Performance of M-ary Soft Fusion Systems Using Simulated Human Responses

Donald J. Bucci, Sayandeep Acharya, and Moshe Kam

Department of Electrical and Computer Engineering, Drexel University, Philadelphia, Pennsylvania, 19104, USA

***Abstract*—A major hurdle in the development of soft and hard/soft data fusion systems is the inability to determine the practical performance gains between fusion operators without the burdens associated with human testing. Drift diffusion models of human responses (i.e., decision, confidence assessments, and response times) from cognitive psychology can be used to gain a sense of the performance of a fusion system during the design phase without the need for human testing. The majority of these models were developed for binary decision tasks, and furthermore, the few models which can operate on M-ary decision tasks are yet unable to generate subject confidence assessments. The current study proposes a method for realizing human responses over an M-ary decision task using pairwise successive comparisons of related binary decision tasks. We provide an example based on the two-stage dynamic signal detection models developed by Pleskac and Busemeyer (2010) where subjects were presented with a pair of lines on a computer screen, asked to determine which of two lines was the longest, and to assess their confidence in their decision using a subjective probability scale. M-ary human opinions were simulated for this line length task and used to assess the performance of several fusion rules, namely: Bayes' rule of probability combination, Dempster's Rule of Combination (DRC), Yager's rule, Dubois and Prade's rule (DPR), and the Proportional Conflict Redistribution rule #5. When taking source reliability into account in the combination, Bayes' rule of probability combination and DRC exhibited the most accurate performance (i.e., the largest amount of specific evidence committed towards the true outcome) for this task. Yager's rule and DPR exhibited inferior performance across all simulated cases.**

***Index Terms*—Data fusion, Dempster-Shafer Theory, Human Simulation, Expert reasoning systems, Belief fusion**

I. Introduction

The use of human opinions in data fusion systems is a current topic of interest. Human-generated data, often categorized as "soft data," may provide a level of insight and intuition that is not always captured by electronic, optical, mechanical, or other "hard" sensors. However, it is not easy to develop a statistical characterization for human decision makers [1]. Methods for determining the performance of data fusion systems involving inputs from humans rely mostly on the use of examples/counterexamples (e.g., [2], [3]) or on predetermined data sets that were developed through direct human testing (e.g., [4]–[6]). Models of human decision-making from cognitive psychology present an opportunity to simulate the performance of soft and hard/soft fusion systems flexibly and accurately without many of the burdens associated with human testing.

The majority of studies which have employed models of human decision makers looked at how task reward structures influence human decision-making strategies when the human acts as a *director* of information (e.g., when humans make choices in response to evolving system performance metrics, as in [7]). Much less work has been done on using human decision making models for assessing the performance of fusion systems in which the human acts as a *source* of information (e.g., when humans make choices regarding the state of a certain phenomenon and assess their level of confidence in these choices, as in [6], [8]).

Drift diffusion models [9] of human responses have been proposed in cognitive psychology as a means of accurately capturing the dynamics and relationships present between human decision-making and response time on both binary (e.g., [10]) and M-ary (e.g., [11]) decision problems. Little work has been done regarding the incorporation of human confidence assessments in such drift diffusion models. Furthermore, the majority of effort in this area has been focused on binary decision problems [12]. We have previously shown how it is possible to assess fusion performance using models of binary human responses in [13], [14]. The current study proposes a method for extending drift diffusion models of human decision making, confidence assessment, and response time to related *multihypothesis* (M-ary) decision tasks. Specifically, we make use of the *two-stage dynamic signal detection* (2DSD) model of [12] to produce subjective probabilities on a M-ary decision task using pairwise successive comparisons of binary decision tasks. As a motivating example, we use the 2DSD human parameters estimated in [12] relating to a *line length discrimination task*, in which the authors positioned subjects in front of a computer monitor, presented the subjects with two lines at a time, and asked them to provide a decision and confidence rating on which line was longer. We apply a successive pairwise comparison algorithm to the binary line length discrimination task to simulate human responses on an M-ary line length discrimination task (i.e., subjects are instructed to choose and assess decision confidence for the longest line amongst M lines). The subjects from the line length discrimination task in [12] and our successive pairwise comparison technique are used to assess the accuracy and precision of combining human responses using Bayes' rule of probability combination (i.e., Bayes' Theorem), Dempster's Rule of Combination (DRC), Yager's Rule, Dubois and Prade's rule (DPR), and the Proportional Conflict Redistribution Rule

#5 (PCR5) under varying numbers of decision alternatives (i.e., sets of lines differing in length).

The remainder of this work is organized as follows. Section II describes the 2DSD human response model employed here as it relates to the line length discrimination task example of [12]. Section III describes the formulation of the M-ary extension methodology using pairwise successive comparisons. Section IV describes an M-ary fusion simulation for the line length discrimination task using Bayes' rule of probability combination, DRC, Yager's Rule, DPR, and the PCR5. Each fusion operator was used to combine belief mass assignments generated using the M-ary extension methodology and 2DSD models provided in [12]. The performance of each operator was determined by calculating the average nearness of the combined BMAs to a BMA which assigns the true outcome full belief. The results of the simulation are described in Section V. After combination of twelve or more sources, Bayes' rule of probability combination and DRC were found to be the most accurate when statistical evidence relating to the subject's ability to make accurate confidence assessments was available. PCR5 was found to be at least as accurate as the best decision-maker in the combination across all fusion cases. Yager's rule and Dubois and Prade's rule exhibited inferior performance.

II. Human Simulation Methodology

A. Two-Stage Dynamic Signal Detection [12]

Two-stage dynamic signal detection (2DSD) is a recently developed model that accounts for a wide range of phenomena in human decision making, while also taking into account the modeling of confidence assessments [12]. Let $\mathcal{A} = \{A, \overline{A}\}$, where $\mathcal{A}$ represents a binary decision task consisting of the alternatives A and $\overline{A}$. In 2DSD, the internal evidence accumulated in favor of the alternative A over $\overline{A}$ at time t (i.e., $L(t)$) is given by the stochastic difference equation

$$\Delta L(t) = \delta \Delta t + \sqrt{\Delta t}\, \epsilon(t + \Delta t), \quad L(0) = L_0, \tag{1}$$

where δ is known as the *drift rate* and $\epsilon(t)$ is a simulated white noise process with zero mean and variance σ^2. The value σ is known as the *drift coefficient*. The drift rate δ is either positive or negative, depending on whether A or $\overline{A}$ is true. To account for trial variability, the drift rate δ and the initial condition L_0 can be chosen on a per trial (or per simulation) basis via $\delta \sim N(\nu, \eta^2)$ (normally distributed) and $L_0 \sim U(-0.5s_z, 0.5s_z)$ (uniformly distributed); here ν and η are the subject mean drift rate and drift rate standard deviation respectively, and $s_z \in [0, \infty)$ is the size of an interval containing the initial condition L_0. The evidence accumulation is simulated until a threshold, either $\theta_A, \theta_{\overline{A}}$, is crossed (where $-\theta_{\overline{A}} < L_0 < \theta_A$). A decision $a \in \mathcal{A}$ is determined such that

$$a = \begin{cases} A & L(t) > \theta_A \\ \overline{A} & L(t) < -\theta_{\overline{A}} \\ \text{wait} & \text{otherwise} \end{cases}. \tag{2}$$

Let $\mathbf{P}^{(a)} = [p_1^{(a)} \cdots p_{K_a}^{(a)}]$ denote the K_a possible confidence values associated with choosing $a \in \mathcal{A}$ at time t_d. The assigned confidence level $p \in \mathbf{P}^{(a)}$ associated with deciding a after waiting $t_c = t_d + \tau$ is given as

$$p = p_i^{(a)} \quad \text{when} \quad L(t_c) \in [c_{i-1}^{(a)}, c_i^{(a)}], \tag{3}$$

where $c_0^{(a)} = -\infty$ and $c_{K_a}^{(a)} = \infty$ for each $a \in \mathcal{A}$. The value τ is known as the *interjudgment time*. The remaining confidence bin parameters $\mathbf{C}^{(a)} = [c_1^{(a)} \cdots c_{K_a-1}^{(a)}]$ are chosen such that $c_{i-1} < c_i$ for each $i \in \{1, \ldots, K_a - 1\}$ and each $a \in \mathcal{A}$.

In summary, a 2DSD realization produces the *subjective probability assignment*

$$P_{\mathcal{A}}(a) = p, \tag{4}$$

$$P_{\mathcal{A}}(\overline{a}) = 1 - p. \tag{5}$$

The authors in [12] suggest the following additional parameter restrictions to simplify the 2DSD implementation.

- The decision thresholds for both alternatives can be chosen symmetrically (i.e., $\theta_{\overline{A}} = \theta_A = \theta$).
- The confidence bins for both choices can be set equal (i.e., $\mathbf{C}^{(A)} = \mathbf{C}^{(\overline{A})} = \mathbf{C}$).
- The confidence values can be fixed for all subjects and all alternatives (e.g., $\mathbf{P}^{(A)} = \mathbf{P}^{(\overline{A})} = [0.50, 0.60, \cdots, 1.00]$).
- The drift coefficient can be fixed for each subject (e.g., $\sigma = 0.1$).

Applying these simplifications results in the 10-tuple,

$$\mathcal{S} = \{\nu, \eta, s_z, \theta, \tau, c_1, c_2, c_3, c_4, c_5\}. \tag{6}$$

for each subject. The authors of [12] suggest using the *quantile maximum probability* method [15] to estimate $\mathcal{S}$ using statistics relating to subject decisions, confidence assessments, and response times.

B. Binary Line Length Task Overview [12]

In the *line length discrimination task* modeled in [12], subjects were shown a 32.00 millimeter long line paired with either a 32.27, 32.59, 33.23, 33.87, or a 34.51 millimeter long line. For each given line pairs, the subjects were asked to identify which of the two lines was longer, and assess their decision confidence using the subjective probability scale $\{0.50, 0.60, \cdots, 1.00\}$. The time step of the simulator was fixed in [12] at $\Delta t = 0.001$ for each subject. Five different mean drift rates, ν_1 though ν_5, were found for each subject relating to the side-by-side comparison of the 32 millimeter long line compared with the 32.27, 32.59, 33.23, 33.87, and 34.51 millimeter long lines respectively. The parameter values used to simulate each subject can be found in [12, Tables 3 and 6]. Also, in [12, Table 6] separate decision thresholds θ were determined for two cases of the line length discrimination task, namely when subjects were asked to focus on *fast* responses and when subjects were asked to focus on *accurate* responses. In the present study, the values of θ which represent the subjects focusing on accurate responses were used.

C. Out-of-Sample Prediction

For the line length discrimination task of [12], the five different mean drift rates, ν_1 though ν_5 relate to five specific tasks comparing a 32.00 millimeter long line with either a 32.27, 32.59, 33.23, 33.87, or a 34.51. We perform linear interpolation to estimate mean drift rates for line comparisons not investigated in [12]. Let Δl represent the length difference between each line, such that

$$\Delta l = l_R - l_L, \tag{7}$$

where l_R and l_L represent the lengths of the right and left lines as presented to the subjects. Linear regression was applied to the coordinate pairs $(\Delta l, \nu)$ for each subject of the line length discrimination task in [12] (Figure 1). All subject drifts rates appear to follow a linear relationship. With this relationship in mind, the human parameter sets in equation (6) can be rewritten with $\nu = \nu_m \Delta l$, such that

$$\mathcal{S}(\Delta l) = \{\nu_m \Delta l, \eta, s_z, \theta, \tau, c_1, c_2, c_3, c_4, c_5\}. \tag{8}$$

Here ν_m is the slope of the linear fit as given for each subject in Figure 1.

III. M-ary Extension Methodology

A. Successive Pairwise Comparison Aggregation

Using the out-of-sample prediction method described by the parameter set $\mathcal{S}(\Delta l)$ in (8), we can formulate an M-ary human simulator which determines the longest line using successive pairwise comparisons. Let the lengths of M lines be given by $\mathbf{L}_M = [l_1, \ldots, l_i, \ldots, l_M]$, and let ω_i represent the event that the i^{th} line is deemed the longest. We denote $\Omega_M = \{\omega_1, \ldots, \omega_i, \ldots, \omega_M\}$ and $\mathbb{N}_M = \{1, 2, \ldots, M\}$. Given a group of N human sources, let $P_\Omega^{(n)}(\omega_i)$ represent the *subjective probability function* which describes the n^{th} source's confidence towards each $\omega_i \in \Omega_M$, where $n \in \mathbb{N}_N$. Assuming that the line lengths described by $\mathbf{L}_M$ are distinct, a unique maximum l_{i^*} must exist in $\mathbf{L}_M$. The subjective probabilities $P_\Omega^{(n)}(\omega_i)$ can be represented as

$$P_{\Omega_M}^{(n)}(\omega_i) = P_{\Omega_M}^{(n)}\left((l_i = l_{i^*}) | l_{i^*} \in \mathbf{L}_M\right). \tag{9}$$

Expanding the conditional probability yields

$$\begin{aligned} P_{\Omega_M}^{(n)}(\omega_i) =& \frac{P_{\Omega_M}^{(n)}\left((l_i = l_{i^*}) \bigcap (l_{i^*} \in \mathbf{L}_M)\right)}{P_{\Omega_M}^{(n)}(l_{i^*} \in \mathbf{L}_M)} \\ =& \frac{P_{\Omega_M}^{(n)}\left((l_i = l_{i^*}) \bigcap \left(\bigcup_{\hat{i}=1}^{M} l_{\hat{i}} = l_{i^*}\right)\right)}{P_{\Omega_M}^{(n)}\left(\bigcup_{\hat{i}=1}^{M} l_{\hat{i}} = l_{i^*}\right)}. \end{aligned} \tag{10}$$

Since we assumed the existence of a unique maximum length l_{i^*}, we have that $(l_i = l_{i^*}) \cap (l_{\hat{i}} = l_{i^*}) = \emptyset$ for all $i, \hat{i} \in \mathbb{N}_M$ where $i \neq \hat{i}$. Hence (9) can thus be reduced to

$$P_{\Omega_M}^{(n)}(\omega_i) = \frac{P_{\Omega_M}^{(n)}(l_i = l_{i^*})}{\sum_{\hat{i}=1}^{M} P_{\Omega_M}^{(n)}(l_{\hat{i}} = l_{i^*})}. \tag{11}$$

The event $(l_i = l_{i^*})$ can be thought of as every other line l_j having shorter length then l_i, where $j \neq i$. Hence,

$$(l_i = l_{i^*}) = \bigcap_{\substack{j \in \mathbb{N}_M \\ j \neq i}} (l_i > l_j). \tag{12}$$

Combining (11) and (12), and assuming that belief in $(l_i > l_j)$ is independent for all $i, j \in \mathbb{N}_M$ yields

$$P_{\Omega_M}^{(n)}(\omega_i) = \frac{\prod_{\substack{j \in \mathbb{N}_M \\ j \neq i}} P_{\mathcal{A}}^{(n)}(l_i > l_j)}{\sum_{\hat{i}=1}^{M} \prod_{\substack{j \in \mathbb{N}_M \\ j \neq \hat{i}}} P_{\mathcal{A}}^{(n)}(l_{\hat{i}} > l_j)}, \tag{13}$$

where $\mathcal{A} = \{(l_i > l_j), (l_i < l_j)\}$. The subjective probabilities $P_{\mathcal{A}}^{(n)}(l_i > l_j)$ for any $i, j \in \mathbb{N}_M$ can be realized using the 2DSD human tuples of [12], and applying the linear fits for the mean drift rates as shown in (8). Suppose that $a \in \mathcal{A}$ and $p \in [0, 1]$ are, respectively, the decision and confidence values associated with a realization of the n^{th} subject using the 2DSD algorithm. The probability assignment for the event $(l_i > l_j)$ for any $i, j \in \mathbb{N}_M$ is

$$P_{\mathcal{A}}^{(n)}(l_i > l_j) = \begin{cases} p & a = (l_i > l_j), \\ 1 - p & a = (l_i < l_j). \end{cases} \tag{14}$$

After realizing $P_{\mathcal{A}}^{(n)}(l_i > l_j)$ for every $i, j \in \mathbb{N}_M$, (13) can be used to create the belief probabilities associated with each line length in $\mathbf{L}_M$ being the longest (i.e., $P_{\Omega_M}^{(n)}(\omega_i)$). The longest line can be determined by choosing the ω_i with highest belief, that is

$$w_{i^*} = \arg \max_{\omega_i \in \Omega_M} \left(P_{\Omega_M}^{(n)}(\omega_i)\right), \tag{15}$$

with a corresponding confidence value of $p_{i^*} = P_{\Omega_M}^{(n)}(\omega_{i^*})$. Since 2DSD models choose from a finite set of confidence values [12], the following three cases can occur: ω_{i^*} is unique, ω_{i^*} is not unique, or ω_{i^*} does not exist because the denominator of (13) is zero. In the second case, a decision can be made by choosing one of the ω_{i^*} at random (i.e., assuming all are equally likely). In the third case, a decision cannot be reached and a "no decision" state is returned.

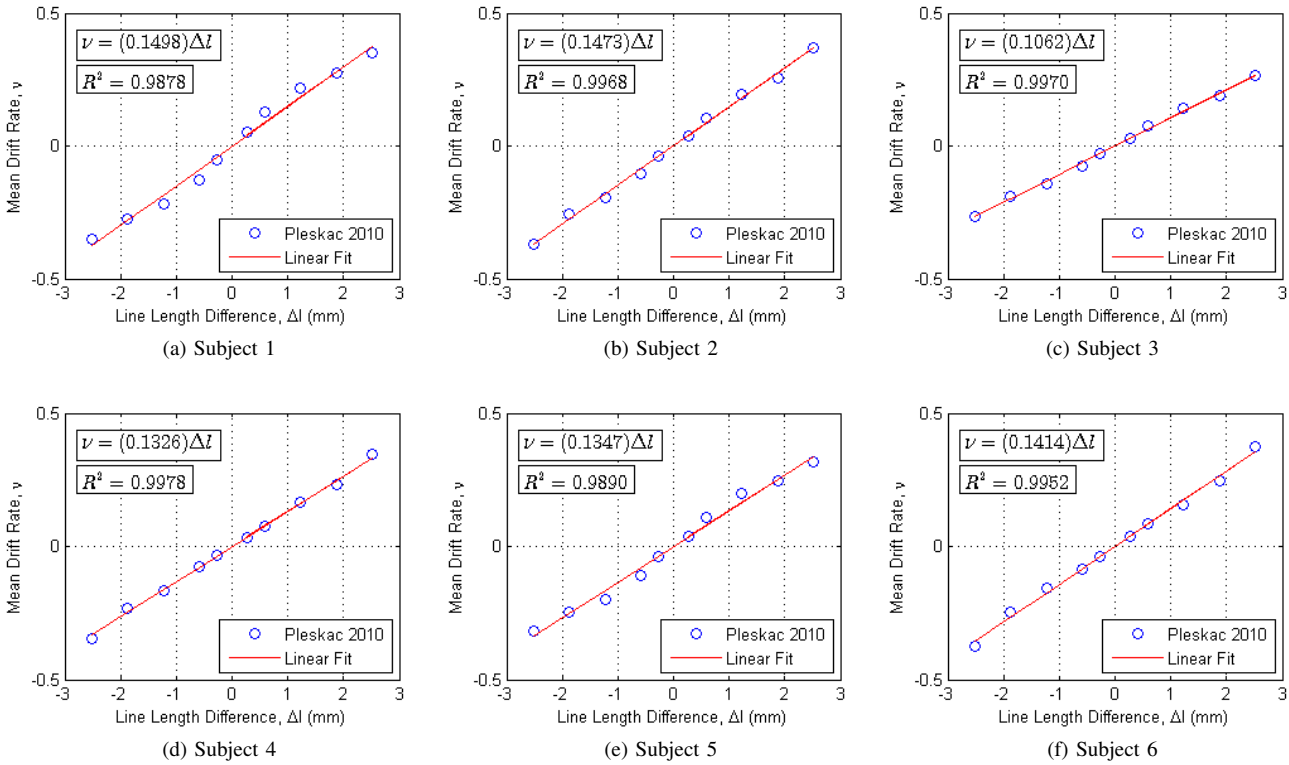
B. Assessing Subject Performance

Similar to equations (4) and (5), the 2DSD-based M-ary line length discrimination task simulator yields a single decision amongst M alternatives (denoted ω_{i^*}), and a corresponding decision confidence (denoted p_{i^*}). Writing this decision and confidence pair as a subjective probability assignment yields

$$P_{\Omega_M}^{(n)}(\omega_{i^*}) = p_{i^*} \tag{16}$$

$$P_{\Omega_M}^{(n)}(\overline{\omega}_{i^*}) = 1 - p_{i^*} \tag{17}$$

for a given subject $n \in \mathbb{N}_N$, where $\overline{\omega}_{i^*} \subset \Omega_M$ represents the negation of ω_{i^*}. Let $\omega_i^* \in \Omega_M$ represent the true outcome of Ω_M. Ideally, subject n should assign full belief (i.e., probability one) to the correct outcome ω_i^*. As subject n assigns less belief to ω_i^*, the quality of this person's opinion



(a) Subject 1 (b) Subject 2 (c) Subject 3

(d) Subject 4 (e) Subject 5 (f) Subject 6

Fig. 1. Linear fits of subject mean drift rates versus line length differences for the line length discrimination task as presented in [12]. Equations and R^2 values shown for each subject.

decreases. Motivated by [12], we denote this idea of subject opinion quality as *evidence strength*, ξ, where

$$\xi(\omega_{i^*}, p_{i^*}|\omega_i^*) = \begin{cases} 1-(1-p_{i^*})^2 & \omega_{i^*} = \omega_i^* \\ 1-(p_{i^*})^2 & \omega_{i^*} \neq \omega_i^* \end{cases}. \quad (18)$$

Evidence strength is derived from the quadratic scoring rule known as *Brier score* [16]. An evidence strength value of one means that the subject has chosen the correct outcome and assigned it probability one. An evidence strength value of zero means that the subject has chosen the incorrect outcome, and has assigned it probability one.

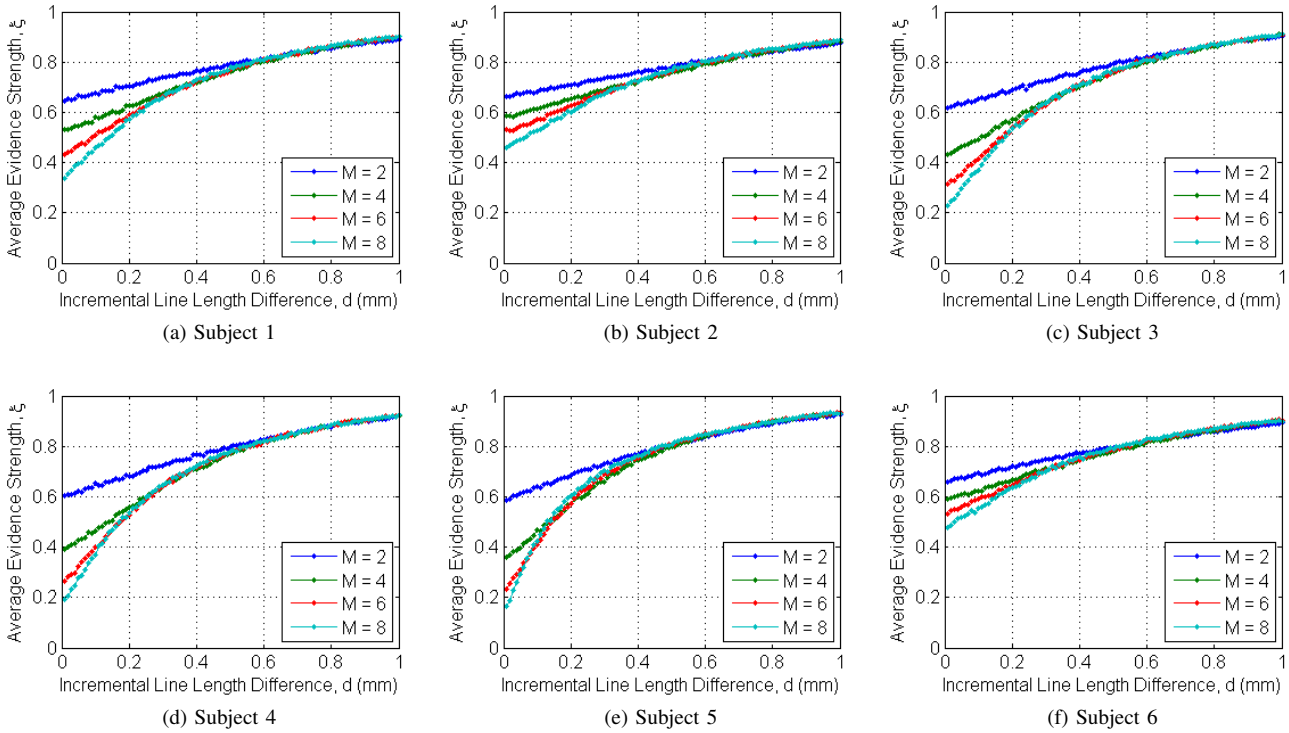
C. Simulated Performance of Subjects

All six subject tuples from [12, Tables 3 and 6] were simulated using the human tuples described by (8) and the mean drift rate regressions of Figure 1. The line lengths presented to the simulated subjects were $\mathbf{L}_M = \{32, 32+d, 32+2d, \ldots, 32+(M-1)d\}$ where M was the number of lines being compared and d was the incremental length difference between lines, in millimeters. Subject decisions and confidence assessments were generated using the successive pairwise comparison aggregation method of Section III-A. The evidence strengths of each subject were determined and averaged over 10,000 trials from $d = 0.01$ to $d = 1.0$ in increments of 0.01 and for $M = 2$, 4, 6, and 8. For each subject, trials which produced the "no decision" state were repeated until a decision and confidence value were reached.

Figure 2 shows the average evidence strength, ξ, of each subject versus the incremental line length difference, d. Evidence strengths for different numbers of alternatives M for each subject are also shown. As expected, increasing the perceptual difficulty of the task (i.e., decreasing d) decreased subject performance. For large enough d (e.g., $d > 0.60$), increasing the number of alternatives was found to have little effect on subject performance. For smaller d (e.g., $d < 0.40$), increasing the number of alternatives caused the largest decrease in performance when going from $M = 2$ to $M = 4$ alternatives. For $M > 4$ however, the subject performance was similar to the $M = 4$ case. This outcome seems logical, as increasing the number of alternatives without changing the task difficulty will result in some alternatives being easier to rule out than others (e.g., the shortest lines will be more easily discernible).

D. Assumptions and Limitations

Our method assumes that the subjects perform pairwise successive comparisons on every single possible pair of alternatives amongst a larger set of alternatives. For the line length discrimination task, any two lines in a set of lines which are clearly different in length would result in larger mean drift rate values, which according to 2DSD will produce exponentially faster response times [12]. According to our extension methodology, simulated subjects will spend less time deliberating between pairs of lines which are clearly different



(a) Subject 1 (b) Subject 2 (c) Subject 3

(d) Subject 4 (e) Subject 5 (f) Subject 6

Fig. 2. Simulated averages of evidence strengths, ξ, for all six 2DSD subject models from [12, Tables 3 and 6] under the 2DSD M-ary human response simulator for the line length discrimination task versus the incremental line length difference, d. Average evidence strengths shown for $M = 2, 4, 6, 8$ alternatives. Averages obtained over 10,000 trials of the M-ary simulation algorithm for each subject, while repeating trials which produced the no decision state for each subject.

in length. In reality, human subjects may not even make use of a pairwise comparison technique for lines which are clearly different in length. The methodology also assumes that a linear relationship for the out-of-sample prediction best describes how changing observations (i.e., line lengths) influences the mean drift rate parameter. Although the information in Figure 1 seems to support this notion, a linear relationship for the out-of-sample prediction may become less accurate as the line length difference approaches zero and the subjects reach their perceptual limit.

IV. Fusion Algorithm Simulation

We used the proposed M-ary human response simulation method to assess the combination of human responses using several belief fusion operators. They are Bayes' rule of probability combination [17]; Dempster's Rule of Combination (DRC) [18]; Yager's Rule [19]; Dubois and Prade's rule (DPR) [20]; and the Proportional Conflict Redistribution Rule #5 [21]. The literature provides an abundance of information on the implementation of various Dempster-Shafer theory concepts (e.g., [18], [21], [22]). For the sake of brevity, we will focus here on explaining only those concepts which are pertinent to the fusion simulation examples presented in this study.

A. Fusion Operator Inputs

For each subject $n \in \mathbb{N}_N$, let the simulated decision and confidence values be given as $\omega_{i^*}^{(n)} \in \Omega_M$ and $p_{i^*}^{(n)} \in [0, 1]$. Excluding Bayes' rule of probability combination, the belief fusion operators investigated here use inputs known as *belief mass assignments* (BMA). BMAs can be thought of as assessing evidence on the powerset of alternatives, allowing the user to specify evidence imprecisely (i.e., evidence towards a disjunction of alternatives rather than the alternatives themselves) [18]. In the current study, the BMAs $m_n(X)$ were formulated for each subject such that

$$m^{(n)}(X) = \begin{cases} p_{i^*}^{(n)} & X = \omega_{i^*}^{(n)} \\ 1 - p_{i^*}^{(n)} & X = \Omega_M \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

for all BMAs. Fusion using Bayes' rule of probability combination was performed on the subjective probability assignments defined for each $\omega \in \Omega_M$ as

$$P^{(n)}(\omega) = \begin{cases} p_{i^*}^{(n)} & \omega = \omega_{i^*}^{(n)} \\ \frac{1}{(M-1)} \left(1 - p_{i^*}^{(n)}\right) & \omega \neq \omega_{i^*}^{(n)} \end{cases}. \tag{20}$$

A vacuous BMA [18] or an equiprobable subjective probability assignment was used whenever the simulated subjects returned the "no decision" state. For each fusion method evaluated, the source combination order was chosen by sampling each of the thirty-six sources with equal probability.

We defined evidence strength ξ in (18). Let $\overline{\xi}^{(n)}$ be the average evidence strength of the n^{th} source. For BMAs,

it is possible to account for source reliability through the *discounting operation* [22]

$$m^{(n)}(X;\overline{\xi}^{(n)}) = \begin{cases} \overline{\xi}^{(n)} m(X) & X \neq \Omega_M \\ \overline{\xi}^{(n)} m(X) + (1-\overline{\xi}^{(n)}) & X = \Omega_M \end{cases}. \tag{21}$$

We define the *discounting operation* analogously for subjective probabilities in [13] as

$$P^{(n)}(\omega;\overline{\xi}^{(n)}) = \overline{\xi}^{(n)} P^{(n)}(\omega) + |\Omega_M|^{-1}(1-\overline{\xi}^{(n)}), \tag{22}$$

where $|\Omega_M|$ is the cardinality of Ω_M.

B. Fusion Operator Performance Metrics

The combination methods we study here can all be thought of as producing a class of subjective probability assignments [23] on each of the $\omega \in \Omega_M$ defined by the belief and plausibility [18] ranges $[\text{Bel}(\omega), \text{Pl}(\omega)]$, where

$$\text{Bel}(X) = \sum_{\substack{Z \subseteq \Omega \\ Z \subseteq X}} m(Z) \Longrightarrow \text{Bel}(\omega) = m(\omega), \tag{23}$$

and

$$\text{Pl}(X) = \sum_{\substack{Z \subseteq \Omega \\ Z \cap X \neq \emptyset}} m(Z) \Longrightarrow \text{Pl}(\omega) = \sum_{\substack{Z \subseteq \Omega \\ \omega \in Z}} m(Z). \tag{24}$$

In the case of Bayes' rule of probability combination, a single subjective probability assignment is produced. Similar to the performance of the subjects in Section III-B, we take the performance of the fusion operators as a measure of the nearness of the subjective probability assignments to one which assigns the truth $\omega^* \in \Omega_M$ probability one. The result is a class of evidence strengths defined by the intervals $[\xi_{\text{Bel}}, \xi_{\text{Pl}}]$, where

$$\xi_{\text{Bel}} = \xi(\omega^*, \text{Bel}(\omega^*)|\omega^*), \tag{25}$$

and

$$\xi_{\text{Pl}} = \xi(\omega^*, \text{Pl}(\omega^*)|\omega^*). \tag{26}$$

The lower envelope ξ_{Bel} can be thought of as a measure of the *accuracy* of the combination operator, and the size of the interval $(\xi_{\text{Pl}} - \xi_{\text{Bel}})$ can be thought of as the *precision* of the combination operator. Accurate belief combination operators will tend to assign probability one to the correct outcome, resulting in values of ξ_{Bel} close to one. Precise belief combination operators will tend to produce more specific evidence, resulting in values of $(\xi_{\text{Pl}} - \xi_{\text{Bel}})$ being close to zero, and hence $1-(\xi_{\text{Pl}} - \xi_{\text{Bel}})$ would be close to one. Since $\text{Bel}(\omega^*) \leq \text{Pl}(\omega^*) \leq 1$ [18], it follows that $\xi_{\text{Bel}} \leq \xi_{\text{Pl}} \leq 1$. Hence systems with high accuracy (i.e., ξ_{Bel} close to one) will also be very precise (i.e., $(\xi_{\text{Pl}} - \xi_{\text{Bel}})$ close to zero). In the Bayesian case, $(\xi_{\text{Pl}} - \xi_{\text{Bel}}) = 0$ since $\text{Bel}(\omega^*) = \text{Pl}(\omega^*) = P(\omega^*)$ [18].

C. Simulation Overview

The M-ary human response simulator of Section III was used to simulate decisions and confidence values over 10,000 trials using six responses from each subject in [12, Tables 3 and 6] under the line length discrimination task (i.e., thirty six total sources). Subjects were simulated using the line length differences $\mathbf{L}_M = \{32, 32+d, 32+2d, \ldots, 32+(M-1)d\}$, using an incremental line length difference $d = 0.20$ mm and $M = 2$, $M = 4$, and $M = 8$ alternatives. The performance metrics ξ_{Bel} and $1-(\xi_{\text{Pl}} - \xi_{\text{Bel}})$ (i.e., accuracy and precision) of each fusion method were determined and averaged over the 10,000 trials of the simulation.

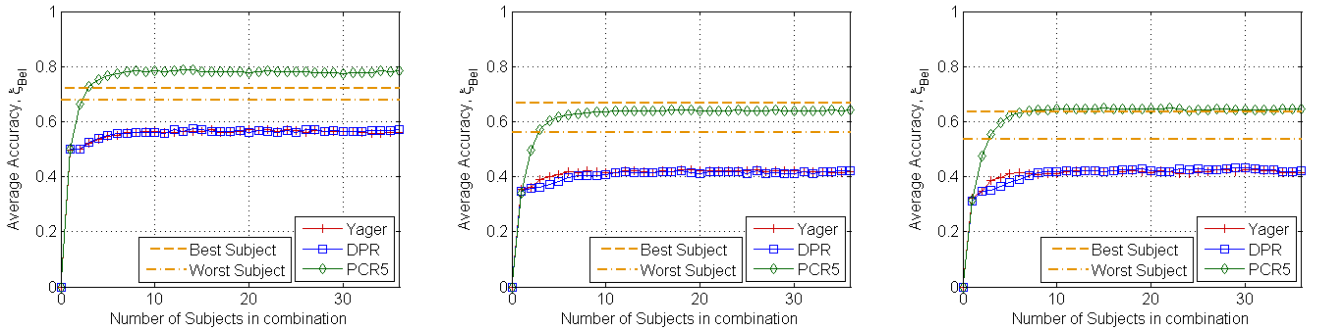
V. Results

Figure 3 shows the accuracy performance (i.e., ξ_{Bel}) for each of the five fusion methods mentioned in Section IV versus the number of sources present in combination. The accuracy performance (i.e., evidence strengths) of the best and worst subjects are shown for comparison. Similarly, Figure 4 shows the precision performance (i.e., $1-(\xi_{\text{Pl}} - \xi_{\text{Bel}})$) for each of the five fusion methods mentioned in Section IV. The subplots of both Figure 3 and Figure 4 show the number of alternatives M simulated and the results obtained from performing or not performing the evidence strength discounting of equations (21) and (22). Higher values of ξ_{Bel} and $1-(\xi_{\text{Pl}} - \xi_{\text{Bel}})$ indicate higher combination accuracy and precision respectively.

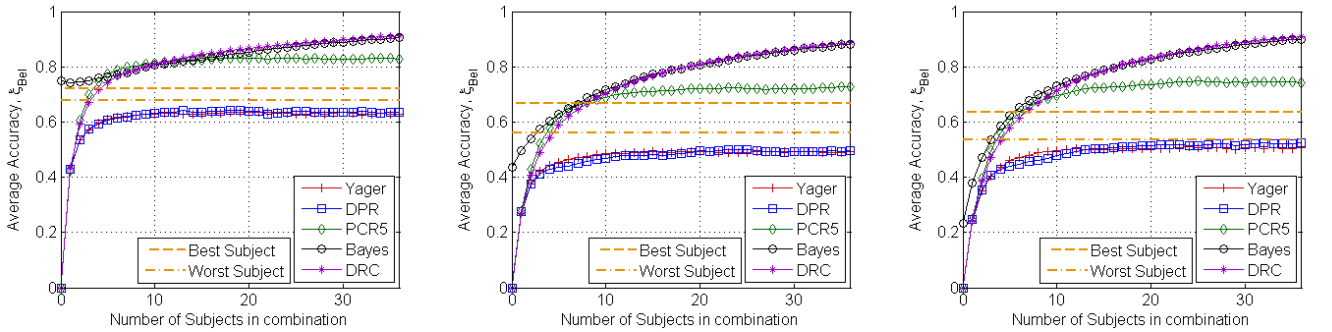
When no source discounting is performed, Bayes' rule of probability combination and DRC could not be used. The reason was that the chances of any two simulated subjects presenting totally conflicting evidence was non-negligible[1]. With this situation in mind, we make note of the following observations.

- When source discounting was performed using average source evidence strength, Bayes' rule of probability combination and DRC exhibited similar accuracy performance (Figures 3d-3f).
- The number of alternatives was observed to have a stronger impact on the accuracy performance when source discounting was not performed (Figure 3). Similar to the subject performance results (Figure 2), the largest decrease in accuracy performance occurred when going from 2 to 4 alternatives. The decrease was smaller when going from 4 to 8 alternatives.
- When source discounting was performed, similar performance was observed by PCR5, Bayes' rule of probability combination and DRC, as long as there were twelve or less human responses in the combination. When we included more than twelve human responses in the combination, Bayes' rule of probability combination and DRC exhibited higher accuracy performance than PCR5 (Figures 3d-3f).
- When source discounting was performed, PCR5 and DRC precision increased as the number of sources present in

[1]Totally conflicting evidence results in a division by zero in the equations for Bayes' rule of probability combination and for DRC.



(a) 2 lines, No evidence strength discounting (b) 4 lines, No evidence strength discounting (c) 8 lines, No evidence strength discounting



(d) 2 lines, Evidence strength discounting (e) 4 lines, Evidence strength discounting (f) 8 lines, Evidence strength discounting

Fig. 3. Average accuracy performance (i.e., ξ_{Bel}) for each of the fusion methods mentioned in Section IV versus the number of sources present in combination (higher is better). The evidence strengths for the best and worst subjects in the combination are also shown for comparison.

the combination increased. Eventually the precision converged to one for both PCR5 and DRC. This convergence was observed to occur more quickly with PCR5 than with DRC. Additionally, it was observed that increasing the number of alternatives decreased the rate of convergence for both PCR5 and DRC (Figures 4d-4f).

- When source discounting was not performed, the precision performance of PCR5 was found to be the same regardless of the number of line length task alternatives (Figures 4a-4c).
- Yager's rule and Dubois and Prade's rule exhibited inferior accuracy performance in all cases (Figure 3). Both rules exhibited accuracy performance that was worse than the worst single source present in the combination. Furthermore, Yager's rule and Dubois and Prade's rule also exhibited the lowest precision performance (Figure 4).

VI. Conclusions

We have shown how the 2DSD human simulator of [12] can be applied to determine the average performance (i.e., accuracy and precision) of fusion operators which use human opinions on M-ary decision problems. We make use of confidence assessment aggregation through successive pairwise comparisons. The same approach can be used with other human-decision models that provide decisions along with assessments of confidence in these decisions. Here, we used an M-ary line length task simulator as an example to evaluate the accuracy and precision of Bayes' rule of probability combination, Dempster's Rule of Combination (DRC), Yager's rule, Dubois and Prade's rule, and the Proportional Conflict Redistribution Rule #5 (Figures 3 and 4). It was observed that the accuracy of Bayes' rule of probability combination and DRC was minimally affected when incorporating subjective data through confidence assessments. After combination of ten to fifteen sources, Bayes' rule of probability combination and DRC were found to exhibit the highest accuracy performance when source discounting was performed. PCR5 was found to exhibit accuracy performance at least as good as the best source in the combination across all fusion cases. Yager's rule and Dubois and Prade's rule were found to exhibit inferior performance, as they exhibited worst accuracy and precision values.

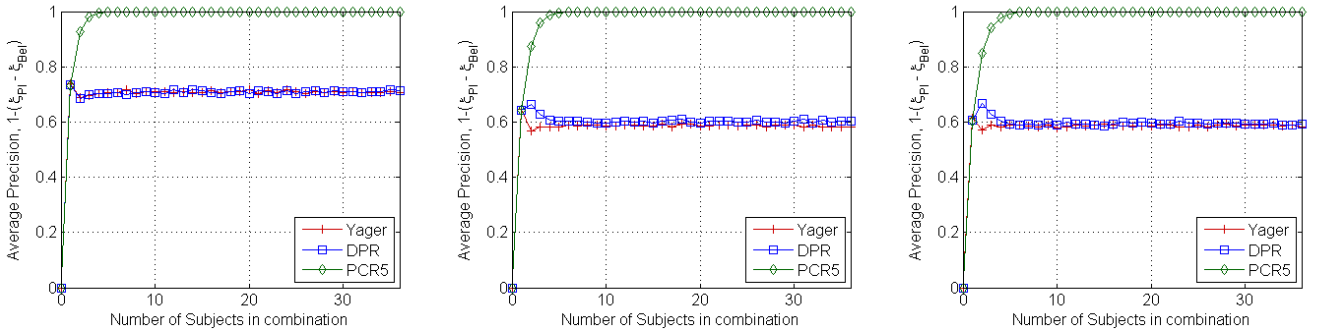
References

[1] D. L. Hall, M. McNeese, J. Llinas, and T. Mullen, "A framework for dynamic hard/soft fusion," in *Information Fusion*, pp. 1–8, 2008.

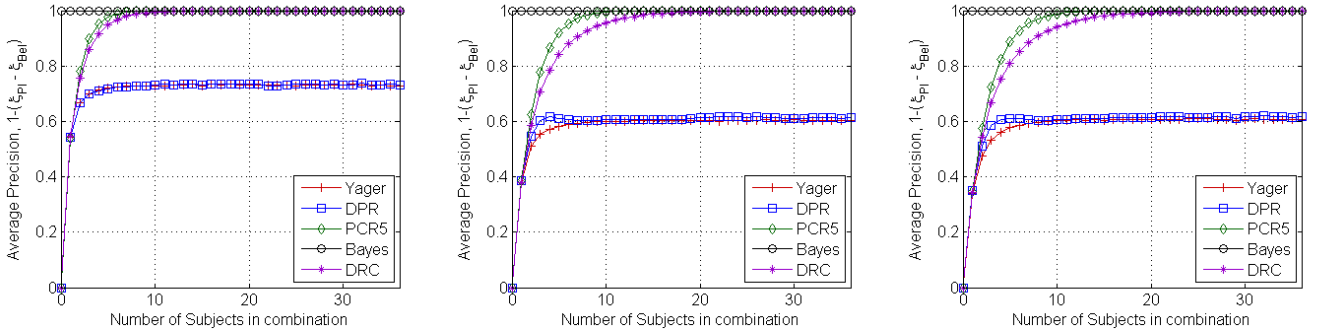
[2] L. A. Zadeh, "On the validity of dempster's rule of combination of evidence," Tech. Rep. 79/24, University of California, Berkely, 1979.

[3] J. Dezert, P. Wang, A. Tchamova, *et al.*, "On the validity of dempster-shafer theory," in *Proceedings of the 15th International Conference on Information Fusion*, 2012.

[4] R. T. Clemen and R. L. Winkler, "Aggregation of expert probability judgments," in *Advances in Decision Analysis: From Foundations to Applications* (W. Edwards, R. F. M. Jr, and D. von Winterfeldt, eds.), vol. 7, ch. 9, pp. 154–176, Cambridge University Press, 2007.



(a) 2 lines, No evidence strength discounting (b) 4 lines, No evidence strength discounting (c) 8 lines, No evidence strength discounting



(d) 2 lines, Evidence strength discounting (e) 4 lines, Evidence strength discounting (f) 8 lines, Evidence strength discounting

Fig. 4. Average precision performance (i.e., $1 - (\xi_{\text{Pl}} - \xi_{\text{Bel}})$) for each of the fusion methods mentioned in Section IV versus the number of sources present in combination (higher is better).

[5] S. Kaplan, "'Expert information' versus 'expert opinions'. Another approach to the problem of eliciting/ combining/using expert knowledge in {PRA}," *Reliability Engineering & System Safety*, vol. 35, no. 1, pp. 61 – 72, 1992.

[6] S. Acharya and M. Kam, "Evidence combination for hard and soft sensor data fusion," in *Proceedings of the 14th International Conference on Information Fusion*, pp. 1–8, July 2011.

[7] M. Cao, A. Stewart, and N. Leonard, "Convergence in human decision-making dynamics," *Systems & Control Letters*, vol. 59, no. 2, pp. 87–97, 2010.

[8] R. T. Clemen and R. L. Winkler, "Combining probability distributions from experts in risk analysis," *Risk analysis*, vol. 19, no. 2, pp. 187–203, 1999.

[9] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. Cohen, "The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks.," *Psychological review*, vol. 113, no. 4, pp. 700–765, 2006.

[10] R. Bogacz and K. Gurney, "The basal ganglia and cortex implement optimal decision making between alternative actions," *Neural computation*, vol. 19, no. 2, pp. 442–477, 2007.

[11] J. Ditterich, "A comparison between mechanisms of multi-alternative perceptual decision making: ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory," *Frontiers in Neuroscience*, vol. 4, 2010.

[12] T. Pleskac and J. Busemeyer, "Two-stage dynamic signal detection: A theory of choice, decision time, and confidence," *Psychological review*, vol. 117, pp. 864–901, July 2010.

[13] D. J. Bucci, S. Acharya, and M. Kam, "Simulating human decision making for testing soft and hard/soft fusion algorithms," in *Proceedings of the 47th Annual Conference on Information Sciences and Systems (CISS)*, 2013.

[14] D. J. Bucci, S. Acharya, T. J. Pleskac, and M. Kam, "Subjective confidence and source reliability in soft data fusion," in *Proceedings of the 48th Annual Conference on Information Sciences and Systems (CISS)*, March 2014.

[15] A. Heathcote, S. Brown, and D. Mewhort, "Quantile maximum likelihood estimation of response time distributions," *Psychonomic Bulletin & Review*, vol. 9, no. 2, pp. 394–401, 2002.

[16] G. Brier, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, pp. 1–3, 1950.

[17] M. Daniel, "On probabilistic transformations of belief functions," Tech. Rep. 934, Institute of Computer Science: Academy of Sciences of the Czech Republic, Sep 2005.

[18] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press: Princeton, NJ, 1976.

[19] R. Yager, "On the dempster-shafer framework and new combination rules," *Information Sciences*, vol. 41, pp. 93–138, 1987.

[20] D. Dubois and H. Prade, "A set-theoretic view of belief functions: Logical operators and approximations by fuzzy sets," in *Classical Works of the Dempster-Shafer Theory of Belief Functions* (L. L. Ronald Yager, ed.), vol. 219 of *Studies in Fuzziness and Soft Computing*, pp. 375–410, Springer-Verlag, Berlin, 2008.

[21] F. Smarandache and J. Dezert, *Advances and Applications of DSmT for Information Fusion*, vol. III. American Research Press, 2009.

[22] P. Smets, "Analyzing the combination of conflicting belief functions," *Information Fusion*, vol. 8, no. 4, pp. 387–412, 2007.

[23] J. Pearl, "Reasoning with belief functions: An analysis of compatibility," *International Journal of Approximate Reasoning*, vol. 4, no. 5-6, pp. 363–389, 1990.