Constructing and model-fitting receiver operator characteristics using continuous data.

Josephine A. Urquhart[1], and Akira R. O'Connor[1]

[1]School of Psychology & Neuroscience, University of St Andrews

Correspondence concerning this article should be addressed to Josephine Urquhart: School of Psychology, University of St Andrews, St Mary's College, South Street, St Andrews, Fife, KY16 9JP, Scotland, UK.

E-mail: jau2@st-andrews.ac.uk

Telephone number: 01334 462037

Keywords: Memory, Recognition, Receiver operator characteristics, Signal detection.

Abstract

Receiver operating characteristics (ROCs) are plots which provide a visual summary of a classifier's decision response accuracy at varying discrimination thresholds. Typical practice, particularly within psychological studies, involves plotting an ROC from a limited number of discrete thresholds before fitting signal detection parameters to the plot. We propose that additional insight into decision-making could be gained through increasing ROC resolution, using trial-by-trial measurements derived from a continuous variable, in place of discrete discrimination thresholds. Such continuous ROCs are not yet routinely used in behavioural research, which we attribute to issues of practicality (i.e. the difficulty of applying standard ROC model-fitting methodologies to continuous data). Consequently, the purpose of the current article is to provide a documented method of fitting signal detection parameters to continuous ROCs. This method reliably produces model fits equivalent to the unequal variance least squares method of model-fitting (Yonelinas et al., 1998), irrespective of the number of data points used in ROC construction. We present the suggested method in three main stages: I) building continuous ROCs, II) model-fitting to continuous ROCs and III) extracting model parameters from continuous ROCs. Throughout the article, procedures are demonstrated in Microsoft Excel, using an example continuous variable: reaction time, taken from a single-item recognition memory. Supplementary MATLAB code used for automating our procedures is also presented in Appendix B, with a validation of the procedure using simulated data shown in Appendix C.

Introduction.

Binary classifications, made under conditions of uncertainty, can be conceptualised as the process of discriminating between target signal and random noise. Signal detection theory describes the process of quantifying a classifier's discrimination accuracy under uncertain conditions. First developed to assess the accuracy of pulse-type radar (Marcum, 1960), signal detection theory was subsequently applied to psychological research in order to examine the dynamics of human decision-making (Green & Swets, 1966). In psychological experimentation signal and noise are often presented to a participant in the form of two stimulus types: i) a target stimulus, the equivalent of the target signal and ii) a lure stimulus, the equivalent of random noise. The participant is required to act as a classifier, indicating whether each stimulus presented is a target or lure. In response to a target stimulus the accuracy of a participant's report can be described as either a hit (accurate classification as a "target") or a miss (inaccurate classification as a "lure"). In response to a lure stimulus the accuracy can be described as either a correct rejection (accurate classification as a "lure") or a false alarm (inaccurate classification as a "target").

In psychological experimentation, signal detection theory users typically assume that participant classification decisions are based upon two underlying Gaussian distributions (MacMillan & Creelman, 2004). Figure 1 shows the probability density functions of such underlying distributions: the lure stimulus distribution is associated with low evidence strength; the target stimulus distribution with higher evidence strength; and there is considerable overlap between the two distributions at intermediate strengths. Within any single classification trial, the stimulus' strength is determined by the probability density function of the distribution from which it is drawn. A stimulus will be classified according to where a participant's decision criterion rests along the evidence strength axis. As such, stimuli with evidence strength below the decision criterion will be classified as “lures”, whilst stimuli with evidence strength above the decision criterion will be classified as “targets”.

A receiver operator characteristic (ROC) graph provides a visual summary of a classifier’s ability to differentiate between signal and noise at varying thresholds of discrimination. In psychological experimentation, an ROC typically provides a visual summary of participant classification performance. In this form, each participant’s ROC consists of points representing the cumulative hit rate against false alarm rate, transitioning to the miss rate against correct rejection rate, at several incrementing thresholds of discrimination (Egan, 1975). Once a participant ROC is constructed the next step in signal detection based analysis involves fitting it with a theoretical model and extracting descriptive model parameters. One of the most widely used models within memory research is the unequal variance (UEV) signal detection model (see Figure 1B; Egan, 1958). The two parameters recovered during UEV model fits are the standard deviation of the target distribution (σ) and the sensitivity (*d'*; pronounced “d prime”). In recognition-based psychological experimentation, σ summarises the spread of predicted target stimulus distribution relative to the lure stimulus distribution, which typically has a standard deviation held at 1. This increased spread is generally attributed to additional variance introduced by encoding procedures in the evidence strength associated with each old item (cf. Koen, Aly, Wang, & Yonelinas, 2013). The *d'* is the displacement of the target stimulus distribution mean from the lure stimulus distribution mean. As such these two parameter estimates derived from the application of a widely used signal detection theory model can be used to summarise and assess different qualities of participant performance.

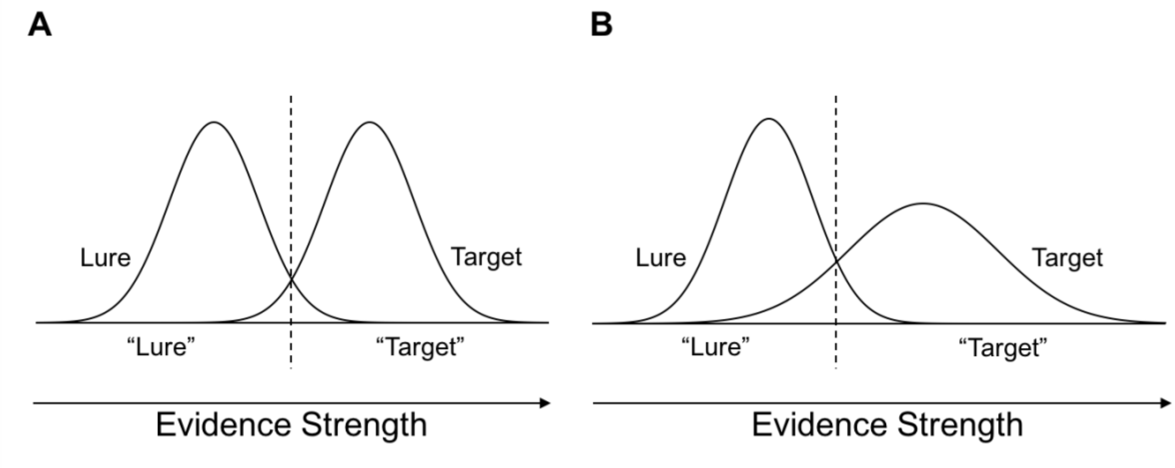

Figure 1. Panel A: Example underlying lure and target distributions, shown as probability density functions along an *x*-axis representing evidence strength. The central dashed line shows the “target”/“lure” decision criterion. In this hypothetical example the criterion is shown in optimal position, such that if a participant were to classify stimuli of evidence strength below it as a “lure” and stimuli of evidence strength above it as a “target”, miss and false alarm rate would both be minimised. **Panel B:** Example lure and target distributions within an unequal variance (UEV) signal detection framework. In this instantiation of signal detection, the lure distribution is a normal distribution, whilst the target distribution has a greater (hence unequal) variance. The central dashed line represents “target”/“lure” criterion, such that stimuli of evidence strength below it will be classified as a “lure” and stimuli of evidence strength above it will be classified as a “target”.

The practice of building and model-fitting ROCs is often confined to a limited range of discrete data points, such that discrimination thresholds are varied a pre-determined number of times. As displayed in Figure 2, psychological tests of recognition memory, are often used to constitute five-point ROC curves, with six numerical confidence ratings indicating, low, medium, and high confidence applied, to “old” and “new” response decisions. In the context of recognition, the range of possible confidence levels (i.e. discrimination thresholds) generally comprises a set number of discrete, experimenter-defined values. Using such values could therefore introduce a level of artificiality and subjectivity into ROC production.

Accordingly, ROC curves derived from a greater range of continuous data points could have a variety of potential applications. Within psychological research, a continuous measure of confidence has been produced on a trial-wise basis by single-unit recording of neural-firing rate in both humans (Rutishauser *et al.,* 2015) and animals (Zhang, Riehle & Requin, 1997), as well as “old”/“new” reaction time (Merkow, Burke & Kahana, 2015; Weidemann & Kahana, 2016). Similarly, physiological methods, including fMRI-measured blood oxygenation level dependent (BOLD) signal, galvanic skin response (GSR), pupil dilation or blood pressure also have the potential to provide continuous confidence information. Such measures offer significant advantage into the study of decision-making as they generate continuous data, without requiring participants to provide a subjective confidence rating on a potentially unwieldy confidence scale (such as the 100-point visual analogue scale described in Mickes, Wixted & Wais, 2007). Thus, a continuous ROC derived from this data could have the potential to reveal nuances in decision sensitivity that would otherwise be masked by coarser ROCs limited by the number of confidence judgments a participant can hold in working memory. Furthermore, in contrast to subjective ratings technique, using a continuous measurement system also allows confidence levels to vary on a trial-by-trial basis within its own natural range. However, in previous instances where continuous data has been collected, including the studies listed above, ROC analysis has still involved reducing data resolution dividing continuous values by size into a fixed number of discrete discrimination thresholds. Again, this reduction procedure introduces a degree of artificiality into ROC production, removing the advantages collecting continuous data offers into ROC analysis. Crucially, we suggest a continuous form of ROC, which does not require data to be reduced into discrete categories, could increase the resolution with which binary decision classifications are interrogated. Furthermore, it might also help to prevent the issue of bias associated with previous research in which arbitrary limits have been placed on the number of discrimination thresholds.

		Hits	False Alarms
"Target" confidence	3	39	2
	2	28	14
	1	17	15
		Misses	**Correct Rejections**
"Lure" confidence	1	15	30
	2	6	27
	3	0	7
Total		105	95

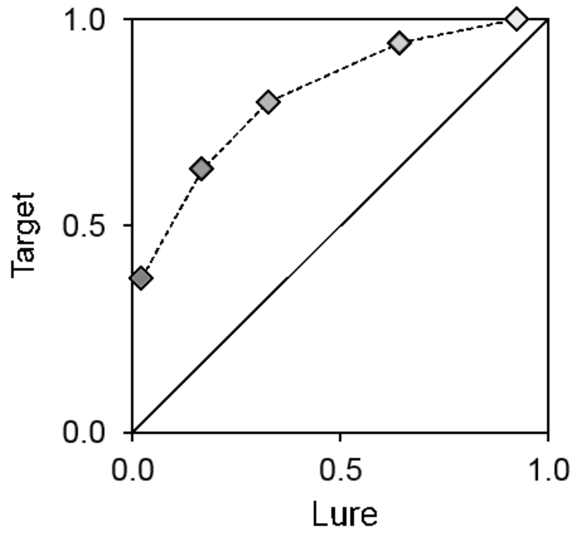

Figure 2. A table of the hypothetical responses and ROC curve, that could have been derived from a recognition memory task in which participants categorise test-items as either old (in this example the target stimuli) or new (in this example the lure stimuli). To generate an ROC, participants would be required to rate their confidence in each "old"/"new" response decision. This is done with an ordinal rating system (i.e. 1: low confidence, 2: medium confidence and 3: high confidence). Data points represent the cumulative hit (*y*-axis) against false alarm (*x*-axis) rates for "old" responses of decreasing confidence transitioning to miss (*y*-axis) against correct rejection (*x*-axis) rates for "new" responses of increasing confidence. The position of the third point from the left shows that responses of at least low confidence "old" (i.e. high, medium or low confidence "old" responses) were given to .800 of all target items (i.e. total hits divided by total target items) and .326 of all lure items (i.e. total false alarms divided by total lure items). For the fourth point from the left, any "old" response, or a low confidence "new" response was given to .943 of all target items and .642 of all lure items. Note that a point is not shown for "lure" high confidence responses because the cumulative calculation process means this point will always lie at [1,1] with *all* target and lure items given at least this response.

Continuous ROCs are not routinely used within psychological research. This can partly be attributed to issues of practicality. For instance, difficulties arise when adapting the commonly used least squares method of fitting the UEV model to continuous data (Yonelinas et al., 1998). The Microsoft Excel Solver function is often used to carry out this least squares fitting on ROCs constructed from discrete data points (as specified by Harris, 1988), but with potentially infinite continuous data points this method will not provide an accurate fit. Other fitting functions can complete more complex least squares fitting procedures. However, these iterative procedures can settle within local minima and require frequent manual parameter estimate re-setting over the course of a single fit. Consequently, carrying out least squares fitting on studies with large numbers of participants or large quantities of data can become both time-consuming and complex.

The purpose of this article is to provide an explanation and guide to a straightforward method for constructing and model-fitting continuous ROCs. Critically, this method does not require data to be condensed into a fixed number of discrete categories, reliably recovering UEV signal detection model parameters σ and d' from continuous data, regardless of the number of points contributing to the ROC. Over the course of the article we describe and demonstrate a series of procedures to: I) calculate the *x*- and *y*- coordinates necessary to plot continuous ROCs using a non-parametric data ordering technique, which can be used under conditions where continuous data fails to meet parametric assumptions; II) transform continuous ROC points into *z*-space and fit a line of best-fit; and III) recover the model parameters necessary for further analysis from the line of best-fit. Details of how to achieve each of these steps will be demonstrated using Microsoft Excel with further examples of code useful for automating the model-fitting process in MATLAB[1] (The MathWorks Inc., Natick, MA, 2000).

Methods.

I. Building continuous ROC curves.

The data used in the following example were generated using a standard single-item recognition task, the type of which is typically used to assess recognition memory for word stimuli (Bayley et al., 2008; Jang et al., 2009 & Khoe et al., 2000). This task required that participants first study a list of serially-presented words. Next, they responded to a test list comprising targets–words presented at study–and lures–words not presented at study. Participants indicated whether they thought a word was "old" (thought to be a target) or "new" (thought to be a lure) by pressing a different key for each response (note that discrete numerical confidence responses were not collected; a sample of these data are shown in Figure 3A and the complete dataset are available in Appendix A). Our proposed methodology is described in terms of this recognition task example, with instructions of how to achieve each step in Excel[2] presented throughout.

---

[1] This code module has been tested using MATLAB versions 2013a-2016b. However, MATLAB versions 2010b onwards should also be capable of running code correctly. Functions called with the module also require that MATLAB Statistics Toolbox be installed.

[2] Excel instructions were generated using Windows Microsoft Office Professional Plus 2010: Microsoft ® Excel ® 2010, Version 14.0.7153.5000 and Microsoft ® Excel ® for Apple Macintosh 2011, Version 14.4.0 (140226). Our procedure does not require the use of Macros or visual basic, therefore methodology should be interchangeable across additional versions of excel and other data processing platforms.

A

	A	B	C	D	E	F
1	Participant	Word	Word Status	Response (ol	Accuracy	RT
2	4	comparison	2	2	1	2.281
3	4	technique	1	1	1	1.111
4	4	receiver	1	1	1	0.863
5	4	princess	1	1	1	1.123
6	4	morality	1	2	0	1.825
7	4	member	1	2	0	1.730
8	4	liberation	1	1	1	1.104
9	4	jungle	1	1	1	1.083
10	4	income	2	1	0	1.135
11	4	hose	2	2	1	2.060
12	4	direction	2	2	1	1.783
13	4	destruction	1	1	1	1.152
14	4	democracy	1	1	1	1.000
15	4	converter	1	2	0	1.841
16	4	chest	2	2	1	2.680
17	4	character	2	2	1	2.176
18	4	bread	1	2	0	2.267
19	4	strength	1	1	1	1.008
20	4	patience	1	1	1	1.145
21	4	conversation	2	2	1	2.659
22	"	"	"	"	"	"

B

	A	B	C	D	E	F
1	Participant	Word	Word Status	Response (ol	Accuracy	RT
2	4	technique	1	1	1	1.111
3	4	receiver	1	1	1	0.863
4	4	princess	1	1	1	1.123
5	4	liberation	1	1	1	1.104
6	4	jungle	1	1	1	1.083
7	4	income	2	1	0	1.135
8	4	destruction	1	1	1	1.152
9	4	democracy	1	1	1	1.000
10	4	strength	1	1	1	1.008
11	4	patience	1	1	1	1.145
12	"	"	"	"	"	"
13	4	comparison	2	2	1	2.281
14	4	morality	1	2	0	1.825
15	4	member	1	2	0	1.730
16	4	hose	2	2	1	2.060
17	4	direction	2	2	1	1.783
18	4	converter	1	2	0	1.841
19	4	chest	2	2	1	2.680
20	4	character	2	2	1	2.176
21	4	bread	1	2	0	2.267
22	4	conversation	2	2	1	2.659
23	"	"	"	"	"	"

C

	A	B	C	D	E	F
1	Participant	Word	Word Status	Response (ol	Accuracy	RT
2	4	destruction	1	1	1	1.152
3	4	patience	1	1	1	1.145
4	4	income	2	1	0	1.135
5	4	princess	1	1	1	1.123
6	4	technique	1	1	1	1.111
7	4	liberation	1	1	1	1.104
8	4	jungle	1	1	1	1.083
9	4	strength	1	1	1	1.008
10	4	democracy	1	1	1	1.000
11	4	receiver	1	1	1	0.863
12	"	"	"	"	"	"

	A	B	C	D	E	F
1	Participant	Word	Word Status	Response (ol	Accuracy	RT
2	4	member	1	2	0	1.730
3	4	direction	2	2	1	1.783
4	4	morality	1	2	0	1.825
5	4	converter	1	2	0	1.841
6	4	hose	2	2	1	2.060
7	4	character	2	2	1	2.176
8	4	bread	1	2	0	2.267
9	4	comparison	2	2	1	2.281
10	4	conversation	2	2	1	2.659
11	4	chest	2	2	1	2.680
12	"	"	"	"	"	"

D

	A	B	C	D	E	F
1	Participant	Word	Word Status	Response (ol	Accuracy	RT
2	4	destruction	1	1	1	1.152
3	4	patience	1	1	1	1.145
4	4	income	2	1	0	1.135
5	4	princess	1	1	1	1.123
6	4	technique	1	1	1	1.111
7	4	liberation	1	1	1	1.104
8	4	jungle	1	1	1	1.083
9	4	strength	1	1	1	1.008
10	4	democracy	1	1	1	1.000
11	4	receiver	1	1	1	0.863
12	"	"	"	"	"	"
13	4	member	1	2	0	1.730
14	4	direction	2	2	1	1.783
15	4	morality	1	2	0	1.825
16	4	converter	1	2	0	1.841
17	4	hose	2	2	1	2.060
18	4	character	2	2	1	2.176
19	4	bread	1	2	0	2.267
20	4	comparison	2	2	1	2.281
21	4	conversation	2	2	1	2.659
22	4	chest	2	2	1	2.680
23	"	"	"	"	"	"

Figure 3. Panel A: A section of the example raw data produced by a single-item recognition task assessing recognition memory for word stimuli. Column A is participant number. Column B is the presented word stimulus. Column C is word status: target (1) or lure (2). Column D is participant response: “old” (1) or “new” (2). Column E is response accuracy: inaccurate (0) or accurate (1). Column F is the time taken to render a response (RT; in seconds). **Panel B:** Data sorted by response. The sort was achieved using the Excel ‘sort’ menu operation (lowest-to-highest) on the basis of Column D content. **Panel C:** Response-sorted data, sorted again by RT. This sort was achieved by ordering each response category on the basis of Column F content (“old” responses from highest to lowest, “new” responses from lowest to highest). **Panel D:** Sorted data concatenated vertically, with “old” responses before “new” responses.

The RT column (Column F) in Figure 3A gives the time taken to make each "old"/"new" response. This is the continuous variable which will be used to construct the ROC presented here. Previous research indicates response times for accurate decision response trials (hits & correct rejections) and inaccurate trials (misses & false alarms) exhibit a varied relationship (Ratcliff & Rouder, 1998). In general, where a binary discrimination task is harder and task instruction places an emphasis on participant accuracy, correct response trials tend to be quicker compared to error trials (Luce, 1986; Swensson, 1972). In accordance with this observation, reaction time has been used previously to construct discrete ROCs, with continuous data binned into several fixed discrimination thresholds (e.g. Merkow *et al.*, 2015; Weidemann & Kahana, 2016). Similarly, continuous ROC construction in the current article is based on the premise that as the likelihood of accurately classifying a word's objective target or lure status increases, the faster an "old"/"new" decision response will be reported. However, the aim of the current article is not to demonstrate that reaction time varies as a function of subjective confidence, as there is well-documented research that supports this theory (see Ratcliff & Starns, 2009, for a general overview). Instead, we intend to demonstrate a novel procedure for continuous ROC construction and analysis that can be used to assess multiple forms of continuous data, including reaction time. Notably, we have chosen to exemplify our procedure with reaction time because, in addition to its known relationship to confidence, it is a variable collected by the majority of recognition researchers. Accordingly, it can be used as a proof of concept, allowing researchers to explore and validate the proposed methodology before extending it to more psychologically interesting, challenging, and rewarding variables.

Classifying an accurate or inaccurate response.

When collecting continuous measurements for two response types (e.g. “old” and “new” responses) the method for determining which decision has been made can be dependent on, or independent of, the continuous data collected. With some data types, including the example RTs, continuous measurements are collected from two sources (separate old and new keys). In this situation, it is the presence of data at one source paired with its absence at the other which denotes decision response type (e.g. a RT of 0.5s for the “new” response and no RT for the “old” response indicates that a “new” response has been made). In experimental contexts where continuous data are collected simultaneously from two sources (e.g. fMRI-measured BOLD response magnitude from two localised brain regions) the greater or lesser of the two values might be used to indicate response type. Where continuous data are collected from one source (e.g. blood pressure) a definitive response decision is not necessarily produced. Consequently, response type should be determined according to an additional nominal-type response method, such as a keyboard response made alongside the collection of continuous data).

Each response (i.e. “old” or “new”) should be compared against its objective status (i.e. target or lure) to determine its accuracy. For the next stage of ROC construction the data should be split according to response type. In abstract terms, two response types are possible: Response 1 (classifying the objective status of a stimulus as a target) and Response 2 (classifying the objective status of a stimulus as a lure). Therefore, in the current example where old words are targets and new words are lures, Response 1 is an “old” response and Response 2 is a “new” response. Figure 3B shows how these data would appear when split using Excel’s ‘sort’ menu operation. The alternative MATLAB code to achieve the split and all subsequent operations is presented in Appendix B.

ROC construction using data sorting and cumulative plotting.

ROC construction is separated into two main stages: i) data sorting and ii) cumulative plotting of ROC points.

Stage i) requires that each data subset be sorted separately on the basis of the continuous data recorded. The experimental context, purpose of the ROC and continuous variable type will determine the sorting direction in line with the principles of ROC construction. All Response 1s should be sorted such that response accuracy likelihood decreases across the data subset (from highest to lowest hypothesised discrimination threshold. All Response 2s should then be sorted so that accuracy likelihood increases across the data subset (from lowest to highest hypothesised discrimination threshold). Therefore, in the current example “old” responses are sorted from lowest to highest RT, whilst “new” responses are sorted from highest to lowest RT. Figure 3C shows how these sorted data subsets would appear in Excel. At this stage, it is important to note that Responses 1 and 2 should be sorted independently of each other. Therefore, where measurements have been drawn from separate and potentially variable sources (i.e. when extracting fMRI-measured BOLD response from two localised brain regions) useful comparison can still be made and a complete ROC curve can be built.

Stage ii) requires that the two sorted arrays derived from stage i) be concatenated vertically such that decision Response 1 and Response 2 subsets are recombined into a single set. Figure 3D presents Excel procedures for the current example. The ROC *x*- and *y*- axis step size should then be calculated according to Formulae 1 and 2. These formulae determine the cumulative increments from [0,0] to [1,1] contributed by hits and misses (Formula 1; where I_y is the *y*-axis increment and N_1 is the number of items for which Response 1 would have been accurate) and correct rejections and false positives (Formula 2; where I_x is the *x*-axis increment and N_2 is the number of items for which Response 2 would have been accurate).

$$I_y = \frac{1}{N_1} \tag{1}$$

$$I_x = \frac{1}{N_2} \tag{2}$$

In the current example the Excel ‘COUNTIF’ function was used to find the number of target status and lure status words for which an “old”/“new” response (accurate or inaccurate) was made. Formulae 1 and 2 applied to the current example (the entire dataset from Appendix A as opposed to the subsample shown in Figure 3) use N_1 and

N_2 values of 56 (56 targets and 56 lures were presented and responded to), hence I_y and I_x are both equal to .018.

The *y*- and *x*- axis increments should now be cumulatively added to each continuous variable-ordered trial. Figure 4 illustrates the addition of I_y to ROC-*y* every time a target is accurately identified (hit). Likewise, I_x is added cumulatively to ROC-*x* when a lure is accurately responded to (correct rejection). I_y is also added to ROC-*y* when a target is inaccurately responded to (miss) and I_x is added to ROC-*x* when an

inaccurate response to a lure is made. The coordinates generated by this process produce a curve, as steps along the *x*- and *y*- axes. This is illustrated in Figure 5.

	A	B	C	D	E	F	G	H
1	Participant	Word	Word Status	Response (ol	Accuracy	RT	ROC-*x*	ROC-*y*
2	4	receiver	1	1	1	0.863	0.000	0.018
3	4	democracy	1	1	1	1.000	0.000	0.036
4	4	strength	1	1	1	1.008	0.000	0.054
5	4	jungle	1	1	1	1.083	0.000	0.071
6	4	liberation	1	1	1	1.104	0.000	0.089
7	4	technique	1	1	1	1.111	0.000	0.107
8	4	princess	1	1	1	1.123	0.000	0.125
9	4	income	2	1	0	1.135	0.018	0.125
10	4	patience	1	1	1	1.145	0.018	0.143
11	4	destruction	1	1	1	1.152	0.018	0.161
12	"	"	"	"	"	"	"	"
13	4	chest	2	2	1	2.680	0.196	0.696
14	4	conversation	2	2	1	2.659	0.214	0.696
15	4	comparison	2	2	1	2.281	0.232	0.696
16	4	bread	1	2	0	2.267	0.232	0.714
17	4	character	2	2	1	2.176	0.250	0.714
18	4	hose	2	2	1	2.060	0.268	0.714
19	4	converter	1	2	0	1.841	0.268	0.732
20	4	morality	1	2	0	1.825	0.268	0.750
21	4	direction	2	2	1	1.783	0.286	0.750
22	4	member	1	2	0	1.730	0.286	0.768
23	"	"	"	"	"	"	"	"

1a Row 2, Column E shows an accurate (1) "old" response (hit), therefore 0.018 is added to ROC-*y* (row 2 column H).

1b Row 2, Column G: IF(C1 = 1, 0.018, 0)
Row 2, Column H: IF(C1 = 0, 0.018, 0)

2a Row 3, Column E again shows an accurate (1) "old" response (hit), therefore 0.018 is added cumulatively to ROC-*y* (row 3 column H).

2b Row 3, Column G: IF(C1 = 1, 0.018 + G2, G2)
Row 3, Column H: IF(C1 = 0, 0.018 + G2, G2)

3a Row 9 Column E shows an inaccurate (0) "old" response to a lure (false alarm) therefore 0.018 is added cumulatively to ROC-*x* (row 8 column G).

3b Row 8, Column G: IF(C1 = 1, 0.018 + G8, G8)
Row 8, Column H: IF(C1 = 0, 0.018 + G8, G8)

Figure 4. Example data section with calculated ROC-*y* and ROC-*x* coordinates at Column H and G, respectively. All -*y* calculations relate to trials for which an "old" response was made, and all -*x* calculations relate to trials for which a "new" response was made. Boxes 1a, 2a and 3a provide examples of ROC coordinate calculation for three separate trials. Boxes 1b, 2b and 3b provide Excel formula inputs to carry out these calculations.

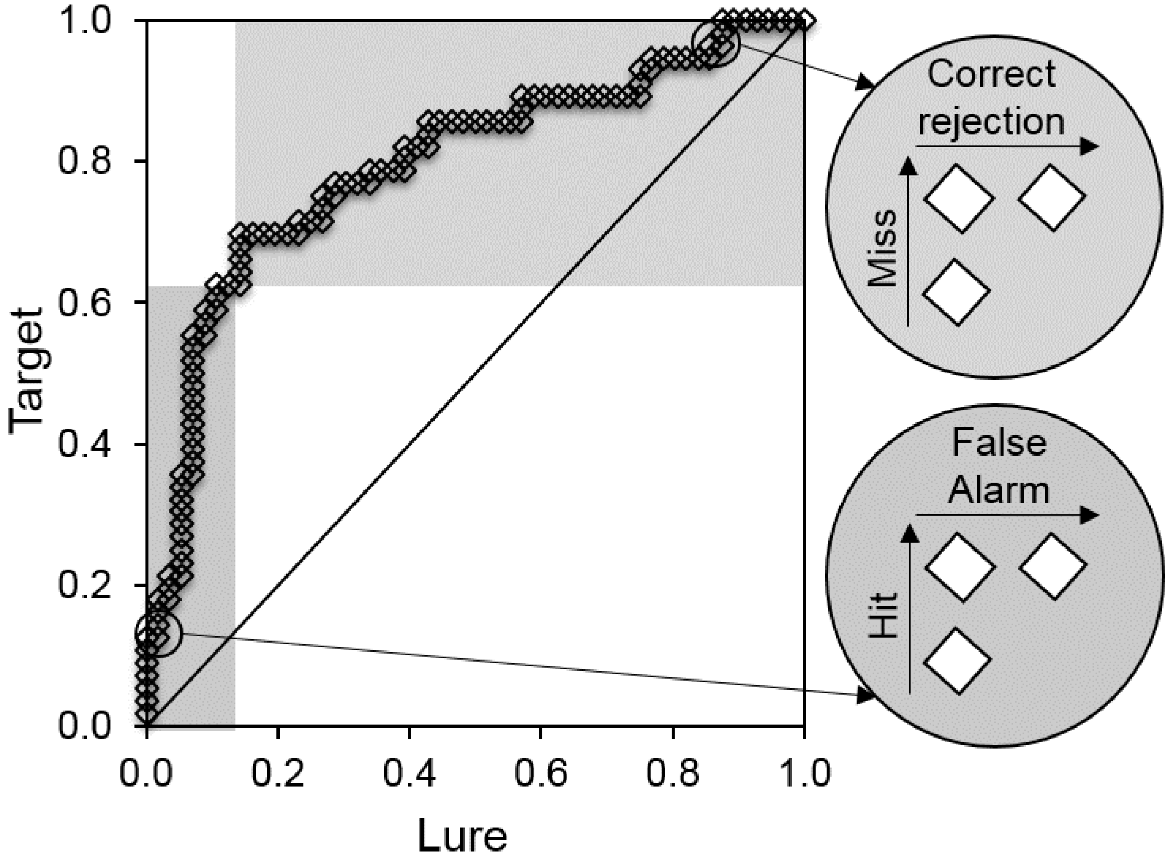

Figure 5. A continuous ROC curve based on the recognition memory data in Appendix A. All “old” responses are plotted to the left of the dashed line; all “new” responses are plotted to the right (their order corresponds to the magnitude of the continuous variable, as detailed in Figure 3). The curve is built by cumulatively adding two formulae-determined increments for each hit, miss, correct rejection and false alarm. The region circled in the bottom quadrant shows a hit (cumulative increment along the *y*-axis) and false alarm (cumulative increment along the *x*-axis) for two items responded to as “old”. The region circled in the top quadrant shows a miss (cumulative increment along the *y*-axis) and correct rejection (cumulative increment along the *x*-axis) for two items responded to as “new”.

II. Model-fitting to continuous ROC curves.

The next stage outlined below can be used to optimally fit continuous ROCs to a signal detection model in which the target stimulus probability distribution variance is free to vary independently of the lure distribution variance (as in the UEV signal detection model of recognition memory). It requires that the ROCs plotted using Step I be transformed into *z*-space (MacMillan & Creelman, 2004), a line of best-fit determined, and the estimated parameters of the signal detection model calculated from the properties of the line of best-fit. Appendix C provides an account of z-transformation model-fitting simulations, which were used to quantify the precision of this methodology. Critically, it should be noted these data suggest our proposed will be most robust when fitting to ROC curves with underlying UEV signal detection parameters *d′* and *σ* within the ranges of, [0.3-2.5] and [0.8-2.0], respectively.

Transforming ROC coordinates into z-space.

First, ROC coordinates must be transformed into *z*-space through normalisation. This is done independently for both the ROC-*x* and ROC-*y* coordinates. Figure 6 illustrates the Excel procedure for the current data using the 'NORMINV' function. For ROC coordinate pairs containing a value equal to 1 or 0, normalisation is not possible. Therefore, a method for disregarding coordinates containing these values should be incorporated into any automated or batch normalisation procedure[3]. The normalised

---

[3] Because ROC coordinates derived from our proposed method will typically contain at least one coordinate pair with a 0 (i.e. the first pair) and a 1 (i.e. the last pair), it is necessary to have an ROC consisting of 4 data points or greater, which do not contain a 1 or 0, in order for an accurate UEV-based model to be fitted. It should be noted that, the precision of the fit will increase with the inclusion of greater numbers of data points. Accordingly, we recommend our proposed method as a means of model-fitting to continuous data, as it is most effective when using data sets containing 50 unique points or more.

ROC (zROC) coordinates provide a linearised version of the continuous ROC curve, shown in Figure 7.

Determining a zROC line of best-fit and recovering its equation

The next stage requires that a line of best-fit be calculated for the zROC data. The equation for this line, in the form $y = mx + c$ (overlaid on the plot in Figure 7), is necessary to calculate the signal detection parameter estimates for the ROC data.

	A	B	C	D	E	F	G	H	I	J
1	Participant	Word	Word Status	Response (ol	Accuracy	RT	ROC-*x*	ROC-*y*	zROC-*x*	zROC-*y*
2	4	receiver	1	1	1	0.863	0.000	0.018		
3	4	democracy	1	1	1	1.000	0.000	0.036		
4	4	strength	1	1	1	1.008	0.000	0.054		
5	4	jungle	1	1	1	1.083	0.000	0.071		
6	4	liberation	1	1	1	1.104	0.000	0.089		
7	4	technique	1	1	1	1.111	0.000	0.107		
8	4	princess	1	1	1	1.123	0.000	0.125		
9	4	income	2	1	0	1.135	0.018	0.125	-2.100	-1.150
10	4	patience	1	1	1	1.145	0.018	0.143	-2.100	-1.068
11	4	destruction	1	1	1	1.152	0.018	0.161	-2.100	-0.992
12	"	"	"	"	"	"	"	"	"	"
13	4	chest	2	2	1	2.680	0.196	0.696	-0.854	0.514
14	4	conversation	2	2	1	2.659	0.214	0.696	-0.792	0.514
15	4	comparison	2	2	1	2.281	0.232	0.696	-0.732	0.514
16	4	bread	1	2	0	2.267	0.232	0.714	-0.732	0.566
17	4	character	2	2	1	2.176	0.250	0.714	-0.674	0.566
18	4	hose	2	2	1	2.060	0.268	0.714	-0.619	0.566
19	4	converter	1	2	0	1.841	0.268	0.732	-0.619	0.619
20	4	morality	1	2	0	1.825	0.268	0.750	-0.619	0.674
21	4	direction	2	2	1	1.783	0.286	0.750	-0.566	0.674
22	4	member	1	2	0	1.730	0.286	0.768	-0.566	0.732
23	"	"	"	"	"	"	"	"	"	"

1 Row 2,, Column I: IF(AND(G2<>0, G2<0.999, NORM.INV)(G2), " ")
Row 2,, Column j: IF(AND(G2<>0, G2<0.999, NORM.INV)(G2), " ")

Figure 6. Normalisation of ROC-*y* and ROC-*x* in order to produce zROC-*y* zROC-*x* (shown in columns J and I, respectively). Box 1 provides an example of the formulae used in Excel to carry out z-transformation. Note: rows 2-8 do not display values because the ROC coordinate pairs (specifically ROC-*x*) contain a 0, which cannot be normalised.

III. Obtaining model parameters from zROCs.

UEV signal detection parameters *σ* (Formula 3) and *d′* (Formula 4) can be calculated

using the zROC line of best-fit equation in the form $y = mx + c$.

$$\sigma = \frac{1}{m} \tag{3}$$

$$d' = c \times \sigma \tag{4}$$

Formulae 3 and 4 applied to the zROC in Figure 7 yield $\sigma = 1.227$ and $d' = 1.241$.

To illustrate that the *σ* and *d′* parameters match the data used to recover them, hypothetical probability density functions from which the obtained ROC data are proposed to originate can be examined, Figure 7B, and theoretical ROCs overlaid upon the experimental data, Figure 7C. Participant-level parameters can then be treated in terms of conventional analysis strategies, for instance taking them to the group level and subjecting them to inferential statistical analysis to determine the effects of manipulations or treatment conditions on discrimination ability and threshold. Initial parameters can also be used to reduce search boundaries of more complex computerised fitting strategies, such that the number of possible iterations to provide an optimal curve fit is reduced.

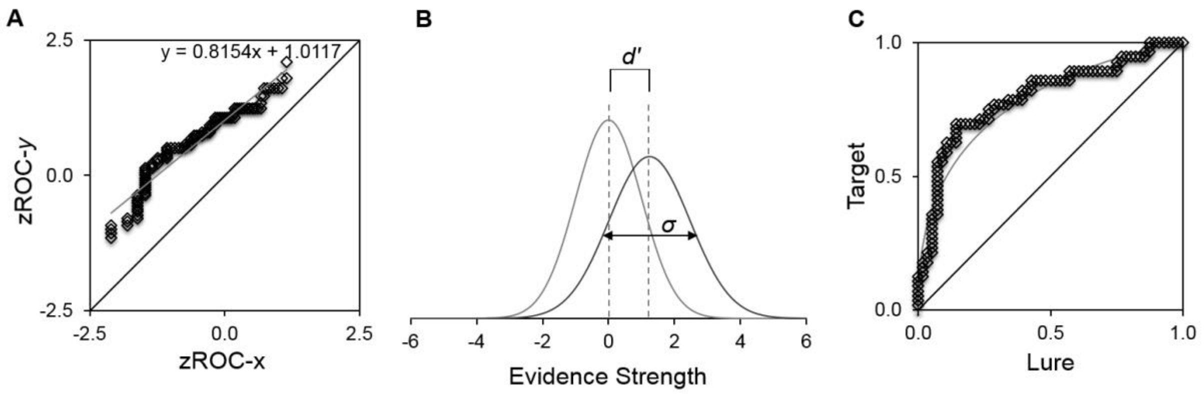

Figure 7. Panel A: A linearised version of the initial continuous ROC (Figure 5), plotted using normalised coordinates: zROC-*y* and zROC-*x*. As with the ROC plot, the zROC is shown with "old" responses on the *y*-axis plotted against "new" responses on the *x*-axis. A line of best-fit overlaid in red (Windows: Chart Tools > Layout > Trendline > More Trendline Options > Trendline, [specify 'Linear']; Apple Macintosh: Chart Layout > Trendline > Trendline Options > Type [specify 'Linear Trendline']) and the equation for the line overlaid as text (Windows: Chart Tools > Layout > Trendline > More Trendline Options > Trendline, [specify 'Display Equation on chart']; Apple Macintosh: Chart Layout > Trendline > Trendline Options > Type [specify 'Display Equation on Chart']; Additionally, the 'SLOPE' and 'INTERCEPT' functions can be used in Excel to calculate *m* and *c* values, respectively**). Panel B and Panel C:** Additional plots derived from recovered UEV signal detection parameters *σ* and *d'*. **Panel B:** Probability density functions (PDFs) for target stimuli (right) and lure stimuli (left). In accordance with the UEV signal detection model, the lure stimulus PDF is a normal distribution (mean = 0; standard deviation = 1). The target stimulus PDF has mean = *d'* (in this case 1.241) and standard deviation = *σ* (in this case 1.227). **Panel C:** A theoretical ROC curve in red, overlaid onto the continuous ROC data used to recover it (Figure 5).

Conclusion.

The practice of building ROCs is generally restricted to using a limited number of discrete discrimination thresholds. However, signal detection theory applied to continuous data has the potential to offer additional insight into decision-making, within psychological experimentation and beyond.

A primary function of a ROC is to act as a visual summary of classificatory performance, which can be compared across classifiers, manipulations or treatment conditions in a straightforward manner. Step I of our proposed methodology demonstrates that by incorporating continuous data, without reducing it into discrete discrimination thresholds, the visual scope of the ROCs can be further augmented. Continuous ROCs provide a complete trial-by-trial summary of discrimination ability. Thus, they incorporate maximal information content, without increasing the complexity of the mathematical framework by which discrete ROCs are generally interpreted. The increased information content provided within a continuous ROC also allows direct visual access to potential micro-trends in discrimination ability, which might otherwise be condensed into averaged data points. As well as highlighting micro-trends relating to individual classifiers and experimental manipulations, continuous ROCs can also help to determine the impact of random noise or non-random bias (introduced by equipment or methodology) on collated data. Potential strategies for noise or bias reduction can then be applied to data before it is subjected to further analysis. Thus, this may be of particular benefit when using measures associated with reduced signal-to-noise ratio (e.g. fMRI BOLD signal; Krüger & Glover, 2001; Weisskoff, 1996).

Despite offering these advantages, continuous ROCs are not routinely used in experimental research, particularly within the field of psychology. We suggest that the practical issues, which arise when adapting standard model-fitting techniques to continuous data, are in part responsible for the lack of continuous ROC use. To this end steps II and III of our proposed methodology set out to address these issues of practicality, describing and exemplifying a straightforward method of fitting SDT parameters to continuous ROCs, which can be applied efficiently to a variety of data types with minimal mathematical assumptions.

Our proposed methodology provides a simple and logical development of traditional discrete ROC methodology. The purpose of the current article is to highlight the potential benefits associated with advancing conventional ROC based methodologies, demonstrating that continuous data in an uncondensed format is central in facilitating this progression. Overall, we suggest that continued refinement in terms of continuous ROC construction and model-fitting methodology will have significant value across a diverse range of experimental fields.

Disclosures.

Conflicts of Interest: The authors declare that they have no conflicts of interest with respect to the authorship or the publication of this article.

Author Contributions: Both JAU and ARO conceived of and designed the methodology reported below, drafted the manuscript, prepared figures, reviewed drafts of the manuscript and prepared the MATLAB code used to enact the methodology. JAU wrote the code used to simulate the model-fitting procedure and analysed simulation data.

Acknowledgments: JAU would like to gratefully acknowledge the support given by the Economic and Social Research Council 1+3 Studentship Scheme (grant number 1649138).

References.

Egan, J. P. (1958). Recognition memory and the operating characteristic. (Tech. Note AFCRC-TN-58-51). Bloomington: Hearing and Communication Laboratory, Indiana University.

Egan, J. P. (1975). Signal detection theory and ROC-analysis. New York: Academic press.

Bayley, P. J., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2008). Yes/no recognition, forced-choice recognition, and the human hippocampus. *Journal of Cognitive Neuroscience*, *20*(3), 505-512.

Green, D. M. & Swets, J. A. (1966). Signal detection theory and psychophysics. New York: Wiley.

Harris, D. C. (1998). Nonlinear least-squares curve fitting with Microsoft Excel Solver. *Journal of chemical education, 75*(1), 119-121.

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1210-1230.

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*(2), 291.

Khoe, W., Kroll, N. E., Yonelinas, A. P., Dobbins, I. G., & Knight, R. T. (2000). The contribution of recollection and familiarity to yes–no and forced-choice recognition tests in healthy subjects and amnesics. *Neuropsychologia*, *38*(10), 1333-1341.

Koen, J. D., Aly, M., Wang, W. C., & Yonelinas, A. P. (2013). Examining the causes of memory strength variability: Recollection, attention failure, or encoding variability?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1726-1741.

Krüger, G., & Glover, G. H. (2001). Physiological noise in oxygenation-sensitive magnetic resonance imaging. *Magnetic resonance in medicine*, *46*(4), 631-637.

Luce, R. D. (1986). Response times (No. 8). New York: Oxford University Press.

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

Marcum, J. (1960). A statistical theory of target detection by pulsed radar. *Information Theory, IRE Transactions on*, *6*(2), 59-267.

Merkow, M. B., Burke, J. F., & Kahana, M. J. (2015). The human hippocampus contributes to both the recollection and familiarity components of recognition memory. *Proceedings of the National Academy of Sciences*, *112*(46), 14378-14383.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review*, *14*(5), 858-865.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347-356.

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological review*, *99*(3), 518-535.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological review*, *116*(1), 59-83.

Rutishauser, U., Ye, S., Koroma, M., Tudusciuc, O., Ross, I. B., Chung, J. M., & Mamelak, A. N. (2015). Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nature neuroscience*, *18*(7), 1041-1050.

Swensson, R. G. (1972). The elusive tradeoff: Speed vs accuracy in visual discrimination tasks. *Perception & Psychophysics*, *12*(1), 16-32.

Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Royal Society open science*, *3*(4), 150670.

Weisskoff, R. M. (1996). Simple measurement of scanner stability for functional NMR imaging of activation in the brain. *Magnetic resonance in medicine*, *36*(4), 643-645.

Yonelinas, A. P., Kroll, N. E., Dobbins, I., Lazzara, M., & Knight, R. T. (1998). Recollection and familiarity deficits in amnesia: convergence of remember-know, process dissociation, and receiver operating characteristic data. *Neuropsychology*, *12*(3), 323-339.

Zhang, J., Riehle, A., & Requin, J. (1997). Analyzing neuronal processing locus in stimulus–response association tasks. *Journal of mathematical psychology*, *41*(3), 219-236.

Appendices

Appendix A – CompleteDataSorted.xlsx. Complete data set used to generate example continuous ROC (Note: data is presented as it would be following all procedures proposed in the current article).

Appendix B – ContinuousROCMatlabCode.zip. An archived module containing supplementary MATLAB code used for automating our proposed methodology.

Appendix C – **SupplementaryMaterials.docx** Supplementary material examining the output precision y of our proposed z-transformation model-fitting methodology across a range of known *d′* and *σ* UEV signal detection input parameters.