Assessing the strength of cognitive states using confidence ratings and response times

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Memory &amp; Cognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>Draft</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Original Manuscript</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>n/a</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Weidemann, Christoph; Swansea University, Department of Psychology Kahana, Michael; University of Pennsylvania, Psychology;</td>
</tr>
<tr>
<td>Keywords:</td>
<td>memory, word recognition, reaction time analyses/methods, recognition, speed-accuracy trade-off</td>
</tr>
</tbody>
</table>
Assessing the strength of cognitive states using confidence ratings and response times

Christoph T. Weidemann
Swansea University

Michael J. Kahana
University of Pennsylvania

Christoph T. Weidemann
Swansea University, Department of Psychology
Singleton Park
Swansea SA2 8PP, Wales, UK e-mail: ctw@cogsci.info
phone: +44-(0)1792-606766
Version of December 30, 2014
Classification of stimuli into categories (such as “old” and “new” in tests of recognition memory or “present” vs. “absent” in signal detection tasks) requires the mapping of continuous signals to discrete responses. Introspective judgments about a given choice response are regularly employed in research, legal, and clinical settings in an effort to measure the signal that is thought to be the basis of the classification decision.

Correlations between introspective judgments and task performance suggest that such ratings often do convey information about internal states that are relevant for a given task, but well-known limitations of introspection call the fidelity of this information into question. We investigated to what extent response times can reveal information usually assessed with explicit confidence ratings. We quantitatively compared response times to confidence ratings in their ability to qualify recognition memory decisions and found convergent results suggesting that much of the information from confidence ratings can be obtained from response times.
Assessing the strength of cognitive states using confidence ratings and response times

Theories of perception, memory, decision making, and other cognitive tasks usually assume that underlying processes operate on continuous signals that are somehow thresholded to produce discrete responses. This assumption is particularly compelling because discrete choices are regularly associated with subjective feelings of strength or confidence. Indeed, introspective judgments of a feeling’s intensity or of an individual’s confidence in a given choice response are regularly employed in research, legal, and clinical settings in an effort to measure the signal that is thought to be the basis of this experience (Busey, Tunnicliff, Loftus, & Loftus, 2000; Jensen & Karoly, 2011). Correlations between introspective judgments and task performance suggest that such ratings often do convey information about internal states that are relevant for a given task (Fleming, Weil, Nagy, Dolan, & Rees, 2010; Katz, 1970; Baranski & Petrusic, 1998; Ackerman & Koriat, 2011), but well-known limitations on introspection call the fidelity of this information into question (Baumgartner & Steenkamp, 2001; Harzing, Brown, Köster, & Zhao, 2012; Lee, Soutar, & Louviere, 2008; Mueller & Weidemann, 2008; Paulhus, 1991).

Research on the nature of confidence judgments has shown that they largely reflect “fluency” of the response (i.e., the ease and speed with which it is generated), even when it is a poor index of performance (Benjamin, Bjork, & Schwartz, 1998; Dougherty, Scheck, Nelson, & Narens, 2005; Kelley & Lindsay, 1993; Schwartz, Benjamin, & Bjork, 1997; Oppenheimer, 2008). It is often important to capture trial-by-trial measures of internal states related to task performance and introspective ratings (usually of response confidence) are almost exclusively used for this purpose. In light of the well-known relationship between such ratings and response fluency (often operationalized by response latency) and the substantial effort required to solicit these ratings (which often take longer
to execute than the response to which they apply), it is remarkable that attempts to quantify the strength of cognitive states on the basis of response latency are rare.

Here we present a general approach to assess to what extent a given measure reflects information underlying a classification response. We use this approach to directly contrast introspective judgments and associated response times with respect to their ability to qualify recognition memory decisions. To foreshadow the results, we show that despite absolute differences in performance indices derived from confidence ratings and response latencies, the relative pattern of these indices across various partitions of the data is remarkably similar.

Measuring classification performance

Signal Detection Theory (SDT) is a common analysis framework for tasks with two response classes (D. M. Green & Swets, 1966). Within this framework, the strength of the (internal) signal on which the classification is based is assumed to vary continuously and a criterion is placed to map the continuous signal onto a binary classification response. Introspective judgments that qualify the classification response are assumed to be based on additional criteria on this same signal and responses at each rated level can be classified as either “hit” (correct response in favor of a standard response class, e.g., “old” in recognition memory tasks or “present” in signal detection tasks) or “false alarm” (FA; incorrect response in favor of a standard response class). With rising endorsement of the standard response class, the way the cumulative hit rates increase relative to the cumulative FA rates indexes a classifier’s performance (a receiver operating characteristic [ROC] function; Figure 1). To the extent that a classifier correctly distinguishes between the two stimulus classes, hit rates should initially rise faster leading to a concave ROC function.

Studies of recognition memory often place a special focus on the shape of the ROC
functions in an effort to distinguish between competing theories of the underlying processes. Rather than investigating the processes giving rise to any particular shape of ROC functions, we explore to what extent ROC functions that are based on response time can provide information that is similar to that obtained from confidence ratings. Specifically, we remain agnostic with respect to the assumptions of SDT or any particular model for task performance and simply use the area under ROC curves (AUC) as a measure of ordinal separation of two the distributions indexing the responses to targets and lures respectively. One advantage of this approach is that it generalizes to any binary classification task and thus is directly applicable across a wide range of domains for psychological inquiry.

In contemporary research on perception and memory, ROC functions are almost exclusively constructed through putative manipulation of response criteria (e.g., by varying instructions, stimulus base-rates, or payoff contingencies) or from introspective judgments. This is somewhat surprising given that historically a wide variety of dependent variables, including response time (Carterette, Friedman, & Cosmides, 1965; Katz, 1970; Norman & Wickelgren, 1969; Emmerich, Gray, Watson, & Tanis, 1972; Yager & Duncan, 1971; Moss, Myers, & Filmore, 1970; Kulics, Carlson, & Werner, 1974; Kulics & Lineberry, 1977; Blough, 1978; M. Green, Terman, & Terman, 1979; Weintraub & Fidell, 1979; Essick & Whitsel, 1985; Herskovic, Kietzman, & Sutton, 1986; K. N. O’Connor, Roitblat, & Bever, 1983; K. O’Connor & Ison, 1991; Gray, 1993; Yin, Fritz, & Shamma, 2010; Thomas & Myers, 1972; see also Luce, 1986), latency of heart rate increase (Kulics et al., 1974), response frequency (Blough, 1967), and firing rates of individual neurons (Essick & Whitsel, 1985; Zhang, Riehle, & Requin, 1997), have been used for ROC construction. To construct an ROC function from a dependent variable, one has to assume or establish a mapping between this measure and the evidence for the classification response. A common way to construct ROC functions is to partition the dependent variable (e.g., response
Assessing the strength of cognitive states

time) by the classification response and to sort it within these partitions on the basis of
this mapping. For the case of response time, the usual assumption is that fast responses
are based on a stronger signal / made with higher confidence than slow responses, such
that, in the case of recognition memory, the inferred evidence that an item has been
studied is weakest for fast “new” responses, strongest for fast “old” responses and
intermediate for slow responses (Emmerich et al., 1972; Jacobs, Hwang, Curran, &
Kahana, 2006; Luce, 1986; Navon, 1975). We refer to response times (RTs) ordered in this
way as “RT strength”. This assumed relationship between signal strength and response
time is well supported by findings that responses that are made with high confidence also
tend to be made faster than those with low confidence (Ratcliff & Starns, 2009; Heathcote,
2003; we address occasional exceptions to this relationship in the Discussion section).1

Classification performance and signal strength

ROC functions constructed in this way are constrained to pass through the
classification point (the point separating, say, “old” from “new” responses in tests of
recognition memory or “signal present” from “signal absent” responses in signal detection
tasks) as well as through the points where both hits and FAs are either zero or one.
Consider, for example, a random measure that bears no relationship to the classification
response such as the throw of a die. Just as with confidence ratings, one could interpret
the outcomes of a die thrown after every classification response in a test of recognition
memory such that the inferred memory strength of an item is weakest for a “new”
response with an outcome of six, strongest for an “old” response with an outcome of six
and intermediate for lower casts. In the limit, the partitions of this variable within a given
classification response contain equal proportions of correct and incorrect responses such that the rate of increase of hits and FAs in the ROC function is constant within each classification response. Thus a “random ROC” that reveals no information beyond that contained in the classification response is bilinear connecting the origin to the classification point and that point to (1, 1) with straight lines (shown together with the main diagonal as dotted lines in Figure 1; Emmerich et al., 1972; Mueller & Weidemann, 2008). The AUC therefore conflates classification performance with the measure’s ability to reflect the signal underlying the classification decision. A relative index of how much information a particular measure contains about the signal underlying the classification response can be obtained by subtracting the area under the random ROC from that under the ROC of interest (Emmerich et al., 1972). Previous applications of this method have shown that response times and other measures contain significant information that qualifies a classification response at levels that sometimes approached but never exceeded that contained in confidence ratings (Emmerich et al., 1972; Herskovic et al., 1986; Thomas & Myers, 1972).

In order to assess to what extent response times can reveal information similar to that obtained by confidence ratings, we administered a recognition memory test that first asked for a binary recognition decision followed by a confidence rating. This set-up allowed us to directly compare the time taken for the recognition decision with the subsequent introspective judgments.

Methods

Participants

We obtained data from a larger memory study that asked participants to return for a total of 20 sessions each. From all participants we selected young adults (maximum age: 35 years) who provided data from at least 7 sessions. We excluded trials with invalid
confidence responses and those with response times for binary old-new judgments below 300 ms or above 3000 ms (a total of 3% of the full data set). From the remaining data, we eliminated 121 sessions (about 3% of the data) that did not contain at least one “old” and one “new” response for both targets and lures. Some analyses partitioned the targets into those that were previously recalled and those that were not previously recalled (see below for details). For those analyses we additionally required that sessions contained at least one “old” and one “new” response for both types of targets which excluded a further 540 sessions (16% of the remaining data set). These exclusion criteria yielded a total of 171 participants (of which only 10 provided data from fewer than 10 sessions for the general analyses with 24 participants providing data from fewer than 10 sessions for the analyses partitioning the data by previous recall). The total number of analyzed sessions was 3120 for the general analyses and 2580 for the analyses partitioning the data by previous recall.

Experimental task

Each session included multiple pairs of study lists followed by recall tasks. Details of study conditions and recall tasks varied across sessions, but in all cases participants studied words presented on a computer screen before being probed to recall the previously presented words. The current study focuses mostly on a final recognition test at the end of each session: Participants were shown one word at a time and asked to indicate whether each word was presented in any of the previous study lists that had been shown in this session. Participants answered verbally by speaking either “pess” or “po” into a microphone to answer in the affirmative or negative respectively (“yes” and “no” were replaced by “pess” and “po” to equalize the initial phoneme in an effort to allow more precise measurements of response latencies; response time was only measured in relation to the initial recognition memory decision and not with respect to the confidence rating). Following the “old”/“new” classification, participants were asked to rate their confidence
in their classification response on a 5-point scale from “very unsure” to “very sure” by either speaking a number between 1 and 5 into the microphone (most participants) or by pressing the corresponding number key on the keyboard. The proportion of lures in the recognition memory test was varied across sessions, but this manipulation was not a focus of the current investigation. Participants indicated that they had finished speaking by pressing the space key on a computer keyboard both after the classification response and after the confidence rating (response times were only measured as the latency of the verbal classification response). Immediately after participants indicated that they had finished the confidence rating they received brief (randomly jittered between 100 and 200 ms) visual (“Correct” in green or “Incorrect” in red) and auditory feedback on their classification decision (feedback was automatically generated with custom speech recognition software). After offset of the feedback, the screen turned blank for a variable interval uniformly distributed between 800 ms and 1200 ms before the next test word was presented. All stimulus presentation and recording of voice and button-press responses was done with PyEPL (Geller, Schleifer, Sederberg, Jacobs, & Kahana, 2007). Some analyses condition recognition memory performance on whether or not a given target item was recalled in any of the recall periods of that session, but we make no other reference to performance in the recall periods of this experiment. Electroencephalography recordings were obtained but are not further discussed in this article.

Results

Assessing memory strength with C- and L-ROC functions

Figure 1 illustrates the construction of the ROC functions based on confidence ratings (C-ROC; top row) and response latency (L-ROC; bottom row). The figure breaks down the steps of constructing and assessing ROC functions to illustrate how this process generalizes across dependent variables. The common case of using confidence ratings for
ROC construction is illustrated in the top row and the bottom row mirrors these steps for response latencies. We followed this procedure separately for each experimental session, averaged the sessions for each participant and show the mean across participants in Figure 1. It is important to note that as long as some relationship between a dependent variable and the strength of the signal underlying the classification response can be assumed or established, the same procedure can be used to evaluate to what extent this variable is able to qualify the classification decision. Similarly, this procedure is completely agnostic with respect to the nature of the classification decision, and, indeed, previous applications of this procedure have almost exclusively focused on signal detection / perceptual discrimination tasks (Emmerich et al., 1972). A particular feature of confidence ratings is that they are usually discrete whereas many other variables that could be used to qualify the signal underlying classification decisions (such as response times and physiological recordings) are continuous. To better illustrate the correspondence with confidence ratings, Figure 1 shows RT strength binned into the same number of bins as there are in our confidence scale (using equally spaced quantiles). All analyses, however, are based on the raw latency data which is why the L-ROC function does not connect the indicated points with straight lines (the curvature reflecting the use of the full resolution of the latency data is difficult to discern, but most prominent for the lowest strength “old” responses).

Though uncommon (and impractical for continuous variables), the response probability plots in the left panels of Figure 1 contain the same information as the corresponding ROC functions (albeit at a lower resolution for our latency data for illustrative purposes as explained above): Cumulatively adding the response probabilities
Assessing the strength of cognitive states

(starting with the strongest “old” responses) for targets and lures to the hit and FA probabilities respectively traces out the ROC function. The response probability plots indicate that for both confidence ratings and RT strength, stronger responses tended to be associated with higher response probabilities for correct responses and lower response probabilities for incorrect responses — a trend that is more easily quantified on the basis of the resulting ROC functions: As is evident from Figure 1, the L-ROC is closer to the random ROC than the C-ROC, resulting in a significant difference between the areas under the C-ROC and the L-ROC functions ($t[170] = 13.609, SE = 0.003, d = 1.041$, $p < .001$). As is also clear from the figure, the areas under both ROC functions exceeded those under the random ROC functions, effects which turned out to be substantially larger than the difference between the two areas ($t[170] = 109.312, SE = 0.004, d = 8.359$, $p < .001$ and $t[170] = 65.243, SE = 0.006, d = 4.989, p < .001$, for C-ROCs and L-ROCs respectively).

To investigate the correspondence between the AUC for C-ROC and L-ROC functions ($\text{AUC}_C$ and $\text{AUC}_L$ respectively), we analyzed the correlation of these and related measures. Figure 2A shows a scatter-plot of $\text{AUC}_C$ and $\text{AUC}_L$ that reveals a very close correspondence between both areas ($r = .93, t[169] = 44.714, p < .001$). This correlation, however, is inflated by the fact that both ROC functions are constrained to pass through the classification point.

Insert Figure 3 about here

Classification-response specific ROC functions

Another way to compare confidence ratings and response times as measures of memory strength, without contamination from classification performance, is by assessing
Assessing the strength of cognitive states
	heir ability to qualify classification responses separately for “old” and “new” judgments. Figure 3 illustrates how the data for “old” and “new” judgments can be separately used as the bases for the calculation of classification-alternative specific ROC functions. This figure shows the same data as Figure 1, but this time response probabilities are conditioned on the classification response. As in Figure 1, RT strength is binned in the response probability panels of Figure 3 and the points corresponding to these bins are indicated on the respective ROC functions. This binning again serves to illustrate the correspondence between the approaches for confidence ratings and response latencies and we used the full resolution of the response time data in the generation of ROC functions and for corresponding analyses (the curvature of the lines connecting points on the L-ROC functions which reflects our use of the raw latency data is clearly discernible in Figure 3).

Whereas the response probabilities for targets and lures across both classification responses each add up to 1.0 in Figure 1, they add up to 1.0 within each classification response in Figure 3. The distribution of target and lure response probabilities across the response categories in Figure 1 reflects the overall classification performance and by conditioning on the classification response in Figure 3, classification performance does not affect the resulting ROC functions. We present this approach as a novel way to analyze measures qualifying classification responses that allows for the separate assessment of such measures for each response class. We comment further on this approach in the Discussion section.

Given that Figure 3 is based on the same data as Figure 1, it is not surprising that it, too, indicates that stronger responses are associated with higher probabilities for correct responses and lower probabilities for incorrect responses — a trend that is more easily discernible in Figure 3 for both response alternatives due to the conditioning. These response probabilities contain the same information as the corresponding ROC functions (albeit at a lower resolution for L-ROCs due to the binning as explained above). They can
be used to generate separate ROC functions for both response alternatives (“old” and “new” in this case), by cumulatively adding response probabilities for targets and lures with decreasing strengths to the hit and FA probabilities respectively for the standard response class (“old” in this case) and vice versa for the other response alternative.

The areas under these ROC functions indicate to what extent the respective dependent variable (confidence ratings or RT in our case) qualifies a given classification response alternative (“old” or “new” in our case). Areas under all ROC functions in Figure 3 clearly exceeded .5 ($t[170] = 18.933–59.921$, $SE = 0.005–0.009$, $d = 1.448–4.582$, all $ps < .001$) confirming that both measures contain additional information about each classification response alternative. A large difference between the areas under the ROC functions for “old” and “new” responses is also apparent ($t[170] = 36.076$, $SE = 0.005$, $d = 2.759$ and $t[170] = 11.337$, $SE = 0.006$, $d = 0.867$, for confidence ratings and response latencies respectively; both $ps < .001$). The differences between the areas under the corresponding C- and L-ROC functions are also substantial ($t[170] = 23.038$, $SE = 0.008$, $d = 1.762$, and $t[170] = 17.781$, $SE = 0.004$, $d = 1.360$ for “old” and “new” responses respectively; both $ps < .001$). Panels B and C of Figure 2 show the relationship between $AUC_C$ and $AUC_L$ for “old” and “new” responses respectively. These scatter-plots illustrate large correlations between AUCs derived from confidence ratings and RTs for both “old” ($r = .45$, $t[169] = 7.860$, $p < .001$) and “new” ($r = .56$, $t[169] = 11.053$, $p < .001$) responses that are not affected by the constraint of the full ROC functions to pass through the classification point.

As illustrated in Figure 3, both confidence ratings and response latencies are much better able to qualify “old” responses than “new” responses. Nevertheless, it seems
reasonable to ask to what extent these measures are correlated across classification responses (e.g., do high areas under the C-ROC [L-ROC] for “old” responses coincide with high areas under the C-ROC [L-ROC] for “new” responses?). Figure 4 illustrates substantial and similar relationships between AUCs for “old” and “new” responses for confidence ratings ($r = .57, t[169] = 11.174, p < .001$) and response latencies ($r = .68, t[169] = 15.548, p < .001$).

Previous recall as a proxy for memory strength

The set-up of the current experiment allowed us to compare performance for targets that have been recalled to that for items that have not been recalled in any of the recall periods that preceded the recognition memory test. Across all included sessions the mean proportion of previously recalled targets was 61% ($SD = 16\%$). Figure 5 shows the mean AUCs based on previously recalled vs. unrecalled targets for confidence ratings, response latencies and the corresponding AUCs for ROC functions that are conditioned on the classification response (cf. Figure 3). It is clear from the figure that recognition memory was lower for targets that were not previously recalled (as indicated by the shorter shaded areas outlining the areas for the random ROCs that are only based on classification performance; $t[170] = 36.265, SE = 0.002, d = 2.773, p < .001$). For both sets of items, confidence ratings and response latencies effectively qualified classification decisions as indicated by AUCs for classification response specific ROC functions exceeding 0.5 ($t[170] = 13.263–63.322, SE = 0.004 – 0.009, d = 1.014–4.842, ps < .001$). The difference between the AUCs corresponding to the C-ROC as well as the L-ROC and the random ROC appears to be similar for AUCs based on recalled vs. unrecalled trials (0.063 vs.
0.060 and 0.023 vs. 0.022 for AUCs based on C-ROCs and L-ROCs respectively). However, this similarity is deceiving, because these differences have different baselines anchored on the random AUCs for recalled (0.897) and unrecalled (0.829) targets. The AUCs based on the ROC functions that are conditioned on the classification response (the four right-most sets of bars in Figure 5) reveal that both confidence ratings and response latencies were somewhat more effective at qualifying recognition decisions for previously recalled items, especially when these were classified as “old”. A 2 (recalled status) × 2 (response) × 2 (dependent measure) repeated measures ANOVA on the AUCs that are conditioned on the classification response (i.e. the four rightmost sets of bars in Figure 5), confirmed large main effects ($F[1, 170] = 332–645$, $MSE = 10.17–91.08$, $\eta^2_p = .66–.79$, $p < .001$) as well as large interactions between recalled status and response ($F[1, 170] = 130$, $MSE = 3.39$, $\eta^2_p = .43$, $p < .001$) and between response and dependent measure ($F[1, 170] = 310$, $MSE = 13.85$, $\eta^2_p = .65$, $p < .001$).

Despite this discrepancy between the effectiveness of C- and L-ROC functions for previously recalled and unrecalled targets, we would expect that in sessions where a given measure is effective at qualifying classification decisions involving previously recalled targets, it should also be effective at qualifying classification decisions involving previously unrecalled targets (and likewise for sessions where a given measure is less effective).

Figure 6 shows the correlations between the AUCs based on previously recalled and unrecalled targets for C- and L-ROC functions. Both correlations are substantial and similar ($r = .89$, $t[169] = 34.076$, and $r = .94$, $t[169] = 50.348$ respectively, both $p < .001$). The corresponding correlations for AUCs based on ROC functions that are conditioned on the classification response are shown in Figure 7 and also show substantial and similar
Assessing the strength of cognitive states

correlations ($r = .89$, $t[169] = 35.412$, and $r = .97$, $t[169] = 72.265$ for “old” responses as well as $r = .57$, $t[169] = 11.158$ and $r = .62$, $t[169] = 12.989$ for “new” responses, all $p < .001$).

Discussion

Much of what we know about human behavior comes from experimental work asking participants to judge a stimulus as belonging to one of two classes. Examples include tasks asking participants to distinguish studied from unstudied items (e.g., to investigate recognition memory — our focus in this paper), tasks asking participants whether two stimuli are identical or not (e.g., to establish a detection threshold), tasks asking which of two stimuli is larger/better on some dimension (e.g., to study inference and/or preference), and tasks asking participants to match stimuli to categorical labels (e.g., to study knowledge and/or perception). Most theories of cognitive processes that are measured with these tasks, assume that a continuous signal (e.g., strength of a memory, percept, or preference) is somehow thresholded to produce a binary response (e.g., “old” vs. “new” in a recognition memory task or “signal present” vs. “signal absent” in a signal detection task). Attempts to characterize this signal, however, have mainly focused on aggregating large numbers of trials (e.g., to compare average levels of evidence across different experimental conditions) or on introspective judgments (e.g., by asking participants to rate their confidence in each classification decision). The latter measures have the distinct advantage of providing trial-by-trial assessments of the evidence underlying the classification decisions, but previous research suggests that they reflect an inference about this evidence (based on response fluency) rather than direct introspective
access to it (Kelley & Lindsay, 1993; Benjamin et al., 1998; Dougherty et al., 2005; Kroriat, 1993; Kroriat, Lichtenstein, & Fischhoff, 1980; Oppenheimer, 2008; Schwartz et al., 1997). Furthermore, a large literature on response biases in survey data has demonstrated reliable limits on introspective judgments (Baumgartner & Steenkamp, 2001; Harzing et al., 2012; Lee et al., 2008; Paulhus, 1991). Potential issues with the use of confidence ratings have also been highlighted in recent debates on the extent to which recognition memory ROC functions are compatible with a dual high threshold model positing discrete “recognize as old,” “recognize as new,” and “uncertain” states, rather than continuously varying evidence assumed by SDT (Bröder & Schütz, 2009; Dube & Rotello, 2012; Dube, Starns, Rotello, & Ratcliff, 2012; Rouder, Province, Swagman, & Thiele, 2014).

Given that introspective judgments are costly to obtain, largely reflect response fluency (which is usually assumed to be inversely related to response latency), and may be subject to response biases, we assessed to what extent response latencies conveyed similar information to that obtained from introspective judgments. Even though response times have previously been used in several studies to generate ROC functions in perceptual tasks (Carterette et al., 1965; Katz, 1970; Emmerich et al., 1972; Yager & Duncan, 1971; Moss et al., 1970; Kulics et al., 1974; Kulics & Lineberry, 1977; Blough, 1978; M. Green et al., 1979; Weintraub & Fidell, 1979; Essick & Whitsel, 1985; Herskovic et al., 1986; K. N. O’Connor et al., 1983; K. O’Connor & Ison, 1991; Gray, 1993; Yin et al., 2010), as far as we are aware, the only published application to recognition memory is a study that investigated performance in a range of memory tasks for four participants (Norman & Wickelgren, 1969). This is particularly surprising given that many studies of recognition memory place a strong focus on the shape of ROC functions in an effort to constrain theoretical accounts and given that response times to recognition decisions feature prominently in attempts to understand recognition memory using sequential sampling models; see Starns and Ratcliff (2014) and Starns (2014) for recent efforts to validate the
Assessing the strength of cognitive states

unequal-variance assumption in recognition memory (which is usually supported by analyses of the shape of ROC functions) with response times in a diffusion model analysis.

We found that that L-ROC functions tended to be closer to the corresponding random ROC functions than C-ROC functions — a common finding among studies using L-ROC functions — but that areas under both types of ROC functions conveyed similar information about relative performance. In all cases we are aware of, the generation of L-ROC functions depends on the assumption that RT tends to be inversely related to evidence strength. See Thomas and Myers (1972) and Navon (1975) for theoretical work assuming that response time should be related to the signal underlying the classification decision, with the former authors providing a formal derivation of (among other things) how variability in the dependent measure (“criterion variability”) affects the shape of the resulting ROC function. Even though there is strong evidence for this assumed relationship between response time and evidence strength, there is also evidence for different relationships in some cases (Ratcliff & Starns, 2009). Examination of L-ROC AUCs for individual sessions (not shown here) revealed that in some cases these AUCs fell considerably below 0.5 (i.e., the level indicating no relationship between RTs and memory strength). This suggests that for these cases our assumptions for the calculation of RT strength are not met. Analyses of some of these cases (not reported in detail here) revealed relatively high FA rates and a tendency for faster responses (particularly for incorrect responses). This is evidence for a substantial proportion of fast guesses in these data which would counter-act the assumed trend of shorter RTs being associated with stronger evidence. Presumably inferential processes that generate confidence ratings on the basis of response fluency can take into account when a response was guessed quickly which may help explain the absolute differences in the AUCs for ROC functions based on confidence ratings and response latencies. Note also that some session contained more targets than lures which may have led participants to preferentially guess “old”. Such a
response pattern could explain the relatively larger AUCs for “old” confidence responses shown in Figure 2B (Panel C in this figure shows no such advantage for “new” confidence responses).

It is likely that even the data without obvious violations of the assumptions for the calculation of RT strength are somewhat contaminated by fast guesses. These tend to reduce the area under L-ROC functions unless the calculation of RT strength is amended to identify these cases to either exclude them or to associate them with low levels of evidence. Modeling the relationship between RT and memory strength in a way that allows for variability between participants and sessions might provide a sensible alternative basis for computing L-ROC functions. Recent accounts of confidence ratings in recognition memory and other binary choice tasks have successfully captured accuracy and response time data with sequential sampling models (Ratcliff & Starns, 2009; Pleskac & Busemeyer, 2010). The RTCON model (Ratcliff & Starns, 2009) is designed to handle confidence ratings only without provisions for a separate binary choice and is thus, in its current form, not able to account for similarities (or lack thereof) between C- and L-ROC functions. The two-stage dynamic signal detection interrogation model of confidence (Pleskac & Busemeyer, 2010), on the other hand, does predict a correlation between confidence ratings and binary choice RTs such that more confident responses are associated with shorter binary choice RTs. The strength of this relationship depends on the exact parameterization of the model and it is conceivable that this model, or a close variant, could provide the basis for a better way to generate L-ROC functions taking into account individual differences.

Our generation of ROC functions that were specific to each classification response (Figure 3) provide a novel way to assess to what extent a given measure qualifies the evidence underlying each possible response. It is conceivable that some measures are differentially sensitive to different levels of evidence or that evidence associated with one
response category is inherently less variable (perhaps because of a threshold process rather than a smooth transition from lower to higher levels of evidence). Indeed for both confidence and latency data we found that the areas under the ROC functions for “new” responses were considerably smaller than the corresponding areas for “old” responses. This indicates that both confidence ratings and response times, are better able to qualify “old” than “new” responses, but these analyses are unable to determine to what extent this is due to differential sensitivity of these measures to evidence levels associated with the two response classes and/or properties of the underlying evidence distributions. In this context it is interesting to note that especially for items classified as “old”, both confidence ratings and response latencies were considerably better able to qualify the classification decision for recalled relative to unrecalled items. Consistent with the comparison between “old” and “new” responses discussed above, this result indicates that both measures are more effective when operating on higher levels of evidence — a finding that, again, could be due to properties of these measures and/or the evidence distributions. A common finding in applications of SDT to recognition memory data is that estimates of the variability of memory strengths for targets tend to be larger than those for lures. Our finding of larger AUCs for “old” responses compared with “new” responses is compatible with this finding. Indeed, the fact that we found this pattern for both confidence ratings and response times, supports this result without relying on the assumptions of a particular modeling framework such as SDT or a drift-diffusion model (Starns & Ratcliff, 2014; Starns, 2014).

Traditionally, researchers have placed a strong focus on the exact shape of the ROC function in an effort to draw inferences about the underlying cognitive processes. Such inferences depend on a number of strong assumptions about the cognitive processes underlying the classification decision as well as about the mapping between latent states (e.g., “memory strength”) and the variables used to measure them (e.g., confidence
ratings). We already alluded to alternatives to our assumptions regarding the mapping between memory strength and response latency — changes in these assumptions will produce corresponding changes in the shapes of the resulting ROC functions that are not related to properties of the cognitive processes underlying the classification decision. Despite the higher face validity for the assumptions regarding the mapping between memory strength and introspective ratings (e.g., a higher confidence rating for an “old” response corresponds to a stronger memory), it is important to remember that this mapping can also be distorted (Baumgartner & Steenkamp, 2001; Benjamin et al., 1998; Harzing et al., 2012; Lee et al., 2008; Paulhus, 1991), which likewise limits the diagnosticity of the corresponding ROC function with respect to the cognitive processes underlying the classification decision. Furthermore, different dependent measures are likely to respond differently to some experimental manipulations — speed stress, for example, is likely to affect response latencies more than confidence ratings whereas instructions to distribute responses equally among the response alternatives may produce a more uniform distribution of confidence ratings at the expense of a reduced correspondence between those ratings and the latent state they are meant to assess. Because the details of the mapping between a given latent state and the dependent variable used to assess it crucially affect the shape of the corresponding ROC function, attempts to draw conclusions about the cognitive processes underlying the classification decision from the shape of the ROC function need to be justified for the particular context from which the ROC function was derived. Indeed, even the assumptions that are not specific to the particular dependent variable used to generate the ROC function have been criticized as “difficult-to-justify and untestable” (Rouder et al., 2014) and some of these assumptions have been shown to not hold up to scrutiny (Balakrishnan, 1999; Mueller & Weidemann, 2008; Ratcliff & Starns, 2009; Van Zandt, 2000). Our approach mostly sidesteps these issues by focusing on the area under the ROC functions to qualify the classification
decision. As described above, violations of the assumptions regarding the mapping between the latent state and the dependent variable used to measure it, will affect the corresponding ROC functions (and thus the respective AUCs), but unless the extent of this violation co-varies with other variables under consideration, this should only affect the absolute value of the AUCs and not the relative pattern of AUCs across those variables.

A common purpose of SDT analyses is to assess to what extent a given experimental manipulations affects the ability to discriminate between the stimulus classes ("sensitivity") and/or the preference for a given response option ("response bias"). The assessment of response biases requires a model of the decision process, but the current method allows the direct assessment of discrimination performance without the need to subscribe to the specific assumptions of SDT. In this context it is important to note that differences in average response latency across the two response classes (perhaps due to a response bias favoring one response class over the other), do not affect the resulting L-ROC functions. This is due to the sorting of the dependent variable for ROC construction being done separately for each response class such that a latency corresponding to strong evidence when it refers to one response option may correspond to weak evidence when it refers to the other response option (because for this response option, there are more faster response latencies). Similarly in cases where different ranges of nominal confidence levels are used for different response classes, these would usually be interpreted such that the largest values within each response class correspond to the most extreme evidence states, without consideration for consistency of the nominal confidence levels across response options.

Several of our analyses correlated different subsets of the data (partitions based on "old" vs. "new" responses and/or recalled vs. unrecalled targets) for measures based on both confidence ratings and response latency (cf. Figures 4, 6, and 7). All of these correlations were strikingly similar for measures based on confidence ratings and response
Assessing the strength of cognitive states 23

latency, in both cases showing strong correlations between AUCs for recalled and unrecalled items (Figure 6), especially for items classified as “old” (Figure 7), as well as between AUCs for “old” and “new” responses (Figure 4). This suggests that the internal structure of the data based on both dependent measures was quite similar, despite the absolute differences in the resulting AUCs — a result that supports our earlier assertion that relative differences in performance are captured similarly well by measures based on confidence ratings and response latencies.

Conclusions

The method we used to generate ROC functions can be (and has been) applied to other classification tasks and to other dependent variables besides confidence ratings and response times. It is likely that measures not based on confidence ratings or response latencies can provide similar or better insights into the evidence underlying classification responses and the current framework allows for the quantitative assessment of such (potential) indices of cognitive states (relative to a random baseline and/or to another index). ROC functions undeniably play an important role in the assessment of choice behavior — a role that generalizes well beyond tests of theories of recognition memory. In some cases, collection of confidence ratings can be impractical and the collection of binary choices across multiple conditions is not always a viable alternative. A strength of the current approach is that it generalizes over a wide range of potential tasks and dependent measures to assess the evidence underlying classification responses. We have shown that L-ROC functions derived from RTs to binary choice responses in a recognition memory task provide similar information to that obtained from confidence ratings. In this process, we have presented a novel method to separately assess the information revealed by a dependent measure with respect to each response class. Importantly, this way of evaluating evidence for classification decisions, either across all responses or separately for
Assessing the strength of cognitive states

each response class, can provide detailed insights into how experimental manipulations affect cognitive processing without relying on assumptions regarding the distribution of evidence or the nature of the decision process. The ability to use incidental measures (such as RT) for this purpose provides a convenient alternative to traditional ways of computing ROC functions that is not subject to the limits of introspection and that can even be applied to experiments “after the fact” as long as RT data (or data from another dependent variable that is thought to reflect the cognitive state on which a classification decision is based) is available.
Assessing the strength of cognitive states

References


Assessing the strength of cognitive states 26


Green, M., Terman, M., & Terman, J. S. (1979). Comparison of yes-no and latency


Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.


Assessing the strength of cognitive states 29


Assessing the strength of cognitive states

Psychophysics, 26, 143–152.


Author Note

Christoph T. Weidemann, Department of Psychology, Swansea University, Wales, UK; Michael J. Kahana, Department of Psychology, University of Pennsylvania. This work was supported by a BIAL Foundation grant to CTW and NIH grant MH55687 to MJK. We would like to thank the members of the Computational Memory Laboratory at the University of Pennsylvania for their assistance with data collection and pre-processing.

Correspondence should be sent to Christoph T. Weidemann, Swansea University, Department of Psychology, Singleton Park, Swansea, SA2 8PP, Wales, UK or by e-mail to ctw@cogsci.info.
Assessing the strength of cognitive states

Footnotes

1 In certain situations the relationship between response times and signal strength is better described by a speed-accuracy trade-off and conditional accuracy functions are sometimes used to characterize this relationship, especially in cases where response speed is not manipulated. Chapter 6 of Luce (1986) discusses these cases and analyses.

2 The interaction between recalled status and dependent measure also reached conventional levels of statistical significance, but that effect was of negligible size: \( \eta_p^2 = .06 \).

3 The exact relationship between the relative variability of memory strengths for targets and the AUCs for ROC functions corresponding to “old” vs. “new” responses depends on a number of assumptions, including the placement of the classification criterion, but simulations confirmed that within the framework of a simple SDT model, larger target variability coincided with larger AUCs for “old” responses.
Figure Captions

Figure 1. Response probabilities for bins of each measure (left) used to compute the ROC functions (middle) with corresponding areas (right). Rows show analyses for confidence ratings and response times respectively. For illustrative purposes, RT strength is shown partitioned into the same number of levels as there are confidence ratings (using equally spaced quantiles) and the points on the ROC functions corresponding to these bins are indicated. However, smooth ROC functions taking advantage of the full resolution of the data are drawn and form the basis of the area calculations. The left-most points on the ROC functions correspond to the rightmost bars in the left panels and subsequent points are calculated by cumulatively adding the probabilities for targets and lures to the hit and FA values respectively. Probabilities for target and lures are shown with overlapping bar graphs with hatching as indicated in the legend (additional shading, blue for targets and yellow for lures, is added to help with the discrimination). The classification point (i.e., the point separating “old” from “new” responses) is shown as a diamond (solid-red and dashed-green parts of the ROC functions indicate the parts corresponding to “old” and “new” responses respectively). Main diagonals as well as random ROC functions are shown as dotted lines in the ROC plots. The lowest value on the ordinate for the bar graphs on the right (0.87) corresponds to the area under the random ROC. Error bars on the area measure show the 95% confidence intervals.

Figure 2. Scatter-plots comparing areas under the ROC curve (AUC) for confidence ratings (C) and response latency (L). Corresponding correlations and the main diagonal are indicated in each panel. Data points correspond to the average AUCs across all sessions for each participant. Individual data points are transparent such that darkness indicates density of points in a given area. **A:** Comparison of the areas under the entire ROC functions.  **B:** Comparison of the areas under the ROC functions for “old” (“O”)
responses only. C: Comparison of the areas under the ROC functions for “new” (“N”) responses only. Note that the scales in Panel A differ from those in the other two panels.

Figure 3. Conditional response probabilities. Same data as in Figure 1, but probabilities are conditioned on the respective classification responses. Separate ROC functions for “old” and “new” judgments are generated by cumulatively adding target and lure probabilities with decreasing strengths to the hit and FA values respectively for “old” responses and vice versa for “new” responses. The right panels show the corresponding AUCs with 95% confidence intervals. As in Figure 1 blue and yellow shadings correspond to data from targets and lures respectively and red and green correspond to data from “old” and “new” responses respectively.

Figure 4. Scatter-plots comparing areas under the “old” (“O”) and “new” (“N”) ROC curves (AUC) for confidence ratings (C; Panel A) and response latency (L; Panel B). Corresponding correlations and the main diagonal are indicated in each panel. Individual data points are transparent such that darkness indicates density of points in a given area.

Figure 5. Mean areas under the ROC curves (AUC) across participants for targets that were previously recalled and those that were previously unrecalled within each session (the same set of lures were used in the calculation of both sets of AUCs). Separate AUCs are shown for ROC functions based on confidence ratings (C), response latency (L) as well as corresponding ROC functions conditioned on the classification response (“O”: “old”, “N”:”new”). The darker shadings in the lower parts of the C and L bars indicate the portions of the total area that is attributable to the area of the random ROC. Error bars show the 95% confidence intervals.

Figure 6. Scatter-plots comparing areas under the ROC functions for targets that were previously recalled and those that were previously unrecalled within each session (the same
set of lures were used in the generation of both ROC functions). Separate scatter-plots are shown for ROC functions based on confidence ratings (A) vs. response latency (B). Corresponding correlations and the main diagonal are indicated in each panel. Individual data points are transparent such that darkness indicates density of points in a given area.

Figure 7. Scatter-plots comparing areas under the ROC functions for targets that were previously recalled and those that were previously unrecalled within each session (the same set of lures were used in the generation of both ROC functions). Separate scatter-plots are shown for ROC functions based on confidence ratings (top row, Panels A and B) vs. response latency (bottom row, Panels C and D) and “old” (left column, Panels A and C) vs. “new” (right column, Panels B and D) responses. Corresponding correlations and the main diagonal are indicated in each panel. Individual data points are transparent such that darkness indicates density of points in a given area.
Assessing the strength of cognitive states, Figure 1
Assessing the strength of cognitive states, Figure 2
Assessing the strength of cognitive states, Figure 3
Assessing the strength of cognitive states, Figure 4
Assessing the strength of cognitive states, Figure 5
Assessing the strength of cognitive states, Figure 6
Assessing the strength of cognitive states, Figure 7

![Graph showing correlation between confidence and latency for "old" and "new" categories.](image)