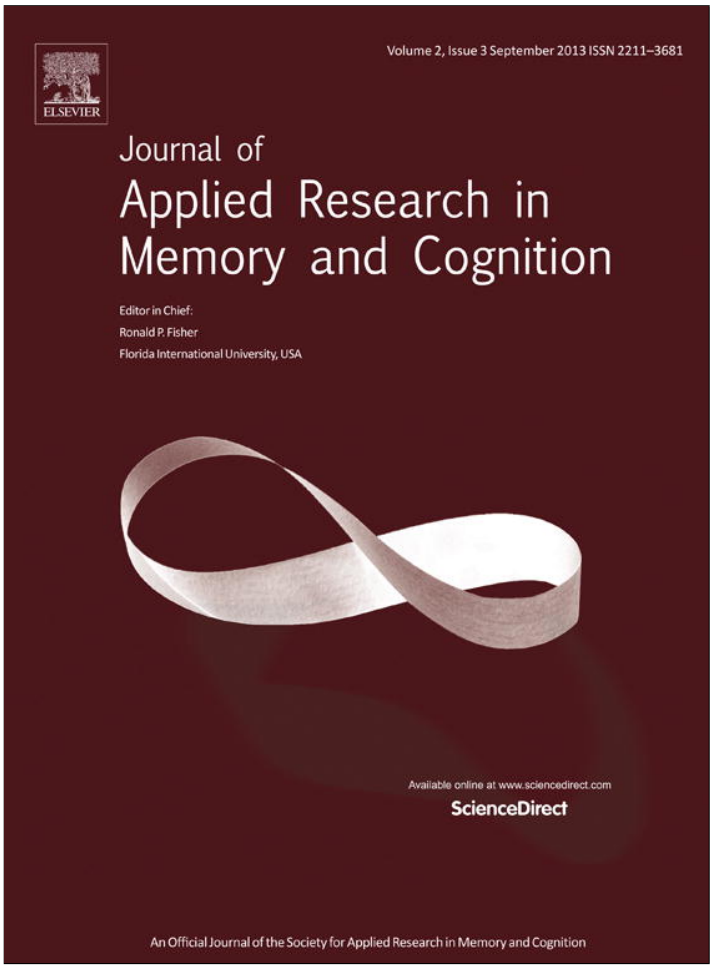

Journal of Applied Research in Memory and Cognition 2 (2013) 137–148

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

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

New Concepts

A quantification of robustness

Matthew M. Walsh[a,*], Evan H. Einstein[b], Kevin A. Gluck[a]

[a] Air Force Research Laboratory, Wright-Patterson Air Force Base, OH, United States
[b] Program in Cognitive Science, Vassar College, NY, United States

ARTICLE INFO

Article history:
Received 20 September 2012
Received in revised form 19 July 2013
Accepted 23 July 2013
Available online 9 August 2013

Keywords:
Robustness
Decision-making
Cognitive systems
Quantification

ABSTRACT

Robustness is an important construct in domains as diverse as evolutionary biology, structural engineering, and decision-making. Unfortunately, in many domains, most relevantly cognitive science, considerations of robustness end with vague semantic references. Little attention is paid to formal analysis. The aim of this paper is to initiate a discussion in the scientific community regarding methods for quantifying and analyzing robustness. To this end, we propose a means for assessing robustness that may supplant the current ambiguous use of the term. We demonstrate our quantitative approach using examples of heuristic-based decision processes, selected due to their explicit association with robustness in the psychological literature. These examples serve to illustrate basic properties of our general methodology for quantifying robustness.

On January 15th, 2009, US Airways Flight 1549 departed LaGuardia airport bound for Charlotte, North Carolina. During its initial climb, the aircraft struck a flock of Canada Geese. The resulting damage to the engines caused an immediate and complete loss of thrust.

Bird strikes have long been a concern in aviation. However, the majority of bird strikes go unnoticed; 85% cause no discernible damage to aircraft (Borrell, 2009). Catastrophe is usually avoided because jet engines are designed to withstand bird strikes, due to the known risks associated with engines and birds sharing the same airspace. This type of dynamic environmental risk is the motivation behind Design For Variation, a strategic initiative launched by Pratt & Whitney (Reinman et al., 2012). Design for Variation embraces the idea of designing a system that can operate in a variable environment. Design for Variation provides aircraft engines with a degree of robustness that increases passenger safety. For example, aircraft engines are constructed to operate across a range of altitudes and weather conditions. Systems cannot be robust against *all* environmental variations, however: it would be impractical to build aircraft that could accommodate *all* conceivable events. The engines in this example, though able to withstand isolated bird strikes, were not able to withstand the stress of an entire flock of geese.

Fortunately, because of his extensive training and experience, captain Chesley Sullenberger was able to land the transformed 'glider' on the frigid water of the Hudson River, saving the lives of everyone onboard. The event has since been labeled the "Miracle on the Hudson" (Fraher, 2011). Captain Sullenberger's piloting skills proved more robust than the aircraft's engines.

1. From physical robustness to cognitive robustness

The juxtaposition between the pilot's robustness and the aircraft's fragility highlights the relevance of cognition to system performance. Looking beyond this one anecdote, evidence for the relevance of robustness to cognition comes from the recently released *Oxford handbook of cognitive engineering* (Lee & Kirlik, 2013), where 21 of 45 chapters mention robustness. Of these, however, only two go on to define the term, and *none* attempt to quantify robustness. This is representative of psychological science as a whole: a review of psychology and behavioral science journals using Web of Science returned 441 articles containing the words 'robust' or 'robustness' within their title (January, 2013). Of these, approximately 46% pertained to statistical methods, 26% to empirical phenomena, 18% to psychosocial assessment and treatment, and 10% to cognitive capacities. The overwhelming majority of articles never defined 'robust'. Definitions from the few articles that did so underscore the varied ways in which the word is used (Table 1). Further, no article provided a quantitative measure of robustness. Apparently, the word robust, though regularly used, is rarely defined and never quantified.

2. From semantic ambiguity to mathematical precision

Gluck et al. (2012) defined robustness as, ". . . the extent to which a system is able to maintain its function when some aspect of the

* Corresponding author at: 711 HPW/RHAC, Cognitive Models and Agents Branch, 2620 Q Street, Building 852, Wright-Patterson AFB, OH 45433, United States. Tel.: +1 937 938 4057; fax: +1 937 904 8810.
E-mail address: mmw188@gmail.com (M.M. Walsh).

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.2013.07.002

Table 1
Sample uses of the terms 'robust' and 'robustness' in the psychological sciences.

Subfield	Example	Interpretation
Statistical	"Robust methods include finding population parameters, estimators, and hypothesis-testing methods that are not drastically affected by small changes in a distribution" (Wilcox, 1998, p. 5).	Stable
Statistical	"The term robust statistics refers to procedures that are able to maintain the Type I error rate of a test at its nominal level and also maintain the power of the test, even when data are nonnormal and heteroscedastic" (Erceg-Hurn & Mirosevich, 2008, p. 593).	General/Strong
Empirical	"In fact, the findings . . . are highly robust and consistent. All attempts at replication have found the same pattern of results" (Wynn, 2000, p. 1535).	Replicable
Empirical	"To demonstrate that the learning effect was robust for older controls, we conducted an analysis on their data only. The effect of behavior remained highly significant, $F(1, 10) = 14.88$, $P < 0.003$, $\eta^2 = 0.60$" (Todorov & Olson, 2008, p. 199).	Strong
Psychosocial	"Another crucial test is whether the prototypes are robust; that is, replicable across different samples" (Eaton, Krueger, South, Simms, & Clark, 2011, p. 1152).	Replicable
Psychosocial	"Investigations of generalizability or 'robustness' of intervention effects are also important. Answers to questions such as for what groups does this intervention, delivered by what types of staff, produce effects on what outcomes and under what conditions are essential" (Glasgow et al., 2008, p. 787).	General
Cognitive	"Robustness is associated with the ability to protect skilled performance from various disturbances, including unexpected events, interruptions, or changing demands" (Taatgen et al., 2008, p. 548).	Stable/General
Cognitive	"The robust satisficer answers two questions . . . of the options that will produce a good enough outcome, which one will do so under the widest range of possible future states of the world" (Schwartz, Ben-Haim, & Dacso, 2010).	General

system is subject to perturbation" (p. 193). This definition resonates with the conception of robustness in biology: a biological system is robust if it continues to function despite perturbations (Kitano, 2004; Larhlimi, Blachon, Selbig, & Nikoloski, 2011; Wagner, 2005). Robustness is quantified by measuring the size or volume of the parameter space that yields viable performance (i.e., *global analysis*), or by evaluating the effects of specific parameter changes on model behavior (i.e., *sensitivity analysis*). Gluck et al.'s definition also resonates with the conception of robustness in structural engineering: a structure is robust if the loss of a load-bearing component does not cause progressive disproportionate collapse (Canisius, 2011; Starossek & Haberland, 2012). Robustness is quantified by comparing the stiffness of damaged and intact structures, or by estimating the progressive collapse resulting from an initial cause.

Control theory provides additional mathematical tools for finding optimal control solutions, which are described as 'robust' (Jagacinski & Flach, 2003). Our quantification of robustness, and the quantification of robustness in biology and engineering, does not directly concern optimization, however (e.g., Dueck, 1993). Rather, we are interested in the cognitive system's ability to function in varying conditions.

Though limited measures of robustness exist in biology and engineering, cognitive scientists have not attempted to quantify robustness. Doing so in basic and applied psychological research is worthwhile for precisely the same reasons that such techniques have been proposed in other fields. Formal quantitative measures, much like formal computational models, enhance the transparency of predictions and testability of theories (Marewski & Olsson, 2009; Tomlinson, Marewski, & Dougherty, 2011). Formal quantitative measures also provide an objective way to express and compare the outcomes of different interventions in applied settings. In short, quantitative measures advance science and enable application in a way that qualitative semantic references cannot.

The purpose of this paper is to introduce a quantitative method for assessing robustness. We describe this method, and apply it to previously published results from decision heuristics research. We chose this particular subset of the scientific literature because of its explicit association with robustness in a variety of recent publications (Gigerenzer, 2008; Gigerenzer & Brighton, 2009; Gigerenzer & Gaissmaier, 2011; Marewski & Gigerenzer, 2012; Marewski & Schooler, 2011). That said, our measure is general and can be applied to a wide range of topics within and beyond cognitive science.

3. Quantification method

Working from Gluck et al.'s (2012) qualitative description of robustness, we can construct a quantitative measure. *Functionality* (a dependent variable) is measured with respect to *perturbations* (independent variables). Perturbations may evoke changes in functionality. *Robustness* is the value that arises from the integration of functionality over the range of evaluated perturbations.

To quantify robustness, one must first specify the system's function. In simple scenarios, the system's function may be determined by a single constraint, for example, to respond correctly or to discover a source of nutrients. In more complex scenarios, the system's function may be determined by multiple constraints, for example, to discover a source of nutrients quickly and with minimal expenditure of metabolic resources. We define functionality as *the ability of a system to achieve its goal*.

To quantify robustness, one must also specify the perturbations the system is liable to encounter. We envision the environment as a multivariate space (Fig. 1). Each scenario the individual may face

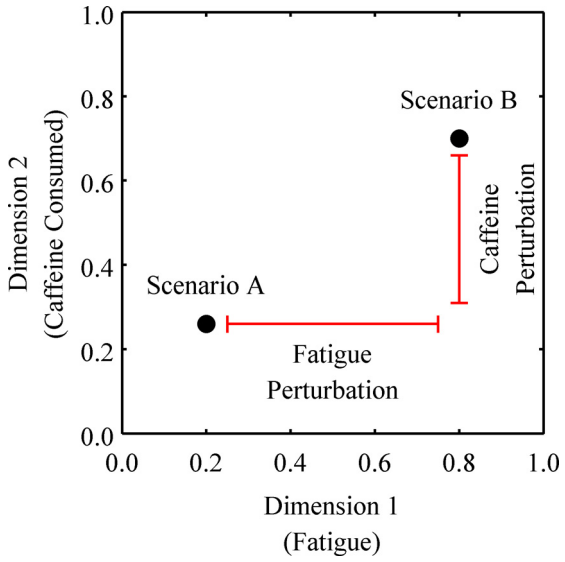

Fig. 1. Environment defined by two dimensions (fatigue and caffeine consumed). Altering the values along the dimensions creates two scenarios within the environment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

constitutes a point in space defined by the values of the variables, or dimensions, that comprise the environment. Variables can be internal to the individual, for example, the number of hours slept during the previous night, or the amount of caffeine consumed. Variables can also be external to the individual; for example, time of day, ambient lighting, and the frequency of critical events in jobs that require sustained attention (e.g., air traffic control). Perturbations involve manipulating the values of one or more variables in order to move the individual from one scenario to another.

After defining a system's function and identifying the perturbations the system is liable to encounter, one can quantify the robustness of the system against those perturbations. Broadly, this entails three steps: calculate functionality, assess robustness, and measure stability.

3.1. Method

Step 1: Calculate functionality

Starting from a single scenario, we ask, did the system achieve its function? Functionality is defined as,

$$Functionality = \left(S - \frac{F}{T}\right) \quad (1)$$

Success (S) is how frequently, or the degree to which, the system achieved its function in the scenario, and failure (F) is how frequently, or the degree to which, the system did not achieve its function. Success and failure can be defined by a single objective (e.g., *Did the system respond correctly?*), or by multiple objectives (e.g., *Did the system respond correctly and within the allocated time?*).

The setting of tolerance (T) depends on the consequences of failure. If errors result in the loss of lives, tolerance must be low. If errors are inconsequential, tolerance may be somewhat higher. Tolerance relates to the notion of risk-based criteria (e.g., maintain below 3.4 failures per million opportunities, or 3.4×10^{-6}), as used in Six Sigma design. Our measure of functionality takes special meaning when tolerance expresses the ratio of allowable risk to one minus allowable risk, $risk/(1 - risk)$. When tolerance is set in this way, negative functionality indicates that the system exceeds allowable risk, and positive functionality indicates that the system remains safely within allowable risk.

There are several approaches for establishing risk-based criteria (Reinman et al., 2012), and in turn, tolerances. Risk-based criteria can be set to levels associated with previous acceptable performance. For example, the standard for landmine detection specified in the Operational Requirements Document for the AN/PSS-12 handheld mine detector is 92% (US Army, 1990). In terms of a model of landmine detection (Staszewski, 2006), this translates to a tolerance of 0.087. Risk-based criteria can also be based on federal standards. For example, the National Assessment Governing Board (NAGB) issued achievement levels for math skills targeted by the National Assessment of Education Progress (NAEP) in Grade 12 Mathematics Assessment (ACT, 2005). Proficiency in performing basic mathematical operations was gauged by achieving 90% accuracy. This translates to a tolerance of 0.111. The standard for evaluating algebraic expressions, equations, and inequalities was somewhat lower (73%). This translates to a tolerance of 0.370. Lastly, risk-based criteria can be determined using tools such as a risk matrix or economic analysis of total costs. Such tools have been used in civil engineering and medical decision-making to establish target levels of reliability (Huaco, Bowders, & Loehr, 2012; Weinstein & Fineberg, 1980).

Within the same task, different types of errors may be associated with different outcomes. For example, the consequences of failing to detect certain traffic signals, such as vehicle brake lights, are catastrophic (i.e., collision). The consequences of responding prematurely to traffic signals are somewhat less severe (i.e., congestion). Signal detection theory (SDT) accounts for the values of different outcomes: Swets, Dawes, and Monahan (2000) noted that choices should be aligned with the consequences of different decisions. For example, the cost of misdiagnosing cancer in a healthy individual may be less than the cost of failing to diagnosis cancer in a sick individual. These considerations suggest a necessary extension of our method to problems where responses cannot simply be counted as successes or failures. In such cases, different tolerances may be assigned to different response categories,

$$Functionality = \left([S_1 + S_2 + \cdots + S_m] - \left[\frac{F_1}{T_1} + \frac{F_2}{T_2} + \cdots + \frac{F_n}{T_n}\right]\right) \quad (2)$$

where $S_{1\ldots m}$ refer to the percentage of successful responses in each of m categories, and $F_{1\ldots n}$ refer to the percentage of unsuccessful, or failed responses in each of n categories. $T_{1\ldots n}$ describe the unique tolerances for each type of failure.

The range of functionality scores is context specific: its upper bound equals one, and its lower bound equals the negative inverse of the strictest tolerance, $-1/\min(T_{1\ldots n})$.[1] To facilitate interpretation across contexts, we advocate normalizing functionality scores,

$$Functionality_{norm} = \frac{\min(T_{1\ldots n}) \cdot Functionality + 1}{\min(T_{1\ldots n}) + 1} \quad (3)$$

The resulting scores fall between one (maximally functional) and zero (minimally functional).

Step 2: Assess robustness

For all scenarios created by combinations of environment dimensions, we ask, how well did the system achieve its function? Robustness is defined as,

$$Robustness = \int Functionality(x) \cdot Probability(x)\, dx \quad (4)$$

$Probability(x)$ describes the likelihood that scenario x will occur. Without prior information, all scenarios are treated as equally likely. Given that the probability weighting function sums to one, and that the maximum possible value of functionality equals one, the maximum possible value of robustness also equals one. As functionality decreases to zero, so too does robustness.

Step 3: Measure stability

Lastly, we ask, how much did functionality vary across scenarios? Stability is defined as,

$$Stability = 1 - 2 \cdot std(Functionality(x)) \quad (5)$$

Stability depends on the variability among functionality scores. Because normalized functionality scores fall between zero and one, the range of the standard deviation among functionality scores is 0.0 to 0.5. We multiply standard deviation by two to place our measure of stability between zero and one. Further, we assign negative value to the standard deviation because stability is *inversely* related to variability among functionality scores. The resulting stability scores fall between one (maximally stable) and zero (minimally stable).

Fig. 2 illustrates four scenarios created by crossing conditions of high and low robustness and stability. The ideal system is robust *and* stable. The explicit separation of robustness and stability deviates from the typical control theoretic interpretation of robustness (Zhou & Doyle, 1997), which is actually what we call stability.

[1] In terms of Eq. (1), this corresponds to the case where the proportion of successful responses equals zero ($S = 0$) and the proportion of failed responses equals one ($F = 1$).

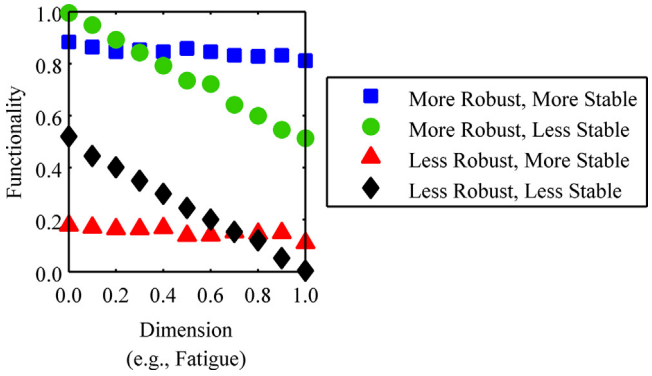

Fig. 2. Functionality of more robust and less robust systems, and of more stable and less stable systems. A more robust system has high functionality over a variation in dimension. A more stable system has consistent functionality over a variation in dimension. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

3.2. Implementation

Robustness can be assessed empirically. For example, the functionality of a cognitive process can be observed as the process is evoked across scenarios. This is called *empirical* functionality because it relates to the observed functionality of a system in different scenarios. Robustness can also be assessed through simulation. For example, a cognitive process can be instantiated computationally to explore its functionality across simulated scenarios. This is called *predicted* functionality because it relates to the expected functionality of a system in different scenarios. In both cases, robustness is quantified by: (1) calculating functionality in isolated scenarios; (2) assessing the maintenance of functionality (i.e., *robustness*) across scenarios; and (3) measuring stability across scenarios.

In the three examples that follow, we use simulations to measure predicted functionality at equally spaced points along a grid spanning the environment dimensions (i.e., a Latin Hypercube; McKay, Beckman, & Conover, 1979). Because our simulations are stochastic, we calculate a distribution of functionality scores at each point. Robustness and stability describe functionality over all points spanning the environment dimensions. To quantify robustness and stability, we generate samples. Each sample contains one functionality score from every point. The mean of functionality within a sample provides an estimate of robustness, and the variance of functionality within a sample provides an estimate of stability. Further, by calculating robustness and stability repeatedly using a large number of samples, we can generate confidence intervals for these measures.

When computational simulations are costly to run, or when performance is evaluated empirically, it may be feasible to assess functionality only in a small number of scenarios. In cases where limited information is available, one could use other simulation or interpolation techniques to estimate robustness and stability (e.g., Gaussian process emulators; Oakley & O'Hagan, 2004). These techniques are not antithetical to our approach. Rather, they efficiently approximate the integral of robustness (Eqs. (1) and (2)) and stability (Eq. (3)) based on fewer scenarios.

4. Robustness of heuristic-based decisions

While describing the emergency landing of US Airways Flight 1549 on the Hudson River, we noted that the pilots' skills were robust against the emergency they faced following engine failure. The pilots quickly realized they could not reach any nearby runways. One might imagine that this realization required complex calculations involving factors such as altitude, airspeed, heading, distance, and wind speed. In fact, co-pilot Jeffrey Skiles stated in an interview, "It's not so much a mathematical calculation as visual, in that when you are flying in an airplane, a point that you can't reach will actually rise in your windshield" (Gigerenzer, Hertwig, & Pachur, 2011, p. 633). Skiles's response contains a description of a visual tracking strategy called the *gaze heuristic*. The gaze heuristic is easy to enact and effective. These two properties, ease and effectiveness, are characteristic of heuristics more generally (Gigerenzer, 2008).

According to the adaptive toolbox theory (Gigerenzer & Gaissmaier, 2011), the mind contains a collection of such heuristics. These heuristics are fast in that they employ simple information processing operations, and they are frugal in that they require little information to enact. For example, the gaze heuristic requires determining whether a target rises, falls, or remains level within the visual field. By capitalizing on statistical regularities in the environment, heuristics perform as well as, or better than, methods that employ more complex calculations and require greater amounts of information. For these reasons, heuristics are thought to be especially robust to environmental variation (Brighton & Gigerenzer, 2011; Gigerenzer, 2008; Gigerenzer & Brighton, 2009; Gigerenzer & Gaissmaier, 2011; Rieskamp & Otto, 2006).

Although researchers have identified and established formal models of decision heuristics, this literature lacks a formal metric and methodology to evaluate the robustness of different decision strategies. To address this deficiency and demonstrate our quantitative method, we now apply it to three examples from the decision heuristics literature. The first example measures the robustness of take-the-best, tally, and regression against variation in the size of the training set (Dawes, 1979; Gigerenzer & Goldstein, 1996). This example involves quantifying robustness over a single dimension. The second example evaluates the robustness of the recognition heuristic against variations in recognition validity and recognition rate (Goldstein & Gigerenzer, 2002). This example extends the methodology to a multi-dimensional case where the likelihoods of different scenarios vary. The third example assesses the robustness of four fast and frugal trees against variations in base rates and payoff structures (Luan, Schooler, & Gigerenzer, 2011). This example centers on manipulations of tolerance. Although each of these examples comes from the literature on heuristic decision-making, our method is not limited to these types of tasks. Rather, our method is extremely general, a point that we return to in the discussion.

4.1. Probabilistic inference

Probabilistic inference involves choosing between alternatives based on several attributes, each of which is differentially associated with an alternative's value. For example, an investor might consider several features of two equity funds before allocating their resources to one. Such decisions are complicated by two factors. First, no single attribute or combination of attributes typically predicts the best alternative with certainty: outcomes are probabilistic rather than deterministic. Second, different attributes typically favor different alternatives: no single alternative is dominant.

The take-the-best (TTB) heuristic is a model of how individuals infer which of two alternatives has a higher value on a criterion (Gigerenzer & Goldstein, 1996). TTB searches attributes in order of their validity, stops upon identifying an attribute that discriminates between alternatives, and selects the alternative with the greater attribute value. The tally (TAL) heuristic is another model of how individuals make probabilistic inferences. TAL evaluates all attributes, counts the number of positive attributes for each alternative, and selects the alternative with the greater number of positive attributes (Dawes, 1979). TTB and TAL are *fast* in that they

use only simple mathematical operations (i.e., binary comparisons and counting). Additionally, TTB is *frugal* in that it can be applied given very little information about alternatives.

Multiple linear regression is the standard, statistical solution to the probabilistic inference problem. This technique involves estimating the set of weights that best predict the scalar value of a dependent variable based on one or more explanatory variables (Brunswik, 1955). In the context of probabilistic inference, the regression model (REGRESS) evaluates all attributes, computes a weighted sum of the attributes for each alternative, and chooses the alternative with the greatest resulting value. REGRESS is not fast. Application of the model requires computing a weighted sum of attributes, which entails a considerably more complex set of mathematical operations. REGRESS is also not frugal. Application of the model requires evaluating all attributes. These considerations notwithstanding, REGRESS constitutes a statistically motived solution to the probabilistic inference problem.

Despite their simplicity, TTB and TAL sometimes perform as well as, or better than, the more sophisticated REGRESS model. For example, Czerlinski, Gigerenzer, and Goldstein (1999) reported that the accuracy of TTB exceeds the accuracy of REGRESS when the number of observations used to train the models is small. This prompted us to ask, how robust are TTB, TAL, and REGRESS against variation in the size of the training set? A robust and stable strategy would yield the correct answer with high probability across all levels of training. A robust but unstable strategy would yield the correct answer with medium probability at low levels of training, and with high probability at high levels of training.

To answer this question, we simulated performance in a city population task (Gigerenzer & Goldstein, 1996). In the task, participants view a pair of cities and are asked which has the larger population. Gigerenzer and Goldstein (1996) identified nine attributes that correlated with the populations of 83 German cities. Attributes differed in their validity; that is, the proportion of times that the attribute predicted the larger city when the attribute's value was positive for one city in the pair and negative for the other. Some attributes had high validity (e.g., *Does the city have a major league soccer team?*), and others had low validity (e.g., *Is the city in the industrial belt?*).

During training, we varied the number of German cities from 8 to 75 (~10% to 90% of the 83 German cities, yielding 68 training set sizes). During test, we generated all pairings of the German cities not included in training, and we recorded the ability of TTB, TAL, and REGRESS to identify which city in each pair had the larger population. In this example, the function of the three models is to respond accurately, and the dimension along which perturbations are applied is the size of the training set. To simplify matters, we set tolerance to one, giving equal weight to correct and incorrect responses.

For all values of training set size, we estimated the functionality of the three models based on 500 simulations.[2] In each simulation, items in the training set were randomly selected from the complete set of available items. The remaining items were used in the test set. To estimate robustness and stability, we created 500 samples that each contained a functionality score from every training set size. Thus, each sample contained 68 functionality scores. The average functionality within a sample provides an estimate of robustness, and the variability among functionality scores within a sample provides an estimate of stability. Because these estimates differ among the 500 samples, their distributions provide information about uncertainty in our measures of robustness and stability, allowing the computation of associated statistics such as effect sizes and confidence intervals.

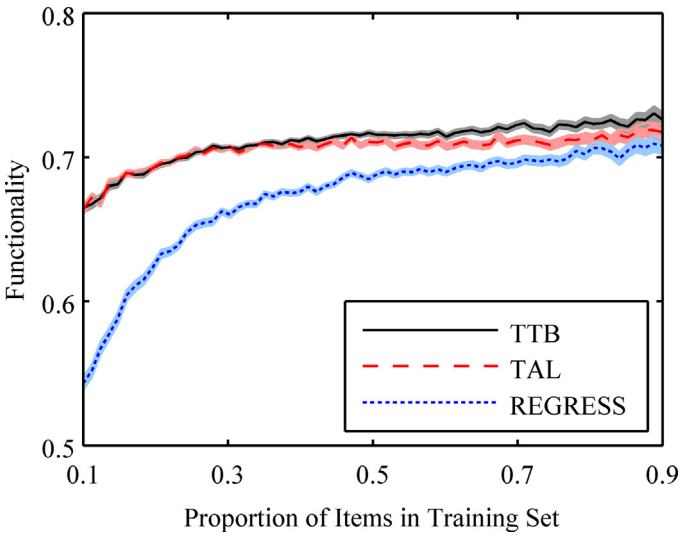

Fig. 3. Functionality of take-the-best (TTB), tally (TAL), and regression (REGRESS) models of probabilistic inference (±1 mean standard error) by the proportion of items included in the training set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Functionality of REGRESS was initially low, but increased with training set size (Fig. 3). Functionality of TTB and TAL, in contrast, began and remained at moderate-to-high levels. Consistent with these observations, robustness was greatest for TTB, followed by TAL, and then by REGRESS (Table 2). Stability was also greatest for TTB, followed by TAL, and then by REGRESS. Although the absolute differences among models appear modest, the effect sizes are substantial. The effect size of the difference between the robustness of TTB and REGRESS is large ($d = 5.65$), as is the effect size of the difference between their stability ($d = 2.72$). Likewise, the effect size of the difference between the robustness of TAL and REGRESS is large ($d = 4.52$), as is the effect size of the difference between their stability ($d = 2.35$). Differences between TTB and TAL were smaller, although TTB exhibited a slight advantage as the size of the training set increased. This is because TTB decides based on the most valid cue that distinguishes between alternatives, whereas TAL assigns equal weight to all cues, including those that are less valid.

TTB and TAL were more robust than their more complex counterpart REGRESS. There are two reasons for this. First, REGRESS must learn the direction and weight of cues whereas TAL must simply learn the direction of cues. Although TTB must also learn the direction and rank of cues, Katsikopoulos, Schooler, and Hertwig (2010) found that TTB and TAL performed well as long as the directions of the three most valid cues were known. These directions can be inferred from very small samples. Second, flexibility in the parameterization of REGRESS allows the model to account for a greater percentage of variance in the training data. Some of this variance relates to the underlying relationship between attributes and the values of alternatives, but some of this variance arises from noise in the sample (Gigerenzer & Brighton, 2009). Consequently, when set size is small, REGRESS is more likely to overfit the training data and to overgeneralize to test items.

Table 2
Robustness and stability of probabilistic inference models.

Model	Robustness	Stability
TTB	.71 ± 0.01 SD	.89 ± 0.01 SD
TALLY	.71 ± 0.01 SD	.88 ± 0.01 SD
REGRESS	.67 ± 0.01 SD	.84 ± 0.02 SD

[2] In these examples and in all later examples, additional simulations did not appreciably reduce error. The standard errors of all mean functionality scores from all simulations were below 0.005.

TTB and TAL were also more stable than their more complex counterpart REGRESS. Indeed, neither TTB nor TAL recorded a functionality score below 0.65. This outcome was not caused by some odd feature of the German cities problem: we obtained similar results with several other data sets (Czerlinski, Gigerenzer, & Goldstein, 1999). In contrast, REGRESS only performed well in circumscribed cases. The key point is that although the choice of strategy matters little when the training set is large, TTB and TAL are attractive because of their weak reliance on the number of training items. Further, to the extent that participants' responses are robust against variations in the size of the training set (Katsikopoulos, Schooler, & Hertwig, 2010), this suggests that REGRESS is an inadequate model of human behavior.

4.2. Recognition heuristic

Heuristics research has inspired the view that the mind contains a toolbox of decision strategies (Gigerenzer & Gaissmaier, 2011). This is important because no single heuristic can be used for all problems. Rather, each heuristic has a cognitive niche, or limited scope of circumstances, in which it can be applied (Marewski & Schooler, 2011). For example, take-the-best and tally can be used only when the individual has background knowledge about choice alternatives. This is unlikely to be the case when alternatives are encountered infrequently. The recognition memory literature holds that people experience a conscious sense of familiarity when they are exposed to infrequently encountered items, however, even if they cannot retrieve specific details about those items (Yonelinas, 2002).

The recognition heuristic exploits this core capacity. According to the recognition heuristic, if the task is to select between two alternatives, and one alternative is recognized but the other is not, select the recognized alternative (Goldstein & Gigerenzer, 1999).[3] The recognition heuristic can be applied to decisions about city populations (Goldstein & Gigerenzer, 1999), stock-market investments (Newell & Shanks, 2004), and even peanut butter preferences (Hoyer & Brown, 1990). Studies show that the recognized option is consistently favored in each of these cases.

Like take-the-best and tally, the recognition heuristic is not always applicable. The heuristic can be used only when one item in the pair is recognized and the other is not. Additionally, the recognition heuristic is not always correct. Its success depends largely on recognition validity: the heuristic is more accurate when recognition correlates with the judgment criterion (e.g., *Which city has a larger population, Zürich or Berne?*), and the heuristic is less accurate when recognition does not correlate with the judgment criterion (e.g., *Which city is located further from the Swiss city of Interlaken, Zürich or Berne?*). Based on these considerations we asked, how robust is the recognition heuristic against variations in item recognition rate and recognition validity?

The recognition heuristic has been studied extensively (for reviews, see Gigerenzer & Goldstein, 2011; Pachur, Todd, Gigerenzer, Schooler, & Goldstein, 2011). Conveniently, this allowed us to base our simulations on 19 papers that collectively included 50 studies of the recognition heuristic.[4] From these, 37 studies reported participants' recognition rates (i.e., the percentage of items that participants recognized). A partially overlapping set of 37 studies reported participants' recognition validity (i.e., the probability that when one item in the pair was recognized and the other was not, the recognized item had greater value for the judgment criterion). Fig. 4 shows the probability densities of recognition rates and validities from the corresponding studies. Each was reasonably approximated by a beta distribution (Recognition rate: $r^2 = 0.76$, mean squared error = 0.001; Recognition validity: $r^2 = 0.73$, mean squared error = 0.001).

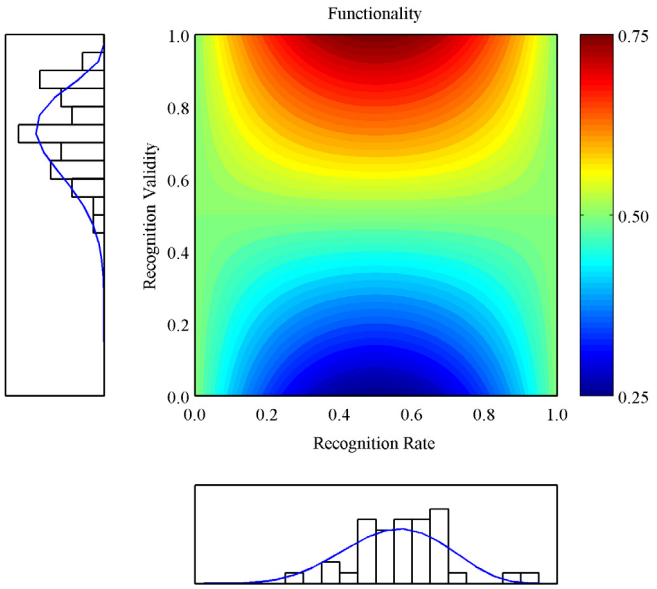

Fig. 4. Functionality of recognition heuristic by recognition rate and recognition validity. Histograms show probability densities of recognition rate and validity from empirical studies. Bin widths equal 0.05. Plotted curves show probability density estimates for recognition rate and validity based on beta distributions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

We varied recognition rate from 0.0 to 1.0 in increments of 0.01 (yielding 101 unique recognition rates). Additionally, we varied recognition validity from 0.0 to 1.0 in increments of 0.01 (yielding 101 unique recognition validities). To simplify matters, we set tolerance to one, giving equal weight to correct and incorrect responses. For each combination of recognition rate and validity, we estimated functionality based on 500 simulations. For each simulation, we calculated the percentage of correct responses in 120 trials, the median number of trials used in studies of the recognition heuristic. In each trial, two items appeared. Items were recognized with probability according to the recognition rate. When one item in the pair was recognized and the other was not, the model selected the recognized item. When both items in the pair were recognized or when neither item was recognized, the model chose randomly.

To estimate robustness and stability, we created 500 samples that each contained functionality scores for every combination of recognition rate and validity. Thus, each sample contained 10,201 functionality scores (101 recognition rates × 101 recognition validities). Average functionality within a sample, weighted by the probability density functions for the beta distributions, provides an estimate of robustness. Variability among functionality scores, weighted by the probability density functions for the beta distributions, provides an estimate of stability. Because these estimates differ among the 500 samples, their spreads provide information about uncertainty in our measures of robustness and stability.

Functionality of the recognition heuristic varied non-monotonically with recognition rate (Fig. 4). When recognition rate was low, the model recognized neither item and chose

[3] This does not imply that the memory system exists in binary states. Contemporary theories of declarative memory assume continuously varying levels of availability or activation, analogous to potentiation at the level of neural processing, along with a threshold above which memory retrievals occur and below which they do not. The use of the recognition heuristic simply requires that when attempting to retrieve two items from memory, one of the retrievals is successful (i.e., it is recognized) and the other is not.

[4] All but three of these papers (Ayton, Onkal, & McReynolds, 2011; Hilbig, Erdfelder, & Pohl, 2011, 2012) are included in the reviews by Gigerenzer and Goldstein (2011) or Pachur et al. (2011).

Table 3
Robustness and stability of the recognition heuristic given different distributions of recognition rates and recognition validities.

Recognition rate	Recognition validity	Robustness	Stability
Uniform (0, 1)	Uniform (0, 1)	.500 ± 0.001 SD	.796 ± 0.001 SD
Uniform (0, 1)	Beta (10.5, 3.95)	.575 ± 0.001 SD	.859 ± 0.002 SD
Beta (7.04, 5.18)	Uniform (0, 1)	.500 ± 0.001 SD	.750 ± 0.002 SD
Beta (7.04, 5.18)	Beta (10.5, 3.95)	.602 ± 0.001 SD	.861 ± 0.002 SD

Note: Values in parenthesis adjacent to uniform distribution correspond to upper and lower limits. Values in parenthesis adjacent to beta distribution correspond to empirically derived shape parameters α and β.

randomly. Likewise, when recognition rate was high, the model recognized both items and chose randomly. When recognition rate was moderate, however, the model sometimes recognized only one item in the presented pair, which it then selected. Functionality was greatest when recognition rate was moderate and recognition validity was high. Conversely, functionality was least when recognition rate was moderate and recognition validity was low. Robustness over the two dimensions was 0.602 ± 0.001 SD, and stability was 0.861 ± 0.002 SD.

To what extent do these results depend on the empirical distributions of recognition rate and validity? To answer this question, we repeated the simulations using different combinations of beta distributions and non-informative uniform priors. Table 3 shows robustness and stability of the recognition heuristic for the resulting combinations of recognition rate and validity. The recognition heuristic was more robust given the empirically derived beta distribution for recognition rate. The recognition heuristic can be applied only when one item in the pair is recognized. This is most likely to be the case when 50% of total items are recognized. Recognition rates in the majority of empirical studies tend to be near 0.50. The recognition heuristic was also more robust given the empirically derived beta distribution for recognition validity. Recognition validities in the majority of empirical studies exceed 0.50.[5] Functionality of the recognition heuristic, in turn, was high.

Although the recognition heuristic is sometimes incorrect, it is accurate in scenarios that individuals are more likely to encounter, and its performance declines gracefully over scenarios that they are less likely to encounter. Further, participants do not always use the recognition heuristic even when it is applicable. Adherence rate, the proportion of times that participants select the recognized item, is positively correlated with recognition validity (Gigerenzer & Goldstein, 2011; Pachur, 2010). Thus, although the recognition heuristic cannot always be used and is not always correct, it is robust over the range of scenarios where its functional contribution is maximal.

4.3. Fast and frugal trees

Fast and frugal trees (F&FTs) are a type of decision tree with the distinguishing trait that they have at least one exit point out of the decision process at each level (cue) in the tree (Martignon, Katsikopoulos, & Woike, 2008). The context of the Coronary Care Unit (CCU) has been a focus for F&FT development, following the proposal that using a simple heuristic approach (i.e., F&FTs) would reduce unnecessary utilization of the CCU (Green & Mehr, 1997). The CCU task requires a choice between admitting and not admitting a patient to the CCU by deciding if that patient is at high risk of myocardial infarction, which is damage to or death of muscle tissue in the heart. There are multiple defining signs of infarction seen through electrocardiographic (ECG), biochemical, and pathological evidence (Van de Werf et al., 2008). The decision in the CCU task is based primarily on visual evidence, consistent with models emphasizing perceptual categorization.

Luan et al. (2011) used signal detection theory (SDT) to analyze the behavior of three-cue F&FTs in the context of a CCU task. They proposed that modifying the exits (decisions that bypass the remaining cues and render a diagnosis) would produce different results by shifting the decision-making along the liberal-to-conservative continuum. This is a reference to how liberal or conservative the decision maker is with regard to assigning patients to the CCU. Exits could occur upon detecting a signal, 'S', and subsequently assigning the patient to a CCU bed. Exits could also occur upon failing to detect a signal, 'N', and subsequently assigning the patient to a regular bed, which involves a lower and less expensive level of monitoring. Four trees ($F\&FT_{SS}$, $F\&FT_{SN}$, $F\&FT_{NS}$, and $F\&FT_{NN}$) were tested in the CCU task (Fig. 5). Hit rates (diagnosing a sick individual) and false alarm rates (misdiagnosing a healthy individual) differed among trees (Fig. 6). F&FTs with more 'S' exits and with earlier 'S' exists ($F\&FT_{SS}$, $F\&FT_{SN}$, and $F\&FT_{S}$) had a liberal diagnostic bias, and F&FTs with more 'N' exists and with earlier 'N' exits ($F\&FT_{NS}$ and $F\&FT_{NN}$) had a conservative diagnostic bias.

The optimal decision bias depends on two factors (Swets, Dawes, & Monahan, 2000): the base rate of illness in the population, and the consequences of negative and positive decisions. When the base rate of illness is high, liberal decision strategies have greater utility. Additionally, when the cost of misses (failing to diagnosis a sick individual) exceeds the cost of false alarms (misdiagnosing a healthy individual), liberal decision strategies have greater utility. If these two factors are known, and if information from cues can be perfectly integrated, SDT can be used to calculate the decision bias that maximizes utility. In the absence of knowledge about these factors, or in environments where the underlying statistical probabilities or the cost–benefit structure are non-stationary, optimal SDT analysis is not possible. These considerations prompted us to ask, how robust are the four F&FTs against variations in the base rate of illness and the penalties for errors?

In this example, responses cannot simply be counted as successes or failures. The different types of failure, misses and false alarms, have different consequences. To incorporate these differences into our analysis, we adopted the expanded functionality equation,

$$Functionality = \left(S - \frac{M}{T_M} - \frac{FA}{T_{FA}}\right) \quad (6)$$

This allows for different contextual tolerance levels to be set for misses, T_M, and for false alarms, T_{FA}. Although both sources of error still reduce functionality, they do so to different extents.

We evaluated the robustness of the four F&FTs based on results from Green and Mehr (1997). Their dataset included information about 89 patients' symptoms and whether they experienced myocardial infarction. We varied the base rate of myocardial infarction from 0.10 to 0.90 in 0.01 increments (yielding 81 unique base rates). To do so, we resampled Green and Mehr's (1997) patient data with replacement to create simulated samples of 100 patients. The numbers of healthy and sick patients included in each sample were set to yield the desired base rates.

We also varied the penalty structure. In the liberal payoff condition, penalties for misses and false alarms were [−5, −1], respectively. This makes it five times more costly to mistakenly diagnose a sick patient as healthy than to falsely diagnose a healthy patient as sick. In the balanced payoff condition, penalties for misses and false alarms were [−1, −1]. In the conservative payoff condition, penalties for misses and false alarms were [−1, −5]. Tolerance values for functionality computations reflected these

[5] These simulations treat recognition rate and validity as independent variables. But recognition rate and validity may covary within individuals (Pachur, 2010). In additional simulations, we found that robustness and stability were minimally affected by correlations between recognition rate and validity.

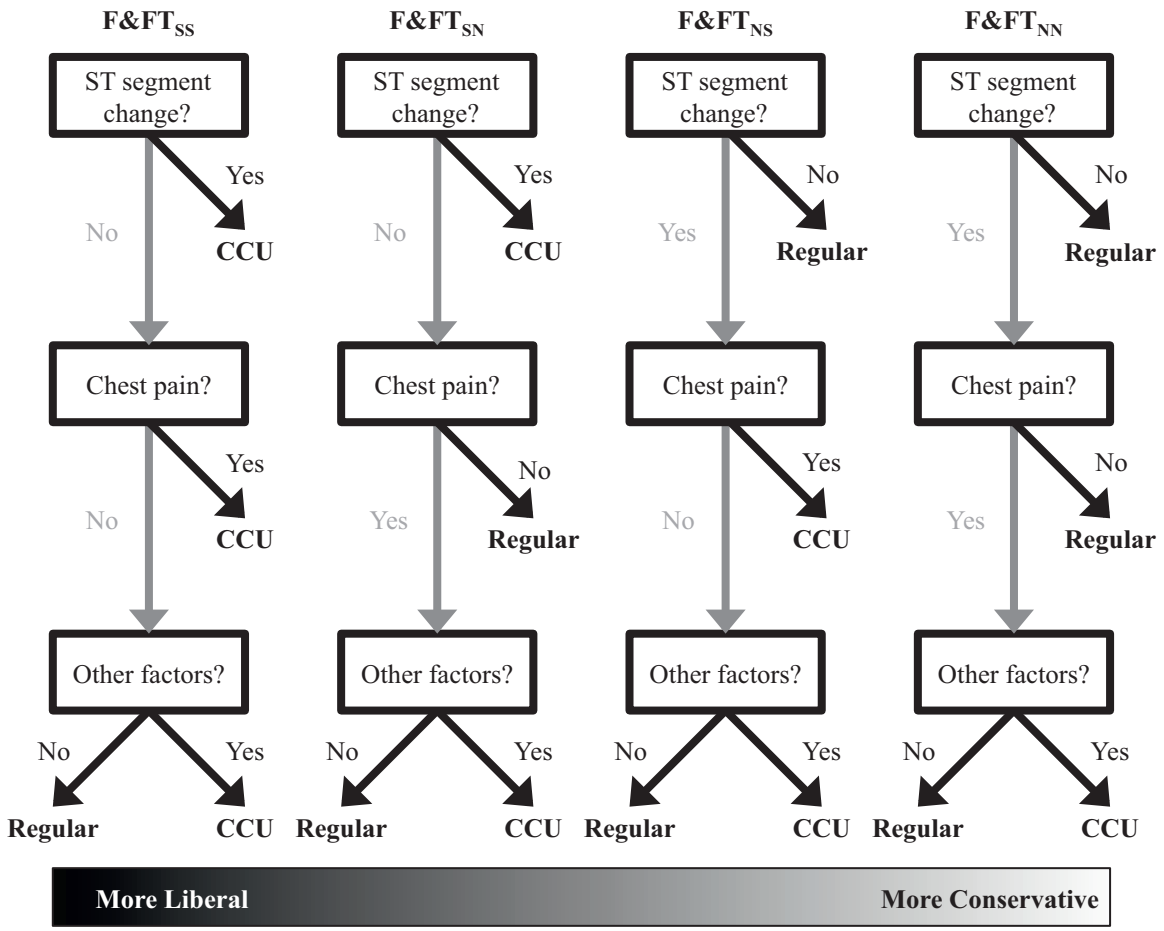

Fig. 5. Four F&FTs with different cue exits. "S" corresponds to signal, and "N" corresponds to noise.
Figure adapted from Luan et al. (2011).

penalties, with tolerance for misses (T_M) set to 0.2, 1.0, and 1.0, in the liberal, balanced, and conservative conditions, respectively. Conversely, tolerance for false alarms (T_{FA}) was set to 1.0, 1.0, and 0.2, across the three conditions. We also created intermediate penalty structures between the liberal and balanced payoff conditions by decreasing the penalty for misses from −5.0 to −1.0 in 0.5 increments, and we created intermediate penalty structures between the conservative and balanced payoff conditions by decreasing the penalty for false alarms from −5.0 to −1.0 in 0.5 increments (yielding 17 unique penalty structures).

For each combination of base rate and penalty structure, we estimated functionality based on 500 simulations. For each simulation, we calculated the percentage of correct responses, misses, and false alarms during 100 trials. To estimate robustness and stability, we created 500 samples that each contained functionality scores from every combination of base rate and penalty structure. Thus, each sample contained 1377 functionality scores (81 base rates × 17 penalty structures). The average of functionality scores within a sample gives an estimate of robustness, and the variance among functionality scores within a sample gives an estimate of stability. Finally, the distributions of estimates from the samples provide information about uncertainty in our measures of robustness and stability.

Fig. 7 shows that the functionality of F&FT$_{SS}$ and F&FT$_{SN}$ dropped in the conservative payoff condition [−1, −5] and when the base rate of illness was low. Because these trees have a liberal bias, they misdiagnosis many healthy individuals when the base rate of illness is low, producing a high rate of false alarms. This is especially problematic when the cost of false alarms is great. Functionality of F&FT$_{NN}$ dropped in the liberal payoff condition [−5, −1] and when the base rate of illness was high. Because F&FT$_{NN}$ has a conservative bias, it fails to diagnose many sick individuals when the base rate of illness is high, producing a high rate of misses. This is especially problematic when the cost of misses is great.

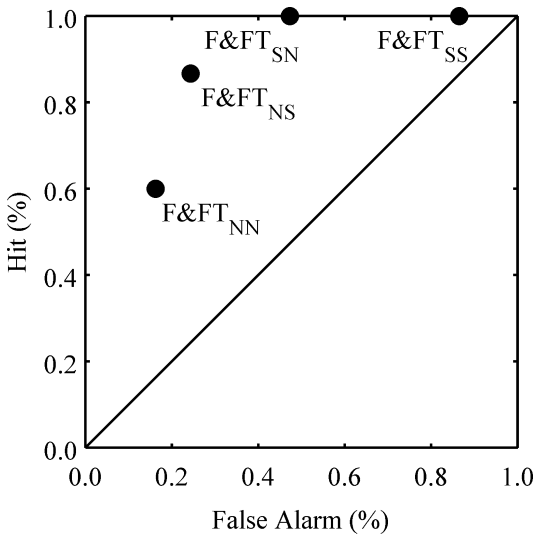

Fig. 6. Hits (diagnosing a sick individual) and false alarms (misdiagnosing a healthy individual) for different F&FTs in the CCU task.
Figure adapted from Luan et al. (2011).

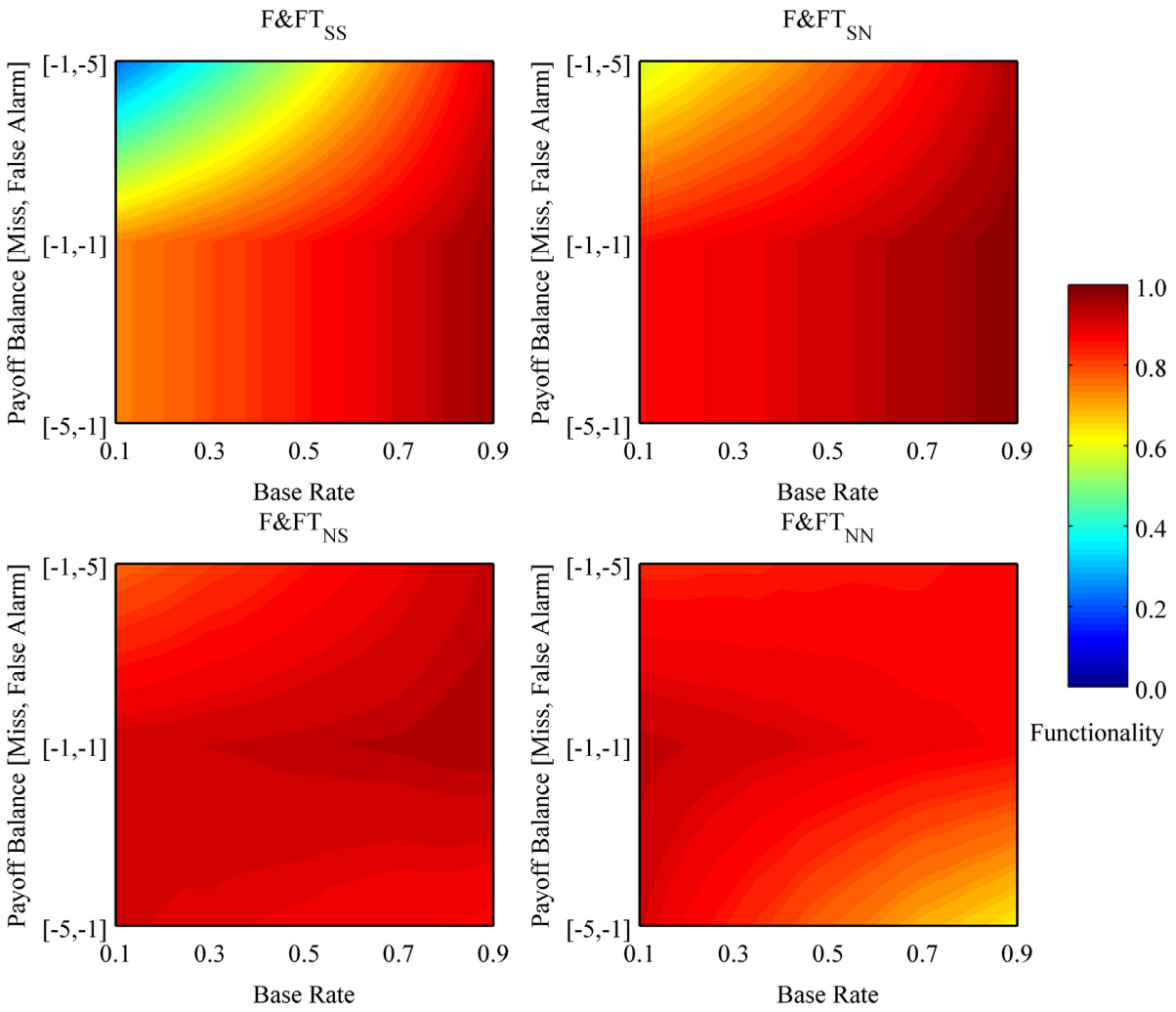

Fig. 7. Functionality of four F&FTs by base rate of illness and payoff balance. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Among the trees, $F\&FT_{NS}$ stands out in that its functionality is relatively high over the range of both dimensions. Of the four trees, $F\&FT_{NS}$ was most robust and stable (Table 4). $F\&FT_{NS}$ did not have the highest functionality in all scenarios. In fact, of the four trees, $F\&FT_{NS}$ had highest functionality in just 41% of scenarios, $F\&FT_{SN}$ in 37% of scenarios, and $F\&FT_{NN}$ in the remaining 22% of scenarios. But unlike the other trees, $F\&FT_{NS}$ never recorded a functionality score below 0.78. Thus, while other trees sometimes perform better, $F\&FT_{NS}$ *always* performs well. In the absence of information about base rate and penalty structure, $F\&FT_{NS}$ can confidently be enacted.

Some previously published research on F&FTs has included comparative analysis with logistic regression models as benchmarks (e.g., Fischer et al., 2002; Martignon et al., 2008). However, the decision optimization available through the use of signal detection theory makes for an even more stringent benchmark assessment. As noted earlier, Luan et al. (2011) mapped F&FT decision processes to SDT. This begs the question: how would the robustness and stability of the SDT model fare in the CCU task relative to the F&FTs? The answer is that, in principle, a SDT model delineates the upper boundary on performance in the CCU task. For each combination of base rate and penalty structure, the SDT model computes and uses a different decision criterion, providing substantial additional flexibility. This makes SDT hard to beat, provided we ignore the complexity of the approach. Indeed, the SDT model was more robust (0.933 ± 0.002 SD) and stable (0.940 ± 0.003 SD) than any of the F&FT models. Thus, it is clear that given complete and certain information about base rates and consequences, and given perfect ability to integrate information from cues, the SDT model produces optimal performance. Yet $F\&FT_{NS}$ achieves 97% of the robustness and 98% of the stability attained by the optimal SDT model without such requirements. An important caveat is that the assumptions regarding information certainty and information processing ability that form the foundation of optimal SDT-based decision-making may not hold in the real world.

Traditionally, operations research professionals have focused on optimal solutions; solutions that are provably best given a set of alternatives, outcomes, probabilities, and constraints that define the choice problem. Rosenhead, Elton, and Gupta (1972), Savage (1954/1972), and Simon (1955) all cautioned against the notion that optimality analyses be used in real-world decision applications. Outcome likelihoods, which can be stated precisely in experimental settings, are typically unknown in real-world

Table 4
Robustness and stability of four F&FTs.

Tree	Robustness	Stability
$F\&FT_{SS}$	.780 ± 0.001 SD	.699 ± 0.003 SD
$F\&FT_{SN}$	.880 ± 0.002 SD	.832 ± 0.004 SD
$F\&FT_{NS}$	.904 ± 0.002 SD	.924 ± 0.003 SD
$F\&FT_{NN}$	.857 ± 0.003 SD	.881 ± 0.004 SD

scenarios. For example, expert climatologists disagree on climate-change models, the inputs to those models, and the likely impacts of different climatic outcomes (Lempert, Nakicenovic, Sarewitz, & Schlesinger, 2004). Because of these sources of uncertainty, traditional decision-analytic methods for risk analysis cannot be applied to climate-change policy. Likewise, in medical decision-making, the prevalence of illnesses in a population is uncertain, and the costs of misdiagnoses are difficult to express as point values (Swets, Dawes, & Monahan, 2000). Further, in eyewitness testimony, the base rate of guilty individuals in police lineups is unspecified, and the consequences of different types of sentencing errors are subjective (Wixted & Mickes, 2012). In these and other scenarios, optimal SDT analysis is not possible.

Swets et al. (2000) expressly called for adaptive statistical prediction rules, which could be trained with one population and that would generalize well to other populations. Our technique for quantifying robustness addresses this issue because it is applied over distributions of values derived from systematic perturbation across dimensions of interest. Our quantification provides a way to predict and assess the extent to which rules (to borrow Swets' terminology) generalize to different populations. Such generalizability is at the heart of our definition and operationalization of robustness.

5. Discussion

Within the psychological literature, the word 'robust' is frequently used, rarely defined, and never quantified. Cognitive science requires sound theory and methods. Although qualitative definitions are useful, quantification is essential to the scientific study of cognition, generally, and robustness, specifically. Without quantification, statements about robustness remain vague at best, and meaningless or even misleading at worst.

The method we propose in this paper makes progress in that direction. This method involves calculating functionality in individual scenarios, assessing the maintenance of functionality (i.e., *robustness*) across scenarios, and measuring stability across scenarios. This method produces two values: robustness, which describes the extent to which the system maintains functionality across the range of scenarios it is liable to encounter, and stability, which describes the extent to which performance of the system is invariant against perturbations. The ideal system is robust *and* stable.

Throughout this paper, we focused on the robustness of decision-making heuristics. A growing literature indicates that heuristics are used not only by participants in laboratory experiments. Pilots use heuristics as they captain aircraft (Gigerenzer, Hertwig, & Pachur, 2011), and judges, customs officers, doctors, and burglars use heuristics to make consequential decisions (Dhami, 2003; Garcia-Retamero & Dhami, 2009; Marewski & Gigerenzer, 2012; Pachur & Marinello, 2013). Our method makes a substantive contribution to this literature by proposing and demonstrating a methodology for moving beyond the realm of semantic references of robustness and into the realm of rigorous measurement of degree of robustness.

The examples provided here demonstrate that decision heuristics are measurably robust and stable against common sources of variation. Additionally, by proposing formal methods and measures, we are able to conduct direct quantitative comparisons of the degree of robustness and stability exhibited by different decision processes. In our first simulation, we found that TTB and TAL are more robust (0.71 and 0.71, respectively) and stable (0.89 and 0.88) against variation in training set size than REGRESS (robustness = 0.67; stability = 0.84). In our second simulation, we found that the recognition heuristic was robust (0.602) and stable (0.861) against empirical variations in recognition rate and validity. In our third simulation, we found that one fast and frugal tree ($F\&FT_{NS}$) for diagnosing myocardial infarction was more robust (0.904) and stable (0.924) against variation in illness base rate and penalty structure than three other possible trees. Importantly, the most robust and stable heuristics did not have highest functionality in all scenarios. But these heuristics also did not have especially low functionality in any one scenario. A key strength of the heuristics-based approach, then, is that these decision strategies can be used with reasonable certainty in a variable, changing world.

Although it is a valid point, our message in this paper is not merely that heuristic-based decision-making can be robust and stable, and demonstrably more so than some more complex decision methods. Rather, the contribution here is that we introduced a method for precisely quantifying degree of robustness and stability. This can serve as the foundation for moving the discourse about robustness in applied cognitive research from qualitative semantic dichotomies to continuously varying quantitative statistics, enabling comparative analyses and informing selection of best courses of action.

5.1. Extensions

Our approach employs an absolute measure of functionality that takes continuous values. Related approaches from biology and engineering also use absolute measures of performance (Bates & Cosentino, 2011). Measures of viability, in contrast, treat performance as a binary variable. Performance that exceeds a pre-established threshold is classified as 'viable', and performance that falls below the threshold is classified as 'non-viable' (Hafner, Koeppl, Hasler, & Wagner, 2009; Larhlimi et al., 2011). Certain problems may warrant quantification in terms of viability, as when exceeding a certain threshold is all that matters. It is a straightforward extension of our approach to define functionality in those terms, and to then compute robustness and stability (Eqs. (4) and (5)).

Yet another, more conservative, approach is to treat robustness as performance in the worst-case scenario (e.g., minimax; Bitmore, 2009). Interestingly, in our comparative simulations, the most robust decision rules happened to have the highest minimum functionality scores, and the least robust decision rules happened to have the lowest minimum functionality scores. But such agreement is not guaranteed. Future studies should compare these approaches to identify their relative strengths and weaknesses.

5.2. Applications

Our approach is not merely descriptive. Robustness analysis can be used to evaluate and prescribe cognitive processes and technologies. For example, in the third simulation, we compared the robustness and stability of four F&FTs, each of which contained different decision biases. A priori, it was unclear which tree was best. $F\&FT_{NS}$ was most robust and stable against variations in base rates and penalties. This suggests that when information about the base rates of illnesses and the consequences of actions is limited, $F\&FT_{NS}$ should be used to make healthcare decisions.

Note, however, that both Green and Mehr (1997) and Luan et al. (2011) recommended the more liberal $F\&FT_{SN}$. Differences in their methodologies and underlying assumptions led to this different conclusion. Green and Mehr did not conduct a robustness analysis, per se. They used sample data from a rural hospital to derive an F&FT ($F\&FT_{SN}$), which they then presented to healthcare providers. Luan et al. (2011) did conduct a type of robustness analysis, but without formally specifying a robustness metric as we have done. Their approach involved generating hypothetical data over variations in cue sensitivity, cue criteria, and intercue correlations. They explicitly assumed lower tolerance for misses (releasing a sick individual)

than false alarms (misdiagnosing a healthy individual), and categorically ruled out more conservative F&FTs ($F\&FT_{NN}$ and $F\&FT_{NS}$). This highlights how the selection of tolerance and dimensions of perturbation affect the computation of robustness and stability. These selections should be tailored to the specific problem. That said, the differences between our conclusions and those of Green and Mehr (1997) and Luan et al. (2011) are more apparent than real. In the analyses presented here $F\&FT_{SN}$ was only slightly less robust than $F\&FT_{NS}$, and both were far superior to $F\&FT_{SS}$ and $F\&FT_{NN}$.

Although the examples in this paper center on decision-making heuristics, the approach is extremely general. Our quantification of robustness can be used to formally quantify the predictive functionality, robustness, and stability of any mathematical, computational, or physical model or system. For example, this methodology may be used to evaluate the outcomes of training and instructional interventions. Key questions for such interventions include whether most students master knowledge, and whether knowledge can be applied in new situations (Koedinger, Corbett, & Perfetti, 2011). Thus, effective interventions must be robust against variations among students, and effective knowledge structures must be robust against variations in the scenarios in which they are needed.

Taatgen, Huss, Dickison, and Anderson (2008) compared the robustness of two sets of instructions in a complex aviation task. The first, list instructions, contained an enumerated list of steps. The second, context instructions, described the conditions for carrying out each step and the results it produced. Although participants in both conditions performed equivalently on easy problems, participants in the list condition performed far worse on medium and difficult problems. Formally, performance was more robust against variation in problem difficulty in the context condition than in the list condition. This and related examples in the areas of unmanned aircraft navigation (Gunzelmann & Gluck, 2009) and landmine detection (Staszewski, 2006), underscore how empirical studies and computational simulations can be used to assess the robustness of training and instructional interventions.

5.3. Conclusion

Robustness is an important construct in domains as diverse as evolutionary biology and structural engineering. Unfortunately, in many domains, most relevantly cognitive science, considerations of robustness end with vague semantic references. Our aim in this research is to supplant qualitative definitions with a quantitative account of robustness. We hope these ideas will stimulate discussion in the scientific community regarding robustness, and ways to quantify it.

Acknowledgements

This research was performed while Matthew Walsh held a National Research Council Research Associateship Award with the U.S. Air Force Research Laboratory's Cognitive Models and Agents Branch and while Evan Einstein held an appointment to the Student Research Participation Program at the U.S. Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Readiness Research Division, Cognitive Models and Agents Branch. The Student Research Participation Program was administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and USAFRL. The authors thank Ron Fisher, Jeremy Knopp, Julian Marewski, John Salerno, John Wixted, Daniel Wright, and an anonymous reviewer for constructive feedback on earlier drafts of this paper.

References

ACT, Inc. (2005). *Developing achievement levels on the 2005 National Assessment of Educational Progress in Grade Twelve Mathematics: Process report*. Iowa City, IA: Author.

Ayton, P., Onkal, D., & McReynolds, L. (2011). Effects of ignorance and information on judgments and decisions. *Judgment and Decision Making*, *6*, 381–391.

Bates, D. G., & Cosentino, C. (2011). Validation and invalidation of systems biology models using robustness analysis. *IET Systems Biology*, *5*, 229–244.

Bitmore, K. (2009). *Rational decisions*. Princeton, NJ: Princeton University Press.

Borrell, B. (2009). What is a bird strike? How can we keep planes safe from them in the future? *Scientific American*,. Retrieved from http://www.scientificamerican.com/article.cfm?id=what-is-a-bird-strike

Brighton, H., & Gigerenzer, G. (2011). Towards competitive instead of biased testing of heuristics: A reply to Hilbig and Richter (2011). *Topics in Cognitive Science*, *3*, 197–205.

Brunswik, E. (1955). Representative design and probabilistic theory in functional psychology. *Psychological Review*, *62*, 193–217.

Canisius, T. D. G. (Ed.). (2011). *Structural robustness design for practising engineers. COST Action TU0601*.

Czerlinski, J., Gigerenzer, G., Goldstein, D. G., & ABC Research Group. (1999). How good are simple heuristics? In G. Gigerenzer, & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 97–118). New York, NY: Oxford University Press.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571–582.

Dhami, M. K. (2003). Psychological models of professional decision making. *Psychological Science*, *14*, 175–180.

Dueck, G. (1993). New optimization heuristics: The great deluge algorithm and the record-to-record travel. *Journal of Computational Physics*, *104*, 86–92.

Eaton, N. R., Krueger, R. F., South, S. C., Simms, L. J., & Clark, L. A. (2011). Contrasting prototypes and dimensions in the classification of personality pathology: Evidence that dimensions, but not prototypes, are robust. *Psychological Medicine*, *41*, 1151–1163.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*, 591–601.

Fischer, J. E., Steiner, F., Zucol, F., Berger, C., Martignon, L., Bossart, W., Altwegg, M., & Nadal, D. (2002). Use of simple heuristics to target macrolide prescription in children with community-acquired pneumonia. *Archives of Pediatrics and Adolescent Medicine*, *156*, 1005–1008.

Fraher, A. L. (2011). Hero-making as a defense against the anxiety of responsibility and risk: A case study of US Airways Flight 1549. *Organisational and Social Dynamics*, *11*, 59–78.

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

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, *3*, 20–29.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristics: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*, 107–143.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669.

Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making*, *6*, 100–121.

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

Glasgow, R. E., Estabrooks, P. A., Marcus, A. C., Smith, T. L., Gaglio, B., Levinson, A. H., & Tong, S. (2008). Evaluating initial reach and robustness of a practical randomized trial of smoking reduction. *Health Psychology*, *27*, 780–788.

Gluck, K. A., McNamara, J. M., Brighton, H., Dayan, P., Kareev, Y., Krause, J., & Wimsatt, W. C. (2012). Robustness in a variable environment. In J. R. Stevens, & P. Hammerstein (Eds.), *Evolution and the mechanisms of decision making* (Strüngmann forum report, vol. 11, J. Lupp, series ed.). Cambridge, MA: MIT Press.

Goldstein, D. G., Gigerenzer, G., & ABC Research Group. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 37–58). London: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90.

Green, L., & Mehr, D. R. (1997). What alters physicians' decisions to admit to the coronary care unit? *Journal of Family Practice*, *45*, 219–226.

Gunzelmann, G., & Gluck, K. A. (2009). An integrative approach to understanding and predicting the consequences of fatigue on cognitive performance. *Cognitive Technology*, *14*, 14–25.

Hafner, M., Koeppl, H., Hasler, M., & Wagner, A. (2009). 'Glocal' robustness analysis and model discrimination for circadian oscillators. *PLoS Computational Biology*, *5*, 1–10.

Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2011). Fluent, fast, and frugal? A formal model evaluation of the interplay between memory, fluency, and comparative judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 827–839.

Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2012). A matter of time: Antecedents of one-reason decision making based on recognition. *Acta Psychologica*, *141*, 9–16.

Hoyer, W. D., & Brown, S. P. (1990). Effects of brand awareness on choice for a common, repeat-purchase product. *Journal of Consumer Research*, *17*, 141–148.
Huaco, D. R., Bowders, J. J., & Loehr, J. E. (2012). Method to develop target levels of reliability for design using LRFD. In *Transportation Research Board 91st annual meeting*. (no. 12-4327).
Jagacinski, R. J., & Flach, J. M. (2003). *Control theory for humans: Quantitative approaches to modeling performance*. Mahwah, NJ: Lawrence Erlbaum Associates.
Katsikopoulos, K. V., Schooler, L. J., & Hertwig, R. (2010). The robust beauty of ordinary information. *Psychological Review*, *117*, 1259–1266.
Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, *5*, 826–837.
Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2011). The knowledge-learning-instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, *36*, 757–798.
Larhlimi, A., Blachon, S., Selbig, J., & Nikoloski, Z. (2011). Robustness of metabolic networks: A review of existing definitions. *Biosystems*, *106*, 1–8.
Lee, J. D., & Kirlik, A. (Eds.). (2013). *The oxford handbook of cognitive engineering*. New York, NY: Oxford University Press.
Lempert, R., Nakicenovic, N., Sarewitz, D., & Schlesinger, M. (2004). Characterizing climate-change uncertainties for decision-makers. *Climatic Change*, *65*, 1–9.
Luan, S., Schooler, L. J., & Gigerenzer, G. (2011). A signal-detection analysis of fast-and-frugal trees. *Psychological Review*, *118*, 316–338.
Marewski, J. N., & Gigerenzer, G. (2012). Heuristic decision making in medicine. *Dialogues in Clinical Neuroscience*, *14*, 77–89.
Marewski, J. N., & Olsson, H. (2009). Beyond the null ritual: Formal modeling of psychological processes. *Journal of Psychology*, *217*, 49–60.
Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review*, *118*, 393–437.
Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology*, *52*, 352–361.
McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*, 239–245.
Newell, B. R., & Shanks, D. R. (2004). On the role of recognition in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 923–935.
Oakley, J. E., & O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *Journal of the Royal Statistical Society: Series B*, *66*, 751–769.
Pachur, T., & Marinello, G. (2013). Expert intuitions: How to model the decision strategies of airport customs officers? *Acta Psychologica*, *144*, 97–103.
Pachur, T. (2010). Recognition-based inference: When is less more in the real world? *Psychonomic Bulletin & Review*, *17*, 589–598.
Pachur, T., Todd, P. M., Gigerenzer, G., Schooler, L. J., & Goldstein, D. G. (2011). The recognition heuristic: A review of theory and tests. *Frontiers in Psychology*, *2*, 1–14.
Reinman, G., Ayer, T., Davan, T., Devore, M., Finley, S., Glanovsky, J., & Yudichak, D. (2012). Design for variation. *Quality Engineering*, *24*, 317–345.
Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207–236.
Rosenhead, J., Elton, M., & Gupta, S. K. (1972). Robustness and optimality as criteria for strategic decisions. *Operational Research Quarterly*, *23*, 413–431.
Savage, L. J. (1972). *The foundations of statistics*. New York, NY: Dover Publications.
Schwartz, B., Ben-Haim, Y., & Dacso, C. (2010). What makes a good decision? Robust satisficing as a normative standard of rational decision making. *Journal for the Theory of Social Behavior*, *41*, 209–227.
Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*, 99–118.
Starossek, U., & Haberland, M. (2012). Robustness of structures. *International Journal of Lifecycle Performance Engineering*, *1*, 3–21.
Staszewski, J. J. (2006). Spatial thinking and the design of landmine detection training. In G. A. Allen (Ed.), *Applied spatial cognition: From research to cognitive technology* (pp. 231–265). Mahwah, NJ: Erlbaum Associates.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.
Taatgen, N. A., Huss, D., Dickison, D., & Anderson, J. R. (2008). The acquisition of robust and flexible cognitive skills. *Journal of Experimental Psychology: General*, *137*, 548–565.
Todorov, A., & Olson, I. R. (2008). Robust learning of affective trait associations with faces when the hippocampus is damaged, but not when the amygdala and temporal pole are damaged. *Social Cognitive and Affective Neuroscience*, *3*, 195–203.
Tomlinson, T., Marewski, J. N., & Dougherty, M. (2011). Four challenges for cognitive research on the recognition heuristic and a call for a research strategy shift. *Judgment and Decision Making*, *6*, 89–99.
US Army. (1990). Military specification: Detecting set, metallic mine, portable. In *Operational Requirements Document MIL-D-0023359G (ME)*.
Van de Werf, F., Bax, J., Betriu, A., Blomstrom-Lundqvist, C., Crea, F., Falk, V., & Weis, M. (2008). Management of acute myocardial infarction in patients presenting with ST-segment elevation. *European Heart Journal*, *29*, 2909–2945.
Wagner, A. (2005). *Robustness and evolvability in living systems*. Princeton, NJ: Princeton University Press.
Weinstein, M. C., & Fineberg, H. V. (1980). *Clinical decision analysis*. Philadelphia, PA: WB Saunders.
Wilcox, R. R. (1998). The goals and strategies of robust methods. *British Journal of Mathematical and Statistical Psychology*, *51*, 1–39.
Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science*, *7*, 275–278.
Wynn, K. (2000). Findings of addition and subtraction in infants are robust and consistent: Reply to Wakeley, Rivera, and Langer. *Child Development*, *71*, 1535–1536.
Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517.
Zhou, K. J., & Doyle, J. C. (1997). *Essentials of robust control*. Upper Saddle River, NJ: Prentice-Hall.