

The ROC Toolbox: A toolbox for analyzing receiver-operating characteristics derived from confidence ratings

Joshua D. Koen¹ · Frederick S. Barrett² · Iain M. Harlow³ · Andrew P. Yonelinas^{3,4}

© Psychonomic Society, Inc. 2016

Abstract Signal-detection theory, and the analysis of receiver-operating characteristics (ROCs), has played a critical role in the development of theories of episodic memory and perception. The purpose of the current paper is to present the ROC Toolbox. This toolbox is a set of functions written in the Matlab programming language that can be used to fit various common signal detection models to ROC data obtained from confidence rating experiments. The goals for developing the ROC Toolbox were to create a tool (1) that is easy to use and easy for researchers to implement with their own data, (2) that can flexibly define models based on varying study parameters, such as the number of response options (e.g., confidence ratings) and experimental conditions, and (3) that provides optimal routines (e.g., Maximum Likelihood estimation) to obtain parameter estimates and numerous goodness-of-fit measures. The ROC toolbox allows for various different confidence scales and currently includes the models commonly used in recognition memory and perception: (1) the unequal variance signal detection (UVSD) model, (2) the dual process signal detection (DPSD)

model, and (3) the mixture signal detection (MSD) model. For each model fit to a given data set the ROC toolbox plots summary information about the best fitting model parameters and various goodness-of-fit measures. Here, we present an overview of the ROC Toolbox, illustrate how it can be used to input and analyse real data, and finish with a brief discussion on features that can be added to the toolbox.

Keywords Signal detection theory · Open source software · Memory · Perception

Signal-detection theory, and the analysis of receiver operating characteristic (ROC) curves in particular, plays an important role in understanding the processes supporting performance in different cognitive domains, such as perception (e.g., Aly & Yonelinas, 2012), working memory (e.g., Rouder et al., 2008), and episodic memory (Egan, 1958; Green & Swets, 1988; Yonelinas, 1999; for review, see Yonelinas & Parks, 2007). Moreover, ROC analysis has been used to shed light on the cognitive process affected in numerous clinical populations, such as individuals with medial temporal lobe damage (Bowles et al., 2007, 2010; Yonelinas et al., 2002), Alzheimer's disease (Ally, Gold, & Budson, 2009; for review, see Koen & Yonelinas, 2014), and schizophrenia (Libby, Yonelinas, Ranganath, & Ragland, 2013). Indeed, the use of ROC analysis is very popular in many different areas of cognitive psychology. However, there is no standard analysis package to fit different signal detection models to ROC data. This paper introduces the ROC Toolbox, which was developed to address this gap in the field and provide a standardized framework for the analysis of ROC data.

The ROC Toolbox is an open source code designed to fit different models to ROCs derived from confidence ratings. The current release of the toolbox can be downloaded at https://github.com/jdkoen/roc_toolbox/releases. The ROC

Electronic supplementary material The online version of this article (doi:10.3758/s13428-016-0796-z) contains supplementary material, which is available to authorized users.

✉ Joshua D. Koen
joshua.koen@utdallas.edu

¹ Center for Vital Longevity, University of Texas at Dallas, Dallas, TX, USA

² Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA

³ Department of Psychology, University of California, Davis, CA, USA

⁴ Center for Mind and Brain, University of California, Davis, CA, USA

toolbox was written in the Matlab programming language because of the widespread use of the Matlab in both behavioral and neuroscience research. Importantly, Matlab is a user-friendly programming language that does not require extensive computer programming expertise to use. There were three primary goals in the development of this toolbox. The first, and most important, goal was to develop code that can flexibly define ROC models across numerous experimental designs. The second goal was to develop a toolbox that uses widely accepted methods, such as maximum likelihood estimation (Myung, 2003), for fitting models to ROC data that are transparent to researchers and controllable by the end-user. The third goal was to develop the toolbox so that it can be expanded in the future to include new models, goodness-of-fit measures, and other advancements related to ROC analysis.

The remainder of this paper will outline the features of the ROC Toolbox in more detail. First, a brief overview of constructing ROCs is presented. Second, the signal-detection models that are currently included in the ROC Toolbox will be outlined. Third, an example that outlines how to format data for use with the ROC Toolbox, and how to define and fit a model to observed data is provided. Some additional features of the ROC toolbox are then discussed, with attention given to how the toolbox might be expanded in the future. Given that some of the models that will be described below were primarily developed to account for various recognition memory phenomena (Yonelinas & Parks, 2007), much of the descriptions and discussion are couched in the context of a simple recognition memory study.

Constructing ROCs from confidence data: A brief overview

In a simple recognition memory experiment, participants are presented with a list of items (e.g., words or pictures) to remember for a subsequent recognition test. At test, the studied items are re-presented and intermixed with new items that have not been studied, and participants are tasked with classifying studied items as old (i.e., targets) and items not from the study list as new (i.e., lures.). Although there are numerous ways for obtaining data tailored to ROC analysis, the most commonly used method is to ask participants to rate how confident they are in their memory decisions (for examples of alternative approaches, see Koen & Yonelinas, 2011; Snodgrass & Corwin, 1988). The most common method is to have participants make their old/new decisions using a 6-point confidence scale (e.g., 6 – sure old, 5 – maybe old, 4 – guess old, 3 – guess new, 2 – maybe new, 1 – sure new). The idea is that the different levels of confidence reflect different levels of response bias, or a tendency to classify a test item as old. An ROC is constructed by plotting the cumulative hit rate (i.e., correctly calling an old item “old”) against the cumulative false alarm rate (i.e., incorrectly calling a new item “old”)

for the different confidence ratings beginning with the “6–Sure Old” response (i.e., the most conservative response bias; see Fig. 1). Specifically, from left to right, the first point on the ROC plots represents the hit and false alarm rates for the most confident old response (e.g., six responses), the second point represents the hit and false alarm rates for the two highest confidence old responses (e.g., six and five responses), and so on. Typically, the highest confidence new (lure) response is not plotted on ROCs because it is constrained to have hit and false alarm rates equal to 1.0.

There are many issues to consider when designing a study for ROC analysis, including the number of target and lure trials, the number of bins in the confidence scale, and the model one wishes to fit to the data. These issues are beyond the scope of the present paper. We refer readers to Yonelinas and Parks (2007) in which these issues are discussed.

Signal-detection models

It is generally accepted that the shape of an observed ROC curve provides information about the cognitive processes that support performance on a particular task. To this end, many different models have been put forth to account for data from various paradigms in cognitive psychology (e.g., recognition memory). These models range from simple threshold and continuous models to more complex models in which multiple

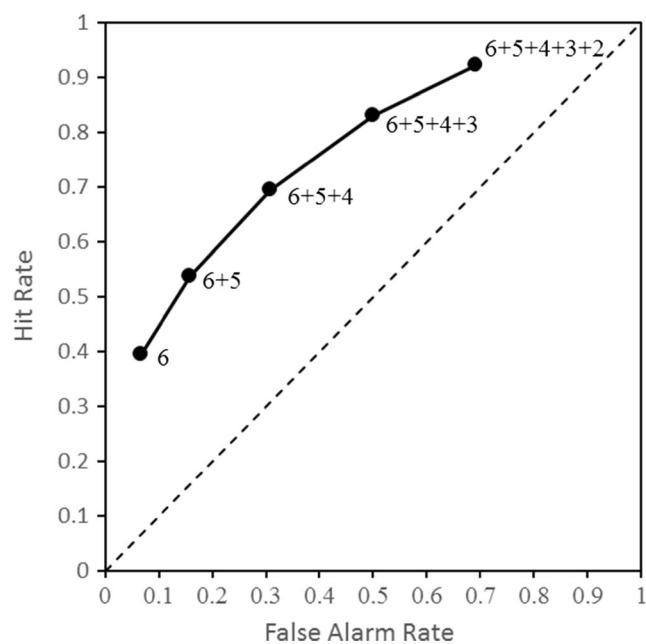


Fig. 1 An example receiver-operating characteristic (ROC) curve derived from standard a 6-point confidence scale where a 6 and 1 response correspond to Sure Target and Sure Lure judgments, respectively. Successive points on the ROC are, starting with the highest confidence decision for targets, the cumulative hit rate plotted against the cumulative false alarm rate

distributions are mixed together (for a detailed description of many different models, see Yonelinas & Parks, 2007). The initial release of the ROC Toolbox includes three of the most commonly used signal detection models: the unequal-variance signal detection (UVSD) model (Egan, 1958; Mickes, Wixted, & Wais, 2007), the dual-process signal detection (DPSD) model (Yonelinas, 1994, 1999), and the mixture signal detection (MSD) model (DeCarlo, 2002, 2003; Onyper, Zhang, & Howard, 2010).

Before outlining the parameterization of each model, there are several issues that bear mention. First, for consistency, the parameter notation utilized in this report and in the ROC Toolbox is consistent with the notation that has been used in the extant literature. In some cases, the notation used to describe the parameters of a model are theoretically loaded. For example, two of the parameters of the DPSD model have been dubbed the recollection and familiarity parameters, which implies that the parameters directly measure the theoretical constructs of recollection and familiarity. However, the parameters of a given signal-detection model are purely a mathematical description of the shape of the (best-fitting) ROC to the data. The value of the best-fitting parameters does not provide confirmation that a particular psychological process does or does not contribute to performance on a given task. Whether a given parameter actually reflects a given theoretical process can at best only be inferred by measuring how the parameter behaves across different experimental conditions (DeCarlo, 2002; e.g., DeCarlo, 2003; Harlow & Donaldson, 2013; for a detailed discussion of this issue, see Koen, Aly, Wang, & Yonelinas, 2013).

Second, the ROC Toolbox defines the UVSD, DPSD, and MSD in their most general form. This provides the greatest flexibility with respect to the models a researcher might want to fit to their data. By defining models in the most general form, this greatly increases the number of models that are available to researchers because any model nested within a more complex model can be fit to a given data set with the application of simple parameter constraints.

The following sections present the mathematical functions of the the UVSD, DPSD, and MSD models. Moreover, the sections below focus on the parameters that are specific to a given model. The parameters for the response criteria – the rate at which individuals classify items as belonging to a target category – are always included in any signal detection model that is fit to ROC data. In a typical model with N different rating bins, there are $N-1$ criteria parameters for each pair of target and lure trials. The parameters for the response criteria are designated as c_i , where i represents the i^{th} rating bin.

The UVSD model

The unequal-variance signal-detection (UVSD) model proposes that the memory strength of target and lure items falls

on a continuum with the target distribution having a higher average strength compared to the lure distribution. The lure distribution is represented by a standard normal distribution with a mean of 0 and a unit standard deviation. The free parameters in this model are the mean (d') and standard deviation (σ) of the target distribution. Specifically, the UVSD model predicts the cumulative proportion of target and lure trials in each rating bin (B_i) using the following formulas:

$$\begin{aligned} p(\text{Target}|B_i) &= \Phi\left[\frac{d'-c_i}{\sigma}\right] \\ p(\text{Lure}|B_i) &= \Phi[-c_i] \end{aligned}$$

The Φ denotes a cumulative normal distribution. The left side of the equations represents the predicted cumulative proportion of target and lure items, respectively, that received a rating greater than or equal to the i^{th} rating bin. This notation is also used in the models reported below. Moreover, the cumulative proportion of trials for both targets and lures in the last rating bin (which corresponds to the highest confidence lure response) is always 1. The UVSD model can be reduced to a simpler equal-variance signal-detection (EVSD) model by constraining σ to unity.

The DPSD model

The dual-process signal detection (DPSD) model is derived from dual-process theories of recognition memory that propose recognition memory is supported by two qualitatively distinct processes – recollection and familiarity (for reviews, see Yonelinas, 2002; Yonelinas, Aly, Wang, & Koen, 2010). The DPSD model in its most complex form is characterized as a mixture of a two-high threshold model and a UVSD model. The two-high threshold component of the DPSD model defines a target threshold parameter being labeled recollection of “oldness” (R_o) and the lure threshold parameter being labeled recollection of “newness” (R_n). The UVSD component of the model – typically labeled the familiarity component of the DPSD model – comprises a parameters for the mean (d'_F) and standard deviation (σ_F) of the familiarity distribution. The DPSD model predicts the cumulative proportion of target and lure trials in each rating bin (B_i) using the following formulas:

$$\begin{aligned} p(\text{Target}|B_i) &= R_o + (1-R_o)\Phi\left[\frac{d'_F-c_i}{\sigma_F}\right] \\ p(\text{Lure}|B_i) &= (1-R_n)\Phi[-c_i] \end{aligned}$$

For target items, the idea is that an item with a strength above the R_o threshold will always be classified as a target, whereas classification of a target item falling below the R_o threshold is determined by the familiarity component of the model. Lure items that have a strength above the R_n threshold

are always classified as new. When a lure item's strength falls below the R_n threshold, classification of new items is governed by a standard normal distribution. The free parameters in the full model are R_o , R_n , d'_F , and σ_F . Although the σ_F parameter is allowed to vary in the ROC Toolbox, in the most common formulation of the DPSD model the σ_F parameter is constrained to equal 1 (equivalent to an equal-variance signal detection, or EVSD, model).

The MSD model

The mixture signal-detection (MSD) model proposes that classification of target and lure items on a recognition test are best modeled as a mixture of different Gaussian distributions. There are numerous theoretical proposals for why targets and lures could be modeled with a mixture of two distributions, but an in depth discussion is beyond the scope of this paper (for discussion, see DeCarlo, 2002, 2003; Harlow & Donaldson, 2013; Koen et al., 2013; Onyper et al., 2010). The ROC toolbox incorporates an expanded version of the models outlined by DeCarlo (2003) and Onyper and colleagues (2010). The MSD predicts the cumulative proportion of target and lure trials in each rating bin (B_i) using the following formulas:

$$p(\text{Target}|B_i) = \lambda_T \Phi \left[\frac{(d'_{\lambda_T} + d'_{(1-\lambda_T)}) - c_i}{\sigma_{\lambda_T}} \right] + (1 - \lambda_T) \Phi \left[\frac{d'_{(1-\lambda_T)} - c_i}{\sigma_{(1-\lambda_T)}} \right]$$

$$p(\text{Lure}|B_i) = \lambda_L \Phi \left[\frac{-d'_{\lambda_L} - c_i}{\sigma_{\lambda_L}} \right] + (1 - \lambda_L) \Phi[-c_i]$$

Note that the subscript for the d' and σ parameters correspond to whether the distribution falls above (i.e., λ_T) or below (i.e., $1 - \lambda_T$) a threshold. Additionally, a T subscript indicates a parameter specific to the distribution for target items whereas a L subscript indicates a parameter specific to the distribution for lure items.

In the above model, the distribution representing target items is modeled as a mixture of two Gaussian distributions with (potentially) different means and standard deviations. Specifically, target items that fall above the threshold for the target distribution (λ_T) are sampled from a Gaussian distribution with a mean that is d'_{λ_T} units larger than the mean of the Gaussian distribution that falls below the target threshold ($d'_{(1-\lambda_T)}$). The variances of the two distributions are σ_{λ_T} and $\sigma_{(1-\lambda_T)}$ for the distribution falling above and below λ_T , respectively.

Similarly, lure items are modeled as a mixture of two distributions. The distribution falling below the lure item threshold λ_L (i.e., $1 - \lambda_L$) is a standard normal distribution. The distribution above the lure item threshold λ_L has a mean and standard deviation of d'_{λ_L} and σ_{λ_L} , respectively. Note that the mean for the above mentioned lure item distribution is

negative, which simply reflects the fact that this distribution is measuring the average strength that an item is a lure (i.e., not a target). The potential free parameters of the MSD model defined above are λ_T , d'_{λ_T} , σ_{λ_T} , $d'_{(1-\lambda_T)}$, $\sigma_{(1-\lambda_T)}$, λ_L , and d'_{λ_L} . Typically, only a subset of these parameters will be relevant to a given experimental question. For example, the standard model fit to data from the simple item recognition we outlined previously only estimates λ_T and d'_{λ_T} , and the remaining parameters are set to fixed values (which are outlined in the manual).

Fitting models to data

This section briefly illustrates how to format data for use with the ROC Toolbox and how to define and fit models to observed data. The example response frequency data used in the example outlined below is in Table 1. This data is from a single participant in an unpublished study conducted by Koen and Yonelinas. Participants in this experiment studied an intermixed list of low and high frequency words for a subsequent recognition memory test. At test, participants discriminated studied words (targets) from new words (lures) using a 20-point confidence scale with the following anchor points: 20 – Sure Old, 11 – Guess Old, 10 – Guess New, 1 – Sure New.

In this example, the most commonly applied form of DPSD model is fit to the example data set with two conditions: high versus low frequency targets and lures. For each condition, separate R_o and d'_F parameters, but not the R_n and σ_F , are estimated.¹ The R_o was allowed to take any value between 0 and 1. There were no bound constraints on the d'_F . The raw data and Matlab code for this example are include as a Supplemental Material and as a file in the ROC Toolbox (https://github.com/jdkoen/roc_toolbox/blob/master/examples/BRM_paper_supp_material.m).² Any reference to variables or functions used in the example code appear in Courier New.

¹ The R_n and σ_F are constrained to values of 0 and 1, respectively, which conforms to the most ubiquitous formulation of the DPSD model (e.g., Yonelinas, 1999).

² The purpose of using data from a 20-point confidence scale in this example with two different conditions is intended to show the flexibility of the toolbox. The ROC Toolbox is (in theory) able to handle any number of different repeated-measures conditions and any number of rating bins ranging from , for instance, a standard 6-point scale to a more continuous, fine-grained confidence scale (e.g., Harlow & Donaldson, 2013; Harlow & Yonelinas, 2016). To illustrate this, those interested can combine the confidence bins in different ways, such as summing frequencies across bins to produce 10 bins or 6 bins (e.g., Mickes et al., 2007), or fit the model to only one condition. Importantly, the example code included in the Supplemental Material will fit the correct version of the DPSD model outlined in the main text by simply changing the format of the input data.

Table 1 Example response frequencies from a single participant in an unpublished study by Koen and Yonelinas

Condition	Confidence bin																			
	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Low Frequency																				
Targets	77	18	13	8	4	3	5	4	8	7	4	5	3	7	1	5	5	7	11	5
Lures	3	2	8	4	4	3	1	4	11	8	15	15	6	7	6	11	14	28	33	17
High Frequency																				
Targets	57	25	13	9	9	3	7	8	13	11	8	11	5	2	4	3	5	6	0	1
Lures	5	9	6	8	4	5	7	4	11	12	20	13	10	7	8	5	15	17	28	6

Note. The confidence bins were labeled as follows: 20 – Sure Old, 11 – Guess Old, 10 – Guess New, 1 – Sure New

Step 1: Formatting data

The ROC Toolbox requires as data the response frequencies in each rating (i.e., confidence) bin separately for target and lure items in each condition. Specifically, two separate matrices – one for the response frequencies to the target items and another lure items – are required for the ROC Toolbox. Each matrix must comprise C columns and R rows. The columns of each matrix represent the different rating bins. The bins must be ordered in a descending fashion such that the rating bin associated with the highest confidence decision for a target item (e.g., 20 – Sure Old) is in the first column and the rating bin associated with the highest confidence lure response is in the last column (e.g., 1 – Sure New). The rows of each matrix represents different within-participant experimental conditions. Before continuing the discussion of the response frequency matrices, we first define the data shown in Table 1 in the appropriate format. The matrix for the response frequencies of each rating bin to target items (the `targf` variable in the [Supplemental Material](#)) is:

$$\begin{bmatrix} 77 & 18 & 13 & 8 & 4 & 3 & 5 & 4 & 8 & 7 & 4 & 5 & 3 & 7 & 1 & 5 & 5 & 7 & 11 & 5 \\ 57 & 25 & 13 & 9 & 9 & 3 & 7 & 8 & 13 & 11 & 8 & 11 & 5 & 2 & 4 & 3 & 5 & 6 & 0 & 1 \end{bmatrix}$$

The corresponding matrix for lure items (the `luref` variable in the [Supplemental Material](#)) is:

$$\begin{bmatrix} 3 & 2 & 8 & 4 & 4 & 3 & 1 & 4 & 11 & 8 & 15 & 15 & 6 & 7 & 6 & 11 & 14 & 28 & 33 & 17 \\ 5 & 9 & 6 & 8 & 4 & 5 & 7 & 4 & 11 & 12 & 20 & 13 & 10 & 7 & 8 & 5 & 15 & 17 & 28 & 6 \end{bmatrix}$$

There are some important issues to consider when formatting data for the ROC Toolbox. First, great care must be taken to format the data correctly as the ROC Toolbox makes assumptions about how the data are organized, but will still provide output, albeit incorrect output, if these assumptions are violated. In addition to the assumption that the data are formatted such that the columns are ordered in the fashion

described above, the ROC Toolbox also assumes that the response frequencies for the target and lure trials from the same condition are in the same row in the target and lure matrices. In the example above, the response frequencies for the low frequency targets and lures are in the first row of the respective target and lure matrices, whereas the high frequency targets and lures are in the second row. Second, the matrices for target and lure items must be the same size else the functions will return an error. Thus, rating bins that are not used for one trial type (e.g., targets) cannot be omitted from the response frequency matrix for targets. Additionally, the target and lure matrices must be the same size even in experimental designs with multiple conditions for target trials relative to a single condition for lure items (e.g., deep vs. shallow processing at encoding).³ To help ensure the data are correctly formatted, the ROC Toolbox includes a function to import data from CSV or text files (`roc_import_data`).

Step 2: Defining the model

The ROC Toolbox requires four pieces of information to define a model to fit to the data: the number of rating bins, the number of different conditions, the type of model to fit to the data (i.e., UVSD, DPSD, MSD), and the model-specific parameters that are allowed to freely vary. This information is passed to the a function called `gen_pars`, which is the primary function used to define the model to fit to the data. The input for number of rating bins and conditions is simply the number of columns and rows of the matrix for target (or lure) items (e.g., the `targf` or `luref` variables in the example script; see [Supplemental Material](#)). The number of rating bins define the number of criteria parameters to estimate. In experiments collecting confidence ratings, this should always be one less than the total number of confidence bins for each within-participant

³ The ROC Toolbox is able to handle these designs with additional options to the `roc_solver` function. Using the ROC Toolbox with this type of experimental design is detailed in the manual.

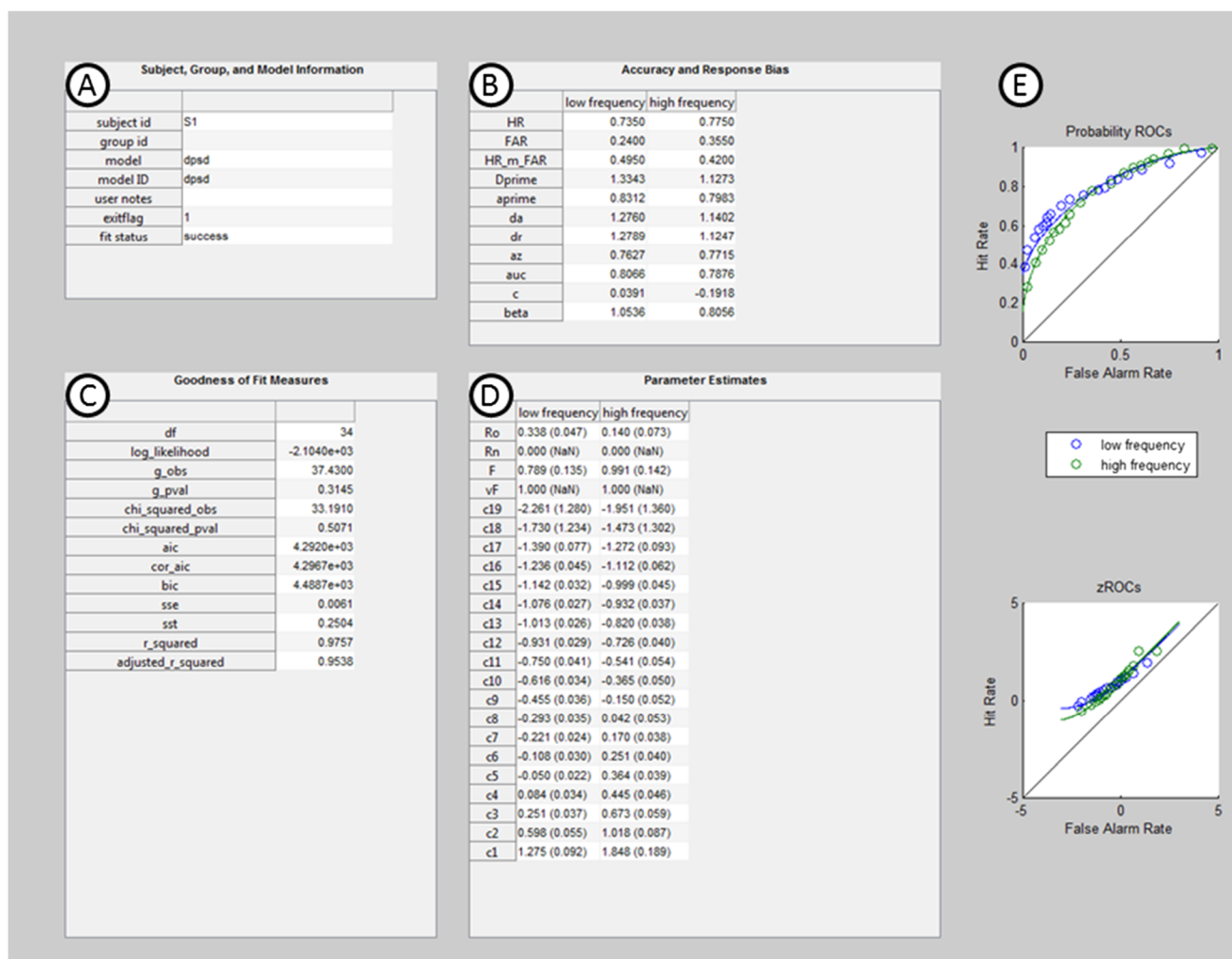


Fig. 2 Example of the summary information output by the `roc_solver` function from fitting the DPSD model to the example data set described in the text (see also [Supplemental Material](#)). The output includes information about the participant, model, and exit status of the optimization routine (a), measures of accuracy/discrimination and response bias (b), goodness-of-fit and model selection measures (c), the

best-fitting parameter estimates and the standard error of the parameter estimates (d), and the observed and model-predicted ROCs and zROCs (e). The standard errors were estimated from 1,000 iterations of the non-parametric bootstrap procedure described in the main text. The NaN values are given for parameters that were not estimated in the model

condition. In this example there are a total of 38 response criteria parameters (19 for each condition). The model specific parameters that are allowed to vary are specified as a cell array of strings corresponding to the labels for the parameters, which in this example are 'Ro' and 'F' for the R_o and d'_F parameters of the DPSD model. Importantly, parameters that are not specified as free parameters are constrained to equal a pre-determined value. The pre-determined values are documented in the manual, and can also be changed by the user.

The `gen_pars` function outputs three matrices: (1) a matrix for the starting values of the optimization routine (x_0), (2) a matrix for the lower bound each parameter can take (LB), and (3) a matrix for the upper bound of each parameter (UB). The latter two matrices are the ones that define which parameters are free, and which parameters are constrained to equal a set

value.⁴ Specifically, parameters are constrained (not estimated) when the lower and upper bounds of parameter have the same value. Note that if not estimated, the parameter takes the value of that is in the matrix defining the upper (and lower) bounds of the parameters. For example, to constrain the R_n parameter to equal 0, the lower and upper bounds should both equal 0.

Step 3: Fitting the model to the data

The `roc_solver` function fits the model defined by the x_0 , LB, and UB matrices to the data defined in `target` and `luref` using either maximum likelihood estimation (Myung, 2003) or

⁴ Additional parameter constraints, such as equality constraints, can also be defined. This is described in detail in the manual.

by minimizing the sum-of-squared errors. The latter approach minimizes the squared differences between the observed and predicted (non-cumulative) proportion of trials in each rating bin. Optimization is carried out using the interior point algorithm (Byrd, Gilbert, & Nocedal, 2000; Byrd, Hribar, & Nocedal, 1999; Byrd, Schnabel, & Shultz, 1988) implemented in the `fmincon` function of Matlab's Optimization Toolbox (<http://www.mathworks.com/help/optim/ug/fmincon.html>). The `roc_solver` function stores data in a structure variable that contains the best fitting parameter values, transformations of the observed frequencies (e.g., cumulative proportions), goodness-of-fit measures, and standard signal detection measures of accuracy and response bias, and also generates a summary plot. The summary plot from the analysis of the example data, which can be reproduced with the example code in the [Supplemental Material](#), is shown in Fig. 2. As can be seen in Fig. 2a-c, the `roc_solver` function outputs numerous accuracy, response bias, and goodness-of-fit measures (see below). Figure 2d lists the best fitting parameter estimates as well as the standard error of each parameter estimate (see below). Figure 2e plots the observed and model predicted ROC and z-transformed ROC (zROC), respectively.

The standard errors are estimated with a non-parametric bootstrap procedure whereby individual target and lure trials are randomly sampled with replacement from the observed data to create a new sample of target and lure trials. The model fit to the observed data is also fit to each of the randomly sampled sets of target and lure trials. The standard error values in Fig. 2 were estimated from 1,000 iterations of the non-parametric bootstrap sampling procedure.⁵ The number of iterations used to estimate the standard error of the parameter estimates can be controlled by the end-user.

An important feature of the ROC Toolbox is that numerous different models that are fit to the same data can be stored in a single data structure (output by the `roc_solver` function). Moreover, there are many additional options for the `roc_solver` function, some of which are used in the example code in the [Supplemental Material](#), that are detailed in the documentation of the ROC Toolbox.

Accuracy, response, goodness-of-fit and model selection measures

The ROC Toolbox outputs numerous measures that index accuracy/discrimination and response bias provided by Macmillan and Creelman (2005), as well as numerous goodness-of-fit measures to evaluate model fit and perform

⁵ The amount of time necessary to run the non-parametric bootstrap procedure can take quite long for a single model, and will depend in part on the complexity of the model and the number of iterations. For this reason, the non-parametric routine to estimate standard errors is not run by default, but can simply be called using the `bootIter` property-value input of the `roc_solver` function.

model selection. The measures of accuracy and discrimination include the hit and false alarm rates, the hit minus false alarm measure, d' , A' , d_a , A_z , and area under the curve (AUC). Additionally, we include the d_r measure described in Mickes et al. (2007), which is similar to the d_a measure. The response bias measures include c and β . The fit and model selection measures included are the log-likelihood value, regular and adjusted R^2 , χ^2 , G (Sokal & Rohlf, 1994), the Akaike Information Criterion (AIC; Akaike, 1974), AICc (Burnham & Anderson, 2004), and Bayesian Information Criterion (Schwarz, 1978).

Concluding remarks and future directions

The ROC Toolbox provides a large set of tools that analyze ROC data. Importantly, the ROC Toolbox can flexibly define statistical models to fit to ROC data derived from many different types of experimental designs. Individual models and goodness-of-fit measures are defined as separate functions within the ROC Toolbox, and these are utilized by generic model-fitting wrappers (e.g., `gen_pars`, `roc_solver`). Thus, the ROC toolbox can easily be expanded with new models and goodness-of-fit measures. Additional future directions for the ROC Toolbox include adding functionality to fit ROCs derived from non-ratings based methods, such as manipulation of response bias with reward (e.g., Fortin, Wright, & Eichenbaum, 2004; Koen & Yonelinas, 2011), and incorporating Bayesian estimation of the models with the ability to flexibly define different priors for each parameter. Our hope is that researchers both with and without experience in ROC analysis will find this toolbox useful in their own research endeavors.

Author Note This work was supported by National Science Foundation Graduate Research Fellowship 1148897 awarded to Joshua D. Koen and by National Institute of Mental Health Grants R01-MH059352-13 and R01-MH083734-05 awarded to Andrew P. Yonelinas. JDK was supported by a Ruth L. Kirschstein National Research Service Award from the National Institute on Aging (F32-AG049583) during the preparation of this manuscript.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi:10.1109/TAC.1974.1100705
- Ally, B. A., Gold, C. A., & Budson, A. E. (2009). An evaluation of recollection and familiarity in Alzheimer's disease and mild cognitive impairment using receiver operating characteristics. *Brain and Cognition*, *69*(3), 504–513. doi:10.1016/j.bandc.2008.11.003
- Aly, M., & Yonelinas, A. P. (2012). Bridging consciousness and cognition in memory and perception: Evidence for both state and strength processes. *PloS One*, *7*(1), e30231. doi:10.1371/journal.pone.0030231

- Bowles, B., Crupi, C., Mirsattari, S. M., Pigott, S. E., Parrent, A. G., Pruessner, J. C., ... Köhler, S. (2007). Impaired familiarity with preserved recollection after anterior temporal-lobe resection that spares the hippocampus. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(41), 16382–16387. doi:10.1073/pnas.0705273104
- Bowles, B., Crupi, C., Pigott, S. E., Parrent, A., Wiebe, S., Janzen, L., & Köhler, S. (2010). Double dissociation of selective recollection and familiarity impairments following two different surgical treatments for temporal-lobe epilepsy. *Neuropsychologia*, *48*(9), 2640–2647. doi:10.1016/j.neuropsychologia.2010.05.010
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304. doi:10.1177/0049124104268644
- Byrd, H. R., Gilbert, C. J., & Nocedal, J. (2000). A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming*, *89*(1), 149–185. doi:10.1007/spl00011391
- Byrd, H. R., Hribar, M. E., & Nocedal, J. (1999). An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, *9*(4), 877–900. doi:10.1137/S1052623497325107
- Byrd, H. R., Schnabel, R. B., & Shultz, G. A. (1988). Approximate solution of the trust region problem by minimization over two-dimensional subspaces. *Mathematical Programming*, *40*(1), 247–263. doi:10.1007/bf01580735
- DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *109*(4), 710–721. doi:10.1037/0033-295X.109.4.710
- DeCarlo, L. T. (2003). An application of signal detection theory with finite mixture distributions to source discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 767–778. doi:10.1037/0278-7393.29.5.767
- Egan, J. P. (1958). *Recognition memory and the operating characteristic (U. S. Air Force Operational Applications Laboratory Technical Note Nos. 58, 51, 32)*. Retrieved from Bloomington, IN.
- Fortin, N. J., Wright, S. P., & Eichenbaum, H. (2004). Recollection-like memory retrieval in rats is dependent on the hippocampus. *Nature*, *431*(7005), 188–191. doi:10.1038/nature02853
- Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics*. Los Altos: Peninsula Publishing.
- Harlow, I. M., & Donaldson, D. I. (2013). Source accuracy data reveal the thresholded nature of human episodic memory. *Psychonomic Bulletin & Review*, *20*(2), 318–325. doi:10.3758/s13423-012-0340-9
- Harlow, I. M., & Yonelinas, A. P. (2016). Distinguishing between the success and precision of recollection. *Memory*, *24*(1), 114–127. doi:10.1080/09658211.2014.988162
- 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. doi:10.1037/a0033671
- Koen, J. D., & Yonelinas, A. P. (2011). From humans to rats and back again: Bridging the divide between human and animal studies of recognition memory with receiver operating characteristics. *Learning & Memory*, *18*(8), 519–522. doi:10.1101/lm.2214511
- Koen, J. D., & Yonelinas, A. P. (2014). The effects of healthy aging, amnesic mild cognitive impairment, and Alzheimer's disease on recollection and familiarity: A meta-analytic review. *Neuropsychology Review*, *24*(3), 332–354. doi:10.1007/s11065-014-9266-5
- Libby, L. A., Yonelinas, A. P., Ranganath, C., & Ragland, J. D. (2013). Recollection and familiarity in schizophrenia: A quantitative review. *Biological Psychiatry*, *73*(10), 944–950. doi:10.1016/j.biopsych.2012.10.027
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A User's guide* (2nd ed.). New York: Lawrence Erlbaum Associates.
- 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. doi:10.3758/BF03194112
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100. doi:10.1016/S0022-2496(02)00028-7
- Onyper, S. V., Zhang, Y. X., & Howard, M. W. (2010). Some-or-none recollection: Evidence from item and source memory. *Journal of Experimental Psychology: General*, *139*(2), 341–364. doi:10.1037/a0018926
- Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences*, *105*(16), 5975–5979. doi:10.1073/pnas.0711295105
- Schwarz, G. (1978). Estimating the Dimension of a Model. 461–464. doi:10.1214/aos/1176344136
- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50.
- Sokal, R. R., & Rohlf, F. J. (1994). *Biometry: The principles and practices of statistics in biological research* (3rd ed.). New York: W.H. Freeman and Company.
- Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1341–1354. doi:10.1037/0278-7393.20.6.1341
- Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1415–1434. doi:10.1037/0278-7393.25.6.1415
- Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517. doi:10.1006/jmla.2002.2864
- Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus*, *20*(11), 1178–1194. doi:10.1002/hipo.20864
- Yonelinas, A. P., Kroll, N. E. A., Quamme, J. R., Lazzara, M. M., Sauvé, M.-J., Widaman, K. F., & Knight, R. T. (2002). Effects of extensive temporal lobe damage or mild hypoxia on recollection and familiarity. *Nature Neuroscience*, *5*(11), 1236–1241. doi:10.1038/nn961
- Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800–832. doi:10.1037/0033-2909.133.5.800