Schooler, & Gigerenzer, 2011; Martignon, Katsikopoulos, & Woike, 2008, 2012), but have not been tested in the mental health domain—to detect depressed mood. FFTs can be surprisingly accurate. In computer simulations based on 30 real-world data sets, Martignon et al. (2008) found their performance to be comparable with that of logistic regression and complex classification trees. FFTs are effective because they refrain from differentially weighting all pieces of information and instead focus on a few best pieces of information. This makes them less susceptible to overfitting—that is, they adjust less to unsystematic variability in the data and thus perform better when applied to new data than do methods that differentially weight many pieces of information (Gigerenzer & Brighton, 2009; Katsikopoulos, 2011; Luan et al., 2011; Martignon et al., 2008). Additionally, FFTs may be more readily accepted by clinicians as decision aids than more complex methods (Adams & Leveson, 2012; Elwyn, Edwards, Eccles, & Rovner, 2001; Katsikopoulos, Pachur, Machery, & Wallin, 2008). In sum, because of their accuracy, transparency, accessibility, and simplicity, FFTs could prove to be promising detection models in the domain of mental health (see Marewski & Gigerenzer, 2012, for a thorough, nontechnical discussion of these issues).

Below, we construct an FFT for detecting depressed mood (as assessed by the BDI) and compare its performance with that of a logistic regression model, which is frequently used as a benchmark for categorization (Dhami & Harries, 2001; Kee et al., 2003; Smith & Gilhooly, 2006), a unit-weight model (Dawes, 1979; Einhorn & Hogarth, 1975), and a naïve maximization model (which predicts all cases to be nondepressed and serves as a baseline model). We conducted this prescriptive test using cross validation—that is, we fitted the models to one data set and then tested their accuracy in predicting outcomes in another.

Our analysis contributes to the literature in several ways. First, although FFTs have been tested as descriptive models (i.e., how well they can describe people's decisions; Dhami & Ayton, 2001; Dhami & Harries, 2001; Kee et al., 2003; Smith & Gilhooly, 2006; Snook, Dhami, & Kavanagh, 2011), only a few investigations have subjected them to prescriptive testing (Fischer et al., 2002; Green & Mehr, 1997; Martignon et al., 2008). In this study, we investigated whether the advantages of using FFTs in medical decision making discussed above also holds for predicting depression. Second, to date, the studies by Martignon et al. (2008) and Luan et al. (2011) are the only ones to have tested FFTs by means of out-of-sample prediction (rather than fitting). Finally, we examine—to our knowledge for the first time—the performance of the various models under different weighting schemes of misses and false alarms using real-world data.

3. Overview of the present study

To this end, we drew on data from the Dresden Predictor Study (Trumpf et al., 2010), a prospective epidemiological study in which a representative sample of young women was surveyed twice at an 18-month interval. The FFT, the logistic regression model, and the unit-weight model were provided with data on a set of five BDI items obtained at the first time point ($t1$) and fitted to the women's depression status according to the full BDI at $t1$. The crucial test was how well the models, using the women's responses to the five BDI items at the second time point ($t2$), would be able to detect depressed mood according to the BDI at $t2$. The correlation of the women's depression status (i.e., the criterion values) between $t1$ and $t2$ was $\Phi = .34$; that of the cue values was $\Phi = .31$. Thus, although the $t1$ and $t2$ data sets stemmed from the same group of people, there was nevertheless considerable variability across the two time points, making the data a good test bed for cross validation (which is often used in judgment and decision making research; e.g., Glöckner & Pachur, 2012).

4. Method

The Dresden Predictor Study data set (Trumpf et al., 2010) comprises a representative sample of young women (average age = 18.8 years, range 18–25 years) from the general population. Respondents were recruited using an unweighted random sampling procedure (see Trumpf et al., 2010, for details). The BDI was completed by 1382 respondents at $t1$ and by 1392 respondents at $t2$ (approximately 18 months later). All but 10 of the respondents at $t2$ also completed the BDI at $t1$.

4.1. Criterion

The models inferred the "depressed mood" status of each respondent, as determined by the full BDI. A respondent was considered as having clinically depressed mood if her BDI score (based on all 21 items) exceeded 17 points (which, according to Beck, Steer, & Garbin, 1988, indicates mildly to severely depressed mood; the maximum number of points is 63). Based on this definition, which was also used in the original Dresden Predictor Study, the base rate of participants with depressed mood was 3.6% (50 cases) at *t*1 and 1.9% (26 cases) at *t*2.[1]

4.2. Cues

We constructed an FFT, a unit-weight model, and a logistic regression model, that categorized a respondent as having depressed mood (or not) based on her responses to items from the German version of the BDI (Hautzinger, 1991). The BDI consists of 21 items, each asking the respondent to indicate which of four statements best describes how she has felt in the last 7 days. Each statement is associated with a certain number of points. The sum of the points of the chosen statements provides the total BDI score, with higher values indicating a higher degree of depressed mood. A sample item reads (a) "I don't feel disappointed in myself" (0 points), (b) "I am disappointed in myself" (1 point), (c) "I am disgusted with myself" (2 points), and (d) "I hate myself" (3 points). Because FFTs are defined to take binary data as input (Luan et al., 2011), we binarized all items using a median split, with higher values coded as 1 and lower values as 0. The median value of each cue was 0.

To determine the maximum number of cues for which weights in a logistic regression model could be estimated reliably, we used the formula developed by Peduzzi, Concato, Kemper, Holford, and Feinstein (1996) $k = Np/10$, where k is the maximum number of cues, N is the number of data points, and p the proportion of positive (i.e., depressed) cases in the data set. With 1382 respondents and a proportion of .036 depressed cases, we obtain $k = 1382 \times 0.036/10 = 4.98$ cues. To ensure comparability, we therefore set the maximum number of cues for all three fitted models to five. We selected those five cues from the full set of 21 BDI items by determining each item's correlation (based on the Φ coefficient) with the BDI score (at *t*1; see below for details) and picking those five items with the highest correlations. The correlations of the five items selected were all similarly high (see Table 1).

4.3. Description of the models

4.3.1. Fast and frugal tree

The FFT was constructed using the *Max* procedure proposed by Martignon et al. (2008). This procedure produces trees that have a bias to categorize cases into the same category. Given that most respondents in our sample did not have clinically depressed mood, using Max seems appropriate. Accordingly, we calculated the *positive validity* of each of the five cues in Table 1, defined as the cue's ability to infer cases with depressed mood according to the BDI:

$$\text{Positive validity} = \frac{H}{H + FA}, \tag{1}$$

as well as each cue's *negative validity*, defined as the cue's ability to infer cases with no depressed mood:

$$\text{Negative validity} = \frac{CR}{CR + M}, \tag{2}$$

where H is the number of hits (cases correctly categorized as depressed), FA is the number of false alarms (cases incorrectly categorized as depressed), CR is the number of correct rejections (cases correctly categorized as not depressed), and M is the number of misses (cases incorrectly categorized as not depressed). To determine the cue hierarchy (the order in which the cues are processed), we used the positive or negative validity of each cue—depending on which was higher—and ordered the cues in descending order. Next, we determined exit locations on each level of the tree. If, on a given level in the cue hierarchy, the positive validity was higher than the negative validity, then the tree was exited given a positive value on that cue, leading to the decision "clinically depressed mood"; if the negative validity was higher than the positive validity, then the tree was exited given a negative value on that cue, leading to the decision "not clinically depressed mood."

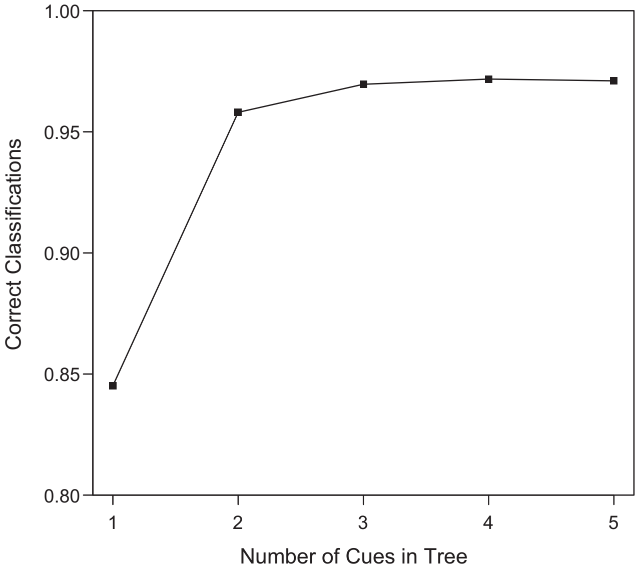

Fig. 1. Accuracies for fast and frugal trees of different tree lengths at *t*1.

The validity of each of the five cues is reported in Table 1. Because most respondents in our sample did not have depressed mood, the negative validity exceeded the positive validity for all cues. To determine the number of cues in the FFT, we constructed trees consisting of different numbers of cues and compared their accuracies, defined as the percentage of correctly categorized cases. As Fig. 1 shows, a four-cue tree had the highest accuracy, at 97.1%.

The FFT used in our subsequent analyses is depicted in Fig. 2. As an example, if a person answers "yes" on the first item ("Have you cried more than usual within the last week?") and "no" to the second item ("Have you been disappointed in yourself or hated yourself within the last week?"), the tree categorizes this person as not having clinically depressed mood. All subsequent items are then ignored. In other words, although the tree consists of multiple cues, it can stop cue inspection at any level (as there is at least one exit on each level).

The resulting tree categorized a person as having clinically depressed mood if she answered "yes" to every question. In principle, this FFT makes the same predictions as a unit-weight model with four cues that categorizes a case as depressed when all four cues have positive values. In contrast to a unit-weight model (which sums up all cue values and makes a categorization based on this sum), however, the FFT is sensitive to cue order, with the order affecting how many cues the FFT has to inspect to make a decision (Luan et al., 2011). Note that the FFT and the unit-weight model we tested do not make the same predictions, because they consist of different numbers of cues (four and five, respectively).

4.3.2. Unit-weight model

As a simple compensatory model, we tested a unit-weight model, which categorizes a person based on the five cues in Table 1 without weighting them differentially. Unit-weight models have been shown to be less susceptible to overfitting than models with

[1] These prevalence rates lie within the range of 1.9–12.0% commonly observed in nonclinical populations (Andrade et al., 2003; Kessler et al., 2003; World Health Organization, 2001).

Table 1
The five items of the Beck Depression Inventory (BDI) used to infer the respondents' depression status (according to the full BDI), ordered by validity.

BDI #	Content of item	Φ[a]	Validity[b]	Regression weight [HDI][c]
10	0 I don't cry any more than usual. 1 I cry now more than I used to. 2 I cry all the time now. 3 I used to be able to cry, but now I can't cry even though I want to.	0.358	0.996 (−)	3.81 [2.58, 5.01]
7	0 I don't feel disappointed in myself. 1 I am disappointed in myself. 2 I am disgusted with myself. 3 I hate myself.	0.351	0.992 (−)	2.26 [1.20, 3.44]
2	0 I am not particularly discouraged about the future. 1 I feel discouraged about the future. 2 I feel I have nothing to look forward to. 3 I feel that the future is hopeless and that things cannot improve.	0.352	0.988 (−)	1.77 [0.93, 2.87]
3	0 I do not feel like a failure. 1 I feel that I have failed more than the average person. 2 As I look back on my life, all I can see is a lot of failures. 3 I feel that I am a complete failure as a person.	0.357	0.986 (−)	1.49 [0.47, 2.43]
12	0 I have not lost interest in other people. 1 I am less interested in other people than I used to be. 2 I have lost most of my interest in other people. 3 I have lost all my interest in other people.	0.364	0.984 (−)	3.05 [1.99, 4.02]

[a] Correlation with the BDI score (as determined by the full BDI) at $t1$.
[b] Validity was computed using Eq. (2).
[c] We report the mean posteriors. 95% highest density intervals (HDIs) indicate "the interval that contains 95% of the distribution such that all points inside the interval have higher believability than points outside the interval" (Kruschke, 2010, p. 665). The constant in the regression model was −8.82 [HDI: −10.40,−7.04].

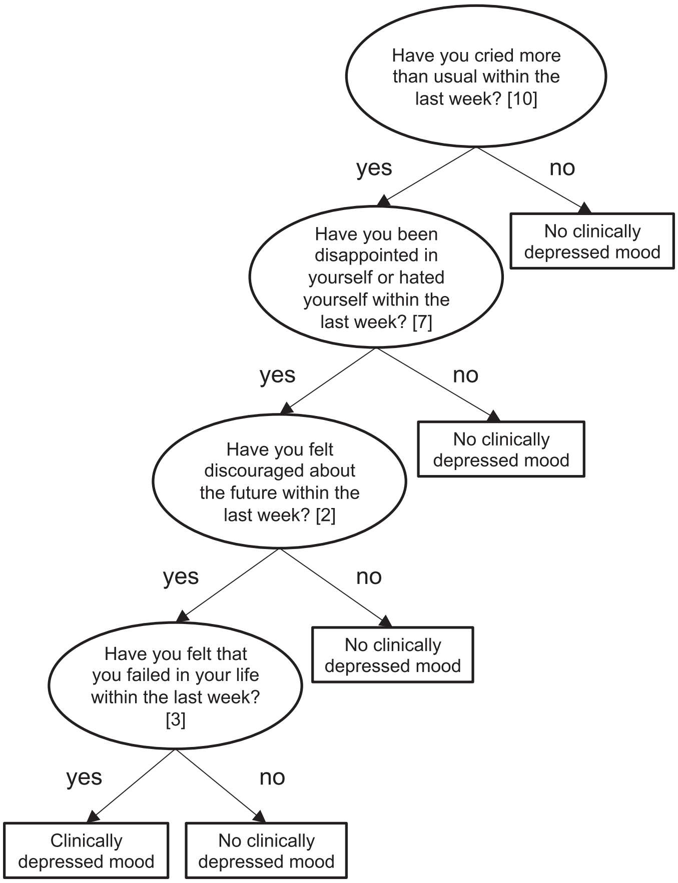
Have you cried more than usual within the last week? [10]
yes
no
No clinically depressed mood
Have you been disappointed in yourself or hated yourself within the last week? [7]
yes
no
No clinically depressed mood
Have you felt discouraged about the future within the last week? [2]
yes
no
No clinically depressed mood
Have you felt that you failed in your life within the last week? [3]
yes
no
Clinically depressed mood
No clinically depressed mood

Fig. 2. Fast and frugal tree screening for depressed mood (according to the total BDI score). The numbers in brackets indicate the position of the respective item in the BDI. The full wording of the BDI cues is presented in Table 1; for this figure, it has been translated into binary questions.

differential weighting (Dawes, 1979; Einhorn & Hogarth, 1975). To implement the unit-weight model, we used the sum of the five cue values as a single cue (plus an intercept) to predict the women's depression status using logistic regression (Armstrong & Cuzan, 2006). Thus, this model determined one coefficient for the sum of the cues. To estimate the regression weights, we used a Bayesian approach.[2] Fitted to the data at $t1$, the unit-weight model categorized a person as having depressed mood if her sum score on the five items was ≥ 4.

4.3.3. *Logistic regression*

The logistic regression model predicts the probability that a woman suffers from depressed mood based on a weighted integration of all five cues. A woman with a predicted probability > .5 was categorized as suffering from depressed mood. We again used a Bayesian approach to estimate the regression weights. Table 1 shows the resulting regression weights for each of the five cues when the regression model was fitted to the data at $t1$. Note that, by definition, the logistic regression model always considers all five cues.

4.3.4. *Naïve maximization model*

Given the low rate of depressed cases, categorizing all cases as nondepressed could lead to a rather high accuracy. As a benchmark, we therefore examined a model that does not use any information about the individual respondents, but only information about the majority status in the sample. As most respondents in our sample were nondepressed, this model categorizes all respondents as nondepressed. This naïve base-rate prediction represents the optimal (i.e., maximizing) strategy in the absence of individual cue information (Shanks, Tunney, & McCarthy, 2002; Vulkan, 2002).

In summary, we fitted an FFT, a unit-weight model, and a logistic regression model to the women's depressed mood status (according to the full BDI) at $t1$. All models used the respondents' responses

[2] Specifically, we fitted a Bayesian logistic regression model using a random walk metropolis algorithm and uniform priors, as implemented in the R-package MCMCpack.

to a set of BDI questions as cues. We also tested a naïve maximization model that predicted all respondents to be nondepressed. The key question was how well the different models fitted to *t*1 would generalize—that is, how well they would be able to infer a respondent's "depressed mood" status according the full BDI at *t*2 based on her responses to the five BDI items at *t*2.

5. Results

The FFT proved to be highly frugal. Across all categorizations, the tree stopped search after inspecting an average of just 1.3 (*SD* = 0.66) and 1.2 (*SD* = 0.69) cues at *t*1 and *t*2, respectively.[3] In contrast, the logistic regression model and the unit-weight model by definition used all five cues for every categorization.

Did the latter two models' taking a greater amount of information into account render better performance? To address this question, we evaluated all models using the signal detection theory (SDT) framework (Green & Swets, 1966; Macmillan & Creelman, 2004). Specifically, we determined the hit and false alarm rates of all models and calculated their *discriminability*, as measured by the index *d*′. The SDT framework also allowed us to determine the *bias* (or decision threshold) of a decision model, as measured by the index *c* (Stanislaw & Todorov, 1999). Both *d*′ and *c* are calculated based on hit rates (*HR*; defined as the proportion of cases correctly categorized as having depressed mood) and false alarm rates (*FAR*; defined as the proportion of cases incorrectly categorized as having depressed mood) as follows:

$$d' = z(HR) - z(FAR) \tag{3}$$

and

$$c = \frac{-(z(HR) + z(FAR))}{2}. \tag{4}$$

d′ quantifies a model's ability to discriminate between signal (depressed mood) and noise (not depressed mood) cases. *c* measures the tendency to make a "signal" or a "noise" decision. A positive value of *c* indicates a conservative decision threshold for making a "signal" decision; a negative value of *c* indicates a lenient threshold. In situations with very low or very high hit and false alarm rates, standard procedures can yield unreliable estimates for *d*′ and *c* (Lee, 2008). Although corrective procedures have been proposed for such situations (Snodgrass & Corwin, 1988), these procedures are based on debatable statistical assumptions. We therefore estimated the SDT indices using a Bayesian approach (Lee, 2008), which is robust in situations with extreme hit and false alarm rates. Additionally, we calculated each model's accuracy, defined as the percentage of correctly categorized cases, using a Bayesian approach (with uniform beta distributed priors).

For all measures, we calculated mean posteriors and 95% highest density intervals (HDI; for an introduction, see Kruschke, 2010). A 95% HDI contains 95% of the posterior distribution, and all values inside the HDI have higher credibility than values outside the HDI. To evaluate the difference between two models, we subtracted each value of the posterior distribution of one model from the corresponding value in the other model's posterior distribution, and calculated the 95% HDI of this difference distribution. It is credible that there is a difference between two models if this distribution does not span the value of 0 (Kruschke, 2011)—in other words, if the HDI includes either only positive or only negative values.

Table 2 reports the accuracy, hit rates, and false alarm rates of all models, and Fig. 3 shows their discriminability and decision threshold, separately for fitting (*t*1) and cross validation (*t*2). In comparing the models, we focus on accuracy and discriminability at cross validation. As Table 2 shows, the unit-weight model achieved the highest accuracy at cross validation, followed by the FFT, the naïve maximization model, and the logistic regression model. Based on the 95% HDIs in Table 3, however, the only credible differences between the models at cross validation were that the unit-weight model and the FFT outperformed the logistic regression model as well as the naïve maximization model (these differences were small, however, as the HDIs bordered on the value of zero). Comparing the models in terms of discriminability yielded a similar pattern: As shown in Fig. 3, the unit-weight model achieved the highest *d*′, followed by the FFT, the logistic regression model, and the naïve maximization model. Based on the 95% HDIs in Table 3, it is highly credible that the unit-weight model outperformed the logistic regression model and the naïve maximization model, and that the FFT and the logistic regression model outperformed the naïve maximization model. There were no credible differences between either the unit-weight model and the FFT, or the FFT and the logistic regression model. Moreover, note that for both the unit-weight model and the FFT, *d*′ increased somewhat between fitting and cross validation, whereas for the logistic regression model, *d*′ decreased at cross validation, indicating overfitting. Overall, these analyses show that although the FFT inspected only about one fourth of the information inspected by the compensatory models, it was able to compete with the latter in cross validation. (In Online Appendix B, we report additional analyses showing that similar conclusions hold when the base rate of critical cases is considerably higher.)

The right panel of Fig. 3 shows the bias of the different models. The logistic regression model and the unit-weight model were most lenient in categorizing a respondent as having clinically depressed mood, whereas the FFT was more conservative. By definition, the naïve maximization model (not shown in Fig. 3 as $c = +\infty$) showed the most conservative bias. (More detailed analyses of the models' bias can be found in Online Appendix A.)[4]

5.1. Differential weighting of false alarms and misses

The standard SDT indices (implicitly) assume that the costs of false alarms and misses are weighted equally. In clinical practice, however, this assumption may be inappropriate. To illustrate, a person erroneously categorized as not being depressed might not receive adequate treatment. It is therefore possible to argue that the focus should be on increasing the number of hits (and thus decreasing the proportion of misses), even if it increases the number of false alarms. Conversely, a person erroneously categorized as depressed may undergo unnecessary treatment. It is therefore also possible to argue that the focus should be on avoiding false alarms.

Does the relative performance of the models differ as a function of how false alarms and misses are weighted? To address this question, we calculated the discriminability (at cross validation) of all models for different relative weights of misses and false alarms. Specifically, with $d' = w \times z(\mathrm{HR}) - (2 - w) \times z(\mathrm{FAR})$ (cf. Eq. (3)), we

[3] Additional analyses showed that the average number of cues used in a tree depends on the base rate of critical cases. As reported in Online Appendix B, in an environment with a considerably higher base rate of critical cases, the tree stopped search after inspecting approximately 3 cues.

[4] We also conducted a *k*-fold cross validation with *k* = 10 (Breiman & Spector, 1992; Hastie, Tishirani, & Friedman, 2009; Kohavi, 1995), separately for t1 and t2. For that purpose, the data set was divided into *k* = 10 bins, the models fitted on 9 of them (randomly selected) and cross-validated on the 10th bin, thus ensuring that fitting and cross validation were conducted on strictly different samples. To obtain stable results, this procedure was repeated 100 times. In this analysis, the average *d*′s of all three tools were even less distinguishable (which probably reflects that the sample size for cross validation is considerably smaller than in the analysis using the data at t2 for cross validation).

Table 2
Performance of the fast and frugal tree (FFT), the logistic regression model (LR), the unit-weight model (UW), and the naïve maximization model (NM) in inferring respondents' depression status (according to the full BDI). 95% highest density intervals are reported in brackets.

Model	Accuracy		Hit rate		False alarm rate	
	Fitting	Cross validation	Fitting	Cross validation	Fitting	Cross validation
FFT	97.04% [96.13, 97.90]	98.57% [97.93, 99.16]	0.231 [0.123, 0.346]	0.464 [0.283, 0.643]	0.002 [0.000, 0.004]	0.004 [0.001, 0.008]
LR	97.47% [96.63, 98.27]	97.56% [96.74, 98.34]	0.692 [0.567, 0.814]	0.571 [0.390, 0.747]	0.015 [0.009, 0.022]	0.017 [0.010, 0.024]
UW	97.47% [96.63, 98.27]	98.92% [98.38, 99.44]	0.577 [0.443, 0.708]	0.750 [0.589, 0.896]	0.011 [0.005, 0.016]	0.007 [0.003, 0.011]
NM	96.31% [95.31, 97.28]	98.06% [97.33, 98.76]	0	0	0	0

Note. The table shows the accuracy, hit rates, and false alarm rates at *t*1 (fitting) and *t*2 (cross validation) and reports mean posteriors.

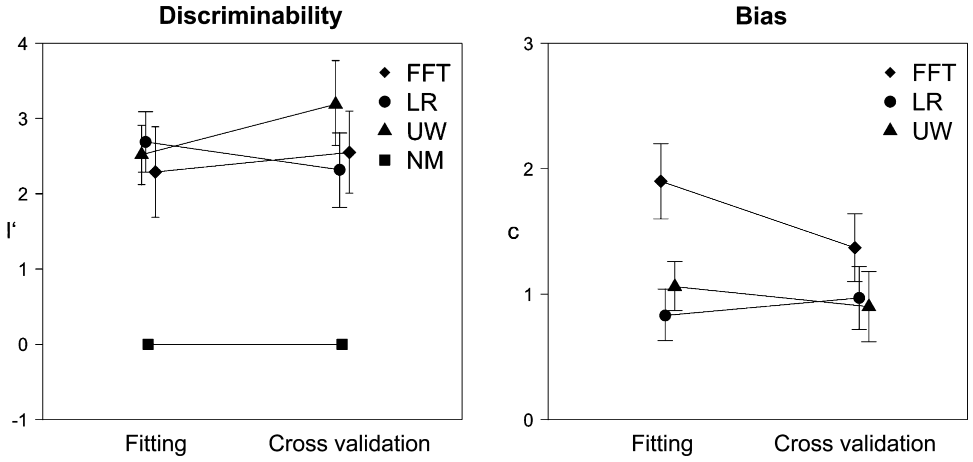

Fig. 3. Discriminability (d') and bias (c) of the models at fitting ($t1$) and cross validation ($t2$). FFT = fast and frugal tree, LR = logistic regression model, UW = unit-weight model, NM = naïve maximization model. Error bars indicate the 95% highest density intervals.

Table 3
95% highest density intervals (HDIs) for the differences between the logistic regression model (LR), the unit-weight model (UW), the fast and frugal tree (FFT), and the naïve maximization model (NM) on categorization accuracy and discriminability (d'). The table shows the HDIs when comparing the model in the row with the model in the column.

	Accuracy				d'			
	LR	UW	FFT	NM	LR	UW	FFT	NM
Fitting								
LR	–	[−0.01, 0.01]	[−0.02, 0.01]	[0, 0.02]	–	[−0.39, 0.73]	[−0.33, 1.11]	[2.29, 3.09]
UW		–	[−0.01, 0.02]	[0, 0.02]		–	[−0.49, 0.95]	[2.12, 2.91]
FFT			–	[−0.01, 0.02]			–	[1.69, 2.89]
NM				–				–
Cross validation								
LR	–	[−0.02, 0]	[−0.02, 0]	[−0.02, 0.01]	–	[−1.63,−0.12]	[−0.98, 0.50]	[1.82, 2.82]
UW		–	[0, 0.01]	[0, 0.02]		–	[−0.14, 1.43]	[2.63, 3.76]
FFT			–	[0, 0.01]			–	[2.01, 3.10]
NM				–				–

varied the value of w to obtain ratios of misses and false alarms ranging from 1:5 to 5:1.[5] To illustrate, with $w = 1$, the ratio is 1:1 and misses (note that the miss rate is the complement of the hit rate) and false alarms are weighted equally; with $w = 1/2$, the ratio is 1:3 and false alarms are given three times as much weight as misses; with $w = 3/2$, the ratio is 3:1 and misses are given three times as much weight as false alarms. When calculating false alarm rates and hit rates, we applied the correction proposed by Snodgrass and Corwin (1988), which has been advocated in situations with extreme hit and false alarm rates (close to 0 or 1; Schooler & Shiffrin, 2005).

Fig. 4 shows the weighted d's for the FFT, the logistic regression, the unit-weight model, and the naïve maximization model across different ratios of misses to false alarms. As can be seen, discriminability of all models is clearly affected by the differential weighting of misses and false alarms: the greater the weight given to false alarms, the better the performance of all models. Differential weighting also influences relative performance. For instance, with increasing weight on false alarms, the discriminability of the FFT converges with that of the unit-weight model. This is because the unit-weight model tends to produce more false alarms than the FFT (HDI$_{\text{difference}}$: [−0.01, 0]; see Table 2). If false alarms are weighted very heavily, the naïve maximization model performs

[5] We used the standard approach to calculate d' for this analysis, as it is currently unclear how to integrate differential weighting of false alarms and misses using the Bayesian approach proposed by Lee (2008).

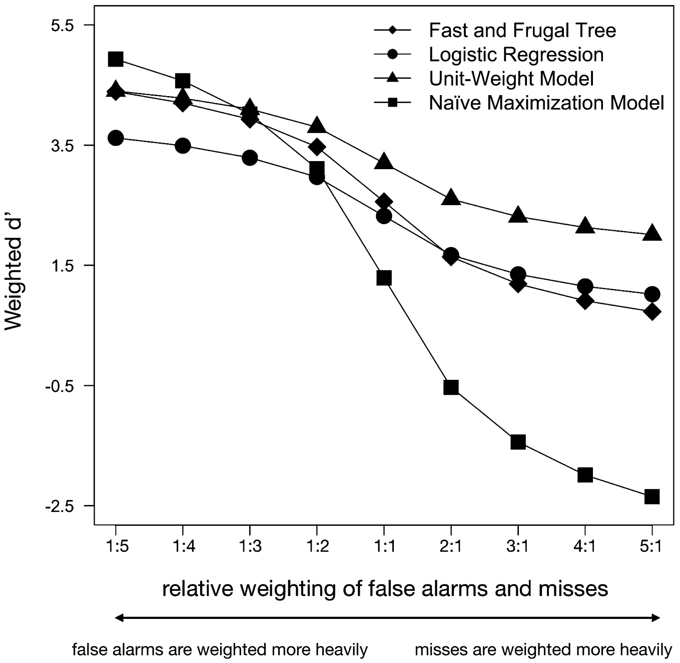

Fig. 4. Results of the weighted d' analysis (for cross validation). The figure shows the models' discriminability (d') for different values of w, reflecting the relative weighting of false alarms and misses.

best overall. With increasing weight on misses, the unit-weight model's edge over the FFT and the naïve maximization model increases, and the FFT performs increasingly worse than the logistic regression. Reconsidering the right panel of Fig. 3, we see that the FFT had the most conservative bias in cross validation, which caused the tree to miss a number of depressed cases, explaining its decreasing relative performance with increased weight on misses. The difference in performance between the unit-weight model and the logistic regression is only slightly affected by the relative weighting of false alarms and misses.

6. Extensions and practical applications

In screening settings, such as in the population examined by general practitioners—who are often the first point of contact for patients with depressive symptoms—base rates of depression are low (Sharp & Lipsky, 2002). The same holds for mental health screening in schools (Barrera & Garrison-Jones, 1988; Jaycox, Reivich, Gillham, & Seligman, 1994; Lewinsohn, Hops, Roberts, Seeley, & Andrews, 1993; Reynolds, 1986; Roberts, Lewinsohn, & Seeley, 1991), in the screening of people recruited for scientific studies, and in the military (Engel et al., 2008; Miller, 2001). As our models were fitted to a data set with a low base rate, our results have implications for application in those settings, in particular. Future studies should examine the generalizability of our findings to clinical samples, where the base rate of depression is considerably higher. Possibly, different algorithms (e.g., an FFT with a different cue hierarchy, a different exit structure, or different cues) are applicable here. In Online Appendix B, however, we show that the result of our model comparison is only slightly affected by a higher base rate of critical cases.

Another extension could be to examine FFTs in a mixed or a male sample, in which the tree might consist of different questions than those used for our female sample as well as in different age groups. Finally, the efficiency and accuracy of simple detection models should also be examined for other mental disorders. Depending on the contexts in which simple models prove to be competitive with more complex methods, existing assessment procedures and diagnostic manuals could be shortened (Margraf, 1994; Whooley et al., 1997; Zimmerman et al., 2010) or transformed into FFTs.

Table 4
Intercorrelations among the BDI items used to infer the respondents' depression status (according to the full BDI).

Item #	12	10	3	2	7
12	–	0.17	0.20	0.23	0.16
10		–	0.16	0.22	0.22
3			–	0.32	0.45
2				–	0.29

More than 50 years ago, Meehl (1954) demonstrated the benefits of using actuarial methods to improve decision making. Nevertheless, the use of these models in practical settings is still met with considerable skepticism (Dawes, Faust, & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000). One reason for the skepticism is that regression models have often been proposed as actuarial methods; these models may be too complex and opaque to be useful in clinical practice (Kleinmuntz, 1990). It has also been speculated that professionals fear being denigrated by their patients for using nonhuman computerized decision aids (Arkes, Shaffer, & Medow, 2007; Shaffer, Probst, Merkle, Arkes, & Medow, 2013). FFTs might offer a solution to these problems by being graphical, simple, transparent, easy to apply (even unnoticeably), and accurate (Adams & Leveson, 2012; Elwyn et al., 2001; Katsikopoulos et al., 2008). Medical professionals are already somewhat familiar with the concept of decision trees, as (more complex) decision trees are used in areas such as asthma (Hong, Dong, Jiang, Zhu, & Jin, 2011).

FFTs are also attractive because they require few cognitive resources. Information can be processed and the decision can be updated one cue at a time; only simple mental operations (answering "yes" or "no" to a few questions) are required. Indeed, spontaneous use of FFTs has previously been found in experts in the areas of medical decision making and mental health (Dhami & Harries, 2001; Smith & Gilhooly, 2006).

On a final note, we propose that anyone categorized as having clinically depressed mood during screening should subsequently see a clinical psychologist or a psychiatrist to receive a full diagnosis. However, in screening situations in which time, resources, or psychological knowledge are scarce, simple models might offer viable detection tools.

7. Discussion

One reason for the FFT's performance being comparable to that of the compensatory logistic regression may be information redundancy in the cues. As Table 4 shows, although the five cues were not highly intercorrelated (the mean intercorrelation was $r = .24$), there was clearly common variance among them. Accordingly, adding further cues does not necessarily add new information, meaning that simple models are able to compete with more complex ones (see Gigerenzer & Goldstein, 1996). Moreover, relative to the logistic regression model, both the unit-weight model and the FFT showed a lower susceptibility to overfit (Martignon et al., 2008; Martignon & Schmitt, 1999; Pitt, Myung, & Zhang, 2002): While the logistic regression model's discriminability decreased from fitting to cross validation (see Fig. 3), the same did not apply to the FFT and the unit-weight model.

Although it never clearly outperformed the FFT, the unit-weight model showed the highest accuracy and discriminability. This result underlines the power of unit-weighting, as highlighted in the seminal investigations by Dawes and Corrigan (1974). Nevertheless, the unit-weight model considers substantially more information than the FFT and might therefore be less appropriate when rapid assessment is key.

With regard to the bias in categorizing a respondent as having depressed mood, note that the bias of the FFT depends on its exit structure, with a higher number of "yes" exits leading to a more lenient threshold. For the present analysis we used a principled approach, namely the Max procedure, to construct the tree, without aiming for a specific type of bias. As the appropriateness of a certain bias depends on the base rate of positive cases and the cost structure of the different types of errors, alternative trees with a specific bias might also be constructed. In Online Appendix C, we report results on the performance of such trees, as well as that of the two-question model by Whooley et al. (1997) mentioned above. In short, the results show that the reversed tree ($d' = 2.68$ [1.75, 3.69]) and the two-question tree ($d' = 2.69$ [1.76, 3.70]) performed similarly to the tree produced by the Max procedure ($d' = 2.55$ [2.01, 3.10]) in terms of discriminability in cross validation. Both had lenient biases (reversed tree: $c = -0.66$ [-1.16; -0.19]; two-question tree: $c = -0.65$ [-1.16; -0.19]).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jarmac.2013.06.001.

References

Adams, S. T., & Leveson, S. H. (2012). Clinical prediction rules. *British Medical Journal*, *344*, d8312. http://dx.doi.org/10.1136/bmj.d8312

Andrade, L., Caraveo-Anduaga, J. J., Berglund, P., Bijl, R. V., De Graaf, R., Vollebergh, W., et al. (2003). The epidemiology of major depressive episodes: Results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys. *International Journal of Methods in Psychiatric Research*, *12*(3–21) http://dx.doi.org/10.1002/mpr.138

Arkes, H. R., Shaffer, V. A., & Medow, M. A. (2007). Patients derogate physicians who use a computer-assisted diagnostic aid. *Medical Decision Making*, *27*, 189–202. http://dx.doi.org/10.1177/0272989X06297391

Armstrong, J. S., & Cuzan, A. (2006). Index methods for forecasting: An application to the American Presidential Elections. *Foresight*, *3*, 10–13.

Barrera, M., & Garrison-Jones, C. V. (1988). Properties of the Beck Depression Inventory as a screening instrument for adolescent depression. *Journal of Abnormal Child Psychology*, *16*, 263–273. http://dx.doi.org/10.1007/BF00913799

Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*, 77–100. http://dx.doi.org/10.1016/0272-7358(88)90050-5

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571. http://dx.doi.org/10.1001/archpsyc.1961.01710120031004

Bowers, J., Jorm, A. F., Henderson, S., & Harris, P. (1992). General practitioners' reported knowledge about depression and dementia in elderly patients. *Australian and New Zealand Journal of Psychiatry*, *26*, 168–174. http://dx.doi.org/10.3109/00048679209072024

Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression. The x-random case. *International Statistical Review*, *60*, 291–319. Retrieved from. http://www.jstor.org/stable/1403680

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *The American Psychologist*, *34*, 571–582. http://dx.doi.org/10.1037/0003-066X.34.7.571

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*, 95–106. http://dx.doi.org/10.1037/h0037613

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674. http://dx.doi.org/10.1126/science.2648573

Dhami, M. K., & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making*, *14*, 141–168. http://dx.doi.org/10.1002/bdm.371

Dhami, M. K., & Harries, C. (2001). Fast and frugal versus regression models of human judgment. *Thinking and Reasoning*, *7*, 5–27. http://dx.doi.org/10.1080/13546780042000019

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, *13*, 171–192. http://dx.doi.org/10.1016/0030-5073(75)90044-6

Elwyn, G., Edwards, A., Eccles, M., & Rovner, D. (2001). Decision analysis in patient care. *The Lancet*, *358*, 571–574. http://dx.doi.org/10.1016/S0140-6736(01)05709-9

Engel, C. C., Oxman, T., Yamamoto, C., Gould, D., Barry, S., Stewart, P., et al. (2008). RESPECT-Mil: Feasibility of a systems-level collaborative care approach to depression and post-traumatic stress disorder in military primary care. *Military Medicine*, *173*, 935–940.

Everson, S. A., Roberts, R. E., Goldberg, D. E., & Kaplan, G. A. (1998). Depressive symptoms and increased risk of stroke mortality over a 29-year period. *Archives of Internal Medicine*, *158*, 1133–1138. http://dx.doi.org/10.1001/archinte.158.10.1133

Fischer, J. E., Steiner, F., Zucol, F., Berger, C., Martignon, L., Bossart, W., et al. (2002). Use of a simple heuristics to target macrolide prescription in children with community-acquired pneumonia. *Archives of Pediatrics and Adolescent Medicine*, *156*, 1005–1008. http://dx.doi.org/10-1001/pubs

Garcia-Retamero, R., & Dhami, M. K. (2009). Take-the-best in expert-novice decision strategies for residential burglary. *Psychonomic Bulletin and Review*, *16*, 163–169.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*, 107–143. http://dx.doi.org/10.1111/j. 1756-8765.2008.01006.x

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669. http://dx.doi.org/10.1037/0033-295X.103.4.650

Gigerenzer, G., Hertwig, R., & Pachur, T. (2011). *Heuristics: The foundations of adaptive behavior*. New York, NY: Oxford University Press.

Gigerenzer, G., & Selten, R. (Eds.). (2001). *Bounded rationality: The adaptive toolbox*. Cambridge, MA: MIT Press.

Gigerenzer, G., Todd, P. M., & The ABC Research Group (Eds.). (1999). *Simple heuristics that make us smart*. New York, NY: Oxford University Press.

Glöckner, A., & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition*, *123*, 21–32. http://dx.doi.org/10.1016/j.cognition.2011.12.002

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford, UK: Robert E. Krieger.

Green, L., & Mehr, D. R. (1997). What alters physicians' decisions to admit to the coronary care unit? *Journal of Family Practice*, *45*, 219–226. Retrieved from:. http://web.missouri.edu/~segerti/capstone/CCUdecisions.pdf

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical generalization: A meta-analysis. *Psychological Assessment*, *12*, 19–30. http://dx.doi.org/10.1037/1040-3590.12.1.19

Hastie, T., Tishirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Hautzinger, M. (1991). Das Beck-Depressioninventar (BDI) in der Klinik [The German version of the Beck Depression Inventory (BDI) in clinical use]. *Der Nervenarzt*, *62*, 689–696.

Herbert, T. B., & Cohen, S. (1993). Depression and immunity: A meta-analytic review. *Psychological Bulletin*, *113*, 472–486. http://dx.doi.org/10.1037/0033-2909.113.3.472

Hong, W. D., Dong, L. M., Jiang, Z. C., Zhu, Q. H., & Jin, S. Q. (2011). Prediction of large esophageal varices in cirrhotic patients using classification and regression tree analysis. *Clinics (Sao Paolo)*, *66*, 119–124. Retrieved from. http://www.ncbi.nlm.nih.gov/pubmed/21437447

Inskip, H. M., Harris, C., & Barraclough, B. (1998). Lifetime risk of suicide for affective disorder, alcoholism, and schizophrenia. *The British Journal of Psychiatry*, *172*, 35–37. http://dx.doi.org/10.1192/bjp.172.1.35

Jaycox, L. H., Reivich, K. J., Gillham, J., & Seligman, M. E. P. (1994). Prevention of depressive symptoms in school children. *Behaviour Research and Therapy*, *32*, 801–816. http://dx.doi.org/10.1016/0005-7967(94)90160-0

Katsikopoulos, K. V. (2010). The less-is-more effect: Predictions and tests. *Judgment and Decision Making*, *5*, 244–257.

Katsikopoulos, K. V. (2011). Psychological heuristics for making inferences: Definition, performance, and the emerging theory and practice. *Decision Analysis*, *8*, 10–29. http://dx.doi.org/10.1287/deca.1100.0191

Katsikopoulos, K. V., Pachur, T., Machery, E., & Wallin, A. (2008). From Meehl to fast and frugal heuristics (and back): New insights into how to bridge the clinical–actuarial divide. *Theory Psychology*, *18*, 443–464. http://dx.doi.org/10.1177/0959354308091824

Kee, J., Jenkins, J., McIllwaine, S., Patterson, C., Harper, S., & Shields, M. (2003). Fast and frugal models of clinical judgment in novice and expert physicians. *Medical Decision Making*, *23*, 293–300. http://dx.doi.org/10.1177/0272989X03256004

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., et al. (2003). The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association*, *289*, 3095–3105. http://dx.doi.org/10.1001/jama.289.23.3095

Kleinmuntz, B. (1990). Why we still use our head instead of formulas: Toward an integrative approach. *Psychological Bulletin*, *107*, 296–310. http://dx.doi.org/10.1037/0033-2909.107.3.296

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on artificial intelligence, Vol. 2* (pp. 1137–1143). Retrieved from. http://dl.acm.org/citation.cfm?id=1643047

Krupinski, J., & Tiller, J. (2001). The identification and treatment of depression by general practitioners. *Australian and New Zealand Journal of Psychiatry*, *35*, 827–832. http://dx.doi.org/10.1046/j. 1440-1614.2001.00960.x

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 658–676. http://dx.doi.org/10.1002/wcs.72

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Oxford, UK: Academic Press.

Lee, M. D. (2008). BayesSDT: Software for Bayesian inference with signal detection theory. *Behavior Research Methods*, *40*, 450–456. http://dx.doi.org/10.3758/BRM.40.2.450

Lewinsohn, P. M., Hops, H., Roberts, R. E., Seeley, J. R., & Andrews, J. A. (1993). Adolescent psychopathology. I. Prevalence and incidence of depression and other DSM-III-R disorders in high school students. *Journal of Abnormal Psychology*, *102*, 133–144. http://dx.doi.org/10.1037/0021-843X.102.1.133

Luan, S., Schooler, L. J., & Gigerenzer, G. (2011). A signal-detection analysis of fast-and-frugal trees. *Psychological Review*, *118*, 316–338. http://dx.doi.org/10.1037/a0022684

Macmillan, N. A., & Creelman, D. C. (2004). *Detection theory: A user's guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

Marewski, J. N., & Gigerenzer, G. (2012). Heuristic decision making in medicine. *Dialogues in Clinical Neuroscience*, *14*, 77–89.

Margraf, J. (1994). *(German version of the Anxiety Disorders Interview Schedule) Mini DIPS: Diagnostisches Kurz-Interview bei psychischen Störungen*. Berlin, Germany: Springer.

Margraf, J., Schneider, S., Soeder, U., Neumer, S., & Becker, E. S. (1996). *(Diagnostic Interview for Psychiatric Disorders (research version)) F-DIPS: Diagnostisches Interview bei Psychischen Störungen (Forschungsversion)*. Berlin, Germany: Springer.

Martignon, L., Katsikopoulus, K. V., & Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology*, *52*, 352–361. http://dx.doi.org/10.1016/j.jmp.2008.04.003

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2012). Naive, fast and frugal trees for classification. In P. Todd, G. Gigerenzer, & the ABC Group (Eds.), *Ecological rationality: Intelligence in the world*. New York, NY: Oxford University Press.

Martignon, L., & Schmitt, M. (1999). Simplicity and robustness of fast and frugal heuristics. *Minds and Machines*, *9*, 565–593. http://dx.doi.org/10.1023/A:1008313020307

MCMCpack [Computer software]. Retrieved from: http://mcmcpack.wustl.edu

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press. http://dx.doi.org/10.1037/11281-000

Miller, G. (2001). Predicting the psychological risks of war. *Science*, *333*, 520–521. http://dx.doi.org/10.1126/science.333.6042.520

Pachur, T. (2010). Recognition-based inference: When is less more in the real world? *Psychonomic Bulletin and Review*, *17*, 589–598.

Pachur, T., & Marinello, G. (2013). Expert intuitions: How to model the decision strategies of airport customs officers? *Acta Psychologica*, *144*, 97–103. http://dx.doi.org/10.1016/j.actpsy.2013.05.003

Pachur, T., Hertwig, R., & Rieskamp, J. (2013). Intuitive judgments of social statistics: How exhaustive does sampling need to be? *Journal of Experimental Social Psychology*, http://dx.doi.org/10.1016/j.jesp.2013.07.004 (in press)

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*, 1373–1379. http://dx.doi.org/10.1016/S0895-4356(96)00236-3

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491. http://dx.doi.org/10.1037/0033-295X.109.3.472

Reynolds, W. M. (1986). A model for the screening and identification of depressed children and adolescents in school settings. *Professional School Psychology*, *1*, 117–129. http://dx.doi.org/10.1037/h0090504

Roberts, R. E., Lewinsohn, P. M., & Seeley, J. R. (1991). Screening for adolescent depression: A comparison of depression scales. *Journal of the American Academy of Child & Adolescent Psychiatry*, *30*, 58–66. http://dx.doi.org/10.1097/00004583-199101000-00009

Schooler, L. J., & Shiffrin, R. M. (2005). Efficiently measuring recognition performance with sparse data. *Behavior Research Methods*, *37*, 3–10. http://dx.doi.org/10.3758/BF03206393

Shaffer, V. A., Probst, C. A., Merkle, E. C., Arkes, H. R., & Medow, M. A. (2013). Why do patients derogate physicians who use a computer-based diagnostic support system? *Medical Decision Making*, *33*(108–118) http://dx.doi.org/10.1177/0272989X12453501

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*, 233–250. http://dx.doi.org/10.1002/bdm.413

Sharp, L. K., & Lipsky, M. S. (2002). Screening for depression across the lifespan: A review of measures for use in primary care settings. *American Family Physician*, *66*, 1001–1009.

Smith, L., & Gilhooly, K. (2006). Regression versus fast and frugal models of decision-making: The case of prescribing for depression. *Applied Cognitive Psychology*, *20*, 265–274. http://dx.doi.org/10.1002/acp.1189

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50. http://dx.doi.org/10.1037/0096-3445.117.1.34

Snook, B., Dhami, M. K., & Kavanagh, J. M. (2011). Simply criminal: Predicting burglars' occupancy decisions with a simple heuristic. *Law and Human Behavior*, *35*, 316–326. http://dx.doi.org/10.1007/s10979-010-9238-0

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments & Computers*, *3*(137–149) http://dx.doi.org/10.3758/BF03207704

Steer, R. A., Cavalieri, T. A., Leonard, D. M., & Beck, A. T. (1999). Use of the Beck Depression Inventory for Primary Care to screen for major depression disorders. *General Hospital Psychiatry*, *21*, 106–111. http://dx.doi.org/10.1016/S0163-8343(98)00070-X

Trumpf, J., Vriends, N., Meyer, A. H., Becker, E. S., Neumer, S. P., & Margraf, J. (2010). The Dresden Predictor Study of anxiety and depression: Objectives, design, and methods. *Social Psychiatry and Psychiatric Epidemiology*, *9*, 853–864. http://dx.doi.org/10.1007/s00127-009-0133-2

Vulkan, N. (2002). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*, 101–118. http://dx.doi.org/10.1111/1467-6419.00106

Whooley, M. A., Avins, A. L., Miranda, J., & Browner, W. S. (1997). Case-finding instruments for depression: Two questions are as good as many. *Journal of General Internal Medicine*, *12*, 439–445. http://dx.doi.org/10.1046/j. 1525-1497.1997.00076.x

World Health Organization. (2001). *The world health report 2001: Mental health: New understanding, new hope* (Annual Report, 2001 edition). Retrieved from: http://www.who.int/whr/2001/chapter2/en/index4.html

Zimmerman, M., Galione, J. N., Chelminski, I., McGlinchey, J. B., Young, D., Dalrymple, K., et al. (2010). A simpler definition of major depressive disorder. *Psychological Medicine*, *40*(451–457) http://dx.doi.org/10.1017/S00332917099