Neuropsychologia 52 (2014) 51–56

Contents lists available at ScienceDirect

Neuropsychologia

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

A novel approach to an old problem: Analysis of systematic errors in two models of recognition memory

Adam J.O. Dede [a,b,*], Larry R. Squire [a,b,c,d], John T. Wixted [b]

[a] Veterans Affairs San Diego Healthcare System, San Diego, CA 92161, USA
[b] Department of Psychology, University of California, San Diego, CA 92093, USA
[c] Departments of Psychiatry University of California, San Diego, CA 92093, USA
[d] Department of Neurosciences, University of California, San Diego, CA 92093, USA

ARTICLE INFO

Article history:
Received 16 July 2013
Received in revised form
30 September 2013
Accepted 22 October 2013
Available online 29 October 2013

Keywords:
HTDP model
CDP model
Confidence ratings
Medial temporal lobe
Hippocampus

ABSTRACT

For more than a decade, the high threshold dual process (HTDP) model has served as a guide for studying the functional neuroanatomy of recognition memory. The HTDP model's utility has been that it provides quantitative estimates of recollection and familiarity, two processes thought to support recognition ability. Important support for the model has been the observation that it fits experimental data well. The continuous dual process (CDP) model also fits experimental data well. However, this model does not provide quantitative estimates of recollection and familiarity, making it less immediately useful for illuminating the functional neuroanatomy of recognition memory. These two models are incompatible and cannot both be correct, and an alternative method of model comparison is needed. We tested for systematic errors in each model's ability to fit recognition memory data from four independent data sets from three different laboratories. Across participants and across data sets, the HTDP model (but not the CDP model) exhibited systematic error. In addition, the pattern of errors exhibited by the HTDP model was predicted by the CDP model. We conclude that the CDP model provides a better account of recognition memory than the HTDP model.

1. Introduction

Dual-process theorists hold that recognition memory depends on two components: familiarity and recollection. Familiarity involves knowing only that an item is old or new, and recollection involves accessing specific details about the episode in which the item was encountered. The relative contribution of these two processes to individual recognition decisions is debated. On one hand, the recognition decision for a particular item may be based on one process or the other, varying from one decision to the next. On the other, the recognition decision for a particular item may be based on both familiarity and recollection. These possibilities are formalized in two models that have been used to characterize recognition memory function, the high-threshold dual-process model (HTDP; Yonelinas, 1994; Yonelinas, 1999) and the continuous dual-process model (CDP; Wixted & Mickes, 2010). In many cases, the CDP model is mathematically equivalent to the single process unequal variance signal detection (UVSD) model (Wixted & Mickes, 2010). However, because of the large body of evidence indicating the existence of separate processes in recognition memory (Diana, Reder, Arndt, & Park, 2006), we focus on the dual process interpretation of the UVSD model (namely, the CDP model).

The HTDP model provides quantitative estimates of familiarity and recollection from confidence ratings made on a standard old/new recognition task, but the CDP model holds that recollection and familiarity cannot be disentangled on the basis of old/new recognition decisions alone. The HTDP model's ability to quantify recollection and familiarity may explain the notable role it has played in guiding investigations of the neural basis of recognition memory. However, it is important to consider that the HTDP model's ability to make these estimates and the CDP model's corresponding inability are derived from the assumptions made by the two models about recognition. If the assumptions that a model makes about recognition memory are accurate, then, when it is fit to recognition data, the only source of error in the fit should be randomly distributed noise. However, if the assumptions that a model makes about recognition memory are inaccurate, then errors in the model's ability to fit data are likely to be systematic (even if the model provides a good fit to the data). Here, we investigate whether the HTDP model or the CDP model produces systematic errors, that is, deviations from what is observed in recognition memory data.

* Corresponding author at: Department of Psychology, University of California, San Diego, CA 92093, USA. Tel.: +1 858 642 3628.
E-mail address: adam.osman.dede@gmail.com (A.J.O. Dede).

0028-3932/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neuropsychologia.2013.10.012

The assumptions of the HTDP model differ from the CDP model in two important respects. First, the HTDP model assumes that recollection is a high-threshold process (Macmillan & Creelman, 2005; Yonelinas, 1994; Yonelinas, 1999), such that recollection is either successful (yielding recognition decisions made with high confidence and high accuracy) or unsuccessful. The CDP model (Wixted, 2007; Wixted & Mickes, 2010), by contrast, assumes that recollection can vary continuously (yielding recognition decisions made with a wide range of confidence and accuracy).

A second difference between the two models follows from the HTDP model's assumption that recollection is a high-threshold process. The HTDP model predicts that if recollection is successful, then familiarity does not contribute to the recognition decision because recollection provides unambiguous evidence of a previous encounter. If recollection is unsuccessful, then the recognition decision is based wholly on the strength of the familiarity signal. By contrast, the CDP model posits that familiarity and recollection are combined during recognition memory decision-making. This feature of the model arises from the proposition that both recollection and familiarity are assumed to be imperfect continuous processes, and combining them can yield a more diagnostic memory signal than relying on either one alone.

A number of studies have compared the CDP model to the HTDP model using receiver operating characteristic (ROC) analysis, a technique based on confidence ratings that allows model-based inferences about the nature of the underlying memory-strength distributions across items (Macmillan & Creelman, 2005). The validity of model-based inferences is typically assessed in ROC analysis by comparing a model's fit to the observed data in order to calculate a goodness-of-fit statistic. Although there have been many studies (e.g. Glanzer, Kim, Hilford, & Adams 1999; Glanzer, Hilford, & Kim, 2004; Healy, Light, & Chung, 2005; Heathcote, 2003; Kelley & Wixted, 2001; Slotnick & Dodson, 2005; Yonelinas, 1994), the evidence based on goodness-of-fit statistics alone has been mixed, with some studies favoring the HTDP model and some favoring the CDP model. The CDP model often provides the better fit to typical recognition memory data (e.g. Slotnick & Dodson, 2005; for review see Wixted, 2007), but some results are better accounted for by the HTDP model (e.g. Howard, Bessette-Symons, Zhang, & Hoyer, 2006; Yonelinas, 1999).

One reason for these inconclusive results may be that both models are able to fit recognition memory data quite well. Indeed, an earlier analysis of a typical data set found that the HTDP model accounted for 99.91% of the variance, and the CDP model accounted for 99.97% of the variance (Glanzer, et al., 1999; Yonelinas, 1999b). Similarly, across four data sets analyzed below, which involved 65 participants, the average percent of variance accounted for was above 90% for both models (HTDP=91%, CDP=96%). The fact that both models fit the data well may explain why both are given credence despite their fundamental differences.

The assumption is often made that models that fit data well are good models. However, this assumption is not necessarily valid (Roberts & Pashler, 2000). Accordingly, model comparisons based on goodness-of-fit may have difficulty deciding which model is best. An alternative, more promising, way to distinguish between the merits of the two models is to first ask whether the models generate systematic errors in their ability to account for recognition memory data. Second, if a model generates systematic errors, then one can ask whether the other model, in fact, predicts these errors.

Fig. 1 illustrates the essential differences between the two models. The models make the same assumptions about the distractor distribution (i.e., the distribution of memory strength signals generated by the foils), but they differ in their assumptions about the target distribution (i.e., the distribution of memory strength signals generated by the targets). The HTDP model has two target distributions, a high-threshold distribution for items that are recollected (these targets have essentially infinite memory strength) and a separate continuous distribution for items judged on the basis of familiarity. The familiarity and distractor distributions are assumed to have equal variance. In contrast, the CDP model has a single target distribution, and that distribution has greater variance than the distractor distribution.

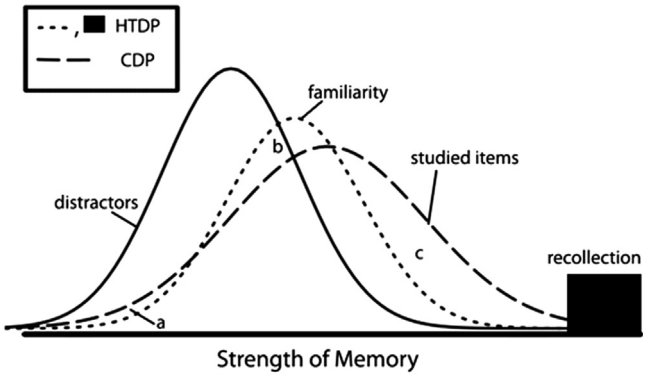


Fig. 1. Schematic representation of the theoretical distributions of items in memory according to both the high threshold dual process model (HTDP; dotted lines) and the continuous dual process model (CDP; dashed lines). The two models share a distribution for new items (distractors; solid line). The HTDP model has separate distributions for study items supported by recollection and familiarity. The CDP model has a single distribution for study items. The *X*-axis represents the strength of memory, proceeding from low memory strength at the left to high memory strength at the right. Areas a, b, and c show areas of non-overlap between the two models where the predicted data differ systematically.

Moving from low memory strength (Fig. 1; left) to high memory strength (Fig. 1; right), visual inspection of the models' target distributions reveals areas where the models do not overlap and where the predicted data differ systematically. At low levels of memory strength (area (a)), the HTDP model predicts a lower frequency of target items than does the CDP model. At medium levels of memory strength (b), the HTDP model predicts a higher frequency of target items than does the CDP model. At moderately high levels of memory strength (c), the HTDP model predicts a lower frequency of target items than does the CDP model. Lastly, at the highest levels of memory strength (represented in the HTDP model by the distribution of recollection responses and in the CDP model by the rightmost tail of the target distribution), the HTDP model predicts a higher frequency of target items than does the CDP model. Thus, if the assumptions of the HTDP model are correct, then one might expect to find that the best-fitting CDP model predicts too many low-confidence responses to targets, too few medium confidence responses to targets, and too many moderately high confidence ratings to targets. If, instead, the assumptions of the CDP model are correct, then the best-fitting HTDP model should exhibit the opposite pattern of systematic error.

Note that Fig. 1 is simply an example illustrating systematic errors that might be observed for a particular set of model parameter values. We chose these parameter values because they correspond to values typically observed in recognition memory experiments. Still, the actual systematic errors could differ across individuals depending on the model parameters that characterize the performance of each individual.

To differentiate between the HTDP and CDP models, we first examined the ability of each model to fit recognition memory data in four data sets from three different laboratories and then investigated whether any systematic errors were evident in their fits to target items. Lure items were also examined but yielded no systematic errors for either model. We then tested whether the systematic errors generated by one model (if any) could be

predicted by the other model. We performed this analysis at the individual level (i.e., fitting the two models to each individual's data separately, and generating predictions of systematic error based on each participant's performance individually). We found that only the HTDP model generated systematic errors in its fit to target items. Moreover these errors were predicted by the CDP model. By contrast, the CDP model did not yield systematic errors, and the HTDP model predicted errors for the CDP model that were not observed. In other words, the predictions of the CDP model were confirmed (validating its assumptions about recognition memory), whereas the predictions of the HTDP model were disconfirmed (invalidating its assumptions about recognition memory). This pattern was observed even though both models fit the data well (as is usually true), underscoring the fact that a good fit does not necessarily imply a valid model (Roberts & Pashler, 2000).

2. Methods

2.1. Data

Four data sets were used involving 65 participants, 32 of whom were tested under two different conditions. All data sets were collected using similar word recognition memory tests. Participants were asked in each case to rate their confidence that a word had been previously presented from 1 (sure new) to 6 (sure old). These data sets were selected because they are based on sufficient data to allow for individual model fitting and because the methods were comparable. The data represent results from a laboratory that has generally supported the CDP model (Dede, Wixted, Hopkins, & Squire, 2013), a laboratory that has generally supported the HTDP model (Koen & Yonelinas, 2010), and a neutral laboratory (Van Zandt, 2000).

Dede et al. (2013). Participants were five memory-impaired patients with bilateral lesions limited to the hippocampus. Eleven age and education-matched controls were also tested. Participants were given three tests of recognition memory. In each test, participants were presented with 50 study words and asked to make pleasantness ratings. After a 3–5 min delay, participants were presented with 100 test words (50 new words and 50 old words). An additional group of seven age and education-matched controls were tested using an identical procedure but with the delay interval between study and test extended to one week.

Koen and Yonelinas (2010). Thirty-two undergraduate participants were presented with a mixed list of 80 words presented for 4 s and 80 words presented for 1 s. Immediately afterwards, participants were presented with 320 test words (160 old words and 160 new words). The data were analyzed as two separate sets, one based on the 80 study words presented for 4 s (plus 160 new words), and the other based on the 80 study words presented for 1 s (plus 160 new words).

Van Zandt (2000). Ten undergraduate participants were presented with 32 study words. Immediately afterwards, participants were presented with 20 of the study words and 20 new words. This procedure was repeated a total of 20 times, using different lists.

2.2. Analysis of systematic error

First, we fit the HTDP and CDP models to the data sets from the three laboratories to determine whether either model yielded a pattern of systematic error. All data were fit with both the HTDP and CDP models at the individual subject level using maximum likelihood estimation. These fits yielded predictions of the frequency with which a participant used each confidence-level response (1–6) for the study items. The observed frequencies of different confidence ratings to target items were then subtracted from the corresponding predicted frequencies derived from the model fits to calculate errors in each model's predictions. If errors for a particular confidence rating are random and non-systematic, then they should have a mean of zero across participants. If errors are systematic, then they should deviate systematically from zero. To test for such systematic error, a series of one-sample *t*-tests determined whether there was significant non-zero error at each confidence level within each of the four data sets. Errors were deemed systematic if they were identified as significant in all four data sets (Dede et al., 2013; Koen & Yonelinas, 2010, 4-s condition; Koen & Yonelinas, 2010, 1-s condition; Van Zandt, 2000).

2.3. Individual prediction analysis

In this analysis, we created predictions of model error that were based on each participant's performance. This analysis was computationally similar to the parametric bootstrap analysis used by Wagenmakers, Ratcliff, Gomez, and Iverson (2004) to assess model mimicry. To understand this analysis conceptually, consider the systematic errors that are produced when the HTDP model is fit to real data. If the same systematic errors are generated when the HTDP model is fit to data generated by the CDP model in simulation, then the inference can be made that the CDP model, having accurately predicted the HTDP model's error, is likely to accurately reflect the phenomenon that produced the real data. Six steps were applied to each participant individually. Step 1: 500 non-parametric bootstrap samples were taken. Step 2: these samples were fit with both the HTDP and CDP models using maximum likelihood estimation. The average error generated in these fits across the 500 bootstrap samples was used to measure systematic error. Step 3: using the parameters obtained in Step 2, simulated data were created by both the HTDP and CDP model simulators described in Section A.1. This step yielded 500 simulated data sets for each model. Step 4: the HTDP model was fit to the CDP model simulation data, and the CDP model was fit to the HTDP model simulation data. Step 5: the fits from Step 4 were used to derive an error prediction at each confidence level for each model. The error prediction was the mean error value for each confidence rating, as predicted by each model individually across the 500 simulated data sets. Step 6: the predicted error values for each model's fit were correlated with the observed error values in each individual's data. This was done in two ways. The predicted error values were correlated with the observed error values found when each model was directly fit to the original data and with the mean error values found when each model was fit to the non-parametric bootstrapped data (Step 2). The bootstrapping procedure was used to obtain a pattern of observed systematic error that was more robust to noise. The histograms of the correlation values across participants were plotted for visual inspection, and the correlation distribution produced by each model was compared to 0 using one-sample *t*-tests (see Section A.2 for a detailed example based on an individual participant and Section A.3 for further analyses of model flexibility).

3. Results

The first objective was to identify systematic errors in the fits of each model to recognition memory data. Accordingly, we fit both models to four sets of data from three studies of recognition memory (Dede et al., 2013; Koen & Yonelinas, 2010; Van Zandt, 2000). Fits were performed using maximum likelihood estimation on an individual participant basis (see Section 2). For the data from Dede et al. (2013), there were no significant differences in the error patterns across groups, so data from the different groups were combined (patients, controls tested with no delay, controls tested with a one-week delay).

3.1. The HTDP model but not the CDP model generated systematic error

Fig. 2a shows the pattern of error when the recognition memory data were fit with the HTDP model. Fig. 2b shows the pattern of error when the same data were fit with the CDP model. The four sets of data indicate that the HTDP model consistently underestimated the frequency of low memory strength responses to target items (i.e., confidence ratings of 1 on the 6-point scale), consistently overestimated the frequency of medium-strength responses to target items (confidence ratings of 3), and consistently underestimated the frequency of high-strength responses to target items (confidence ratings of 5) (Fig. 2a). There was no trend towards systematic error in the fit of the CDP model, and no instance where all four data sets identified a significant error (Fig. 2b).

The systematic errors generated by the HTDP model suggest that the HTDP model did not accurately describe how responses of different memory strength would be distributed in tests of recognition memory. Note that the errors generated by the HTDP model were the same errors predicted from Fig. 1, as outlined in Section 1. That is, the HTDP model generated the errors indicated by areas a, b, and c in Fig. 1.

3.2. The CDP model predicted the errors that were generated by the HTDP model at an individual level

Before presenting the results of this analysis, it is useful to explain the logic of our technique. Consider models A and B. When model A generates simulated data and model B is fit to that

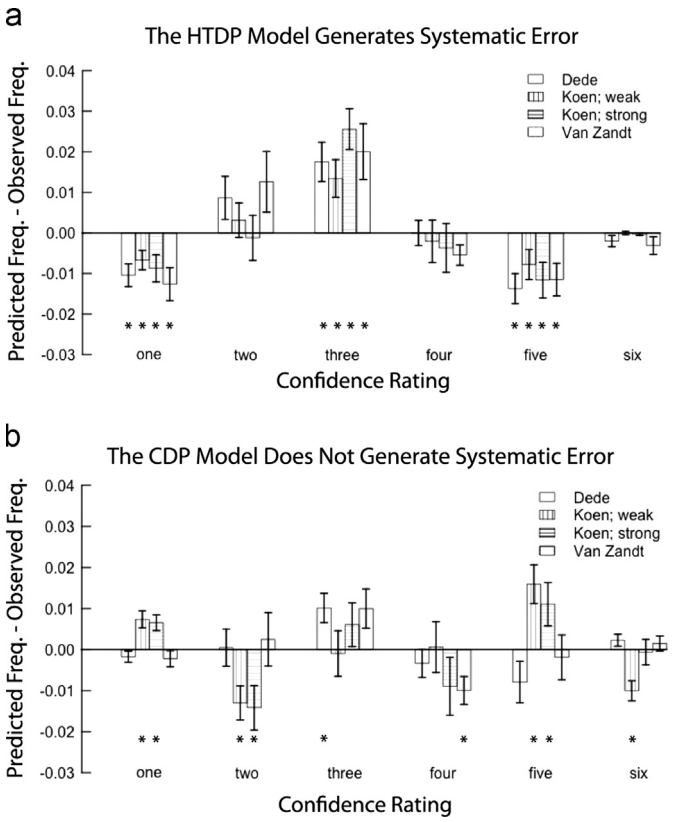


Fig. 2. Errors in the fits of the high threshold dual process model (HTDP) and the continuous dual process (CDP) to study items from four sets of data. (a) The HTDP model consistently underestimated the frequency of 1 and 5 responses and consistently overestimated the frequency of 3 responses. (b) For the CDP model there was no instance where all four data sets identified a systematic error. Error bars indicate SEM. * denotes $p < .05$ in single sample t-tests compared to zero.

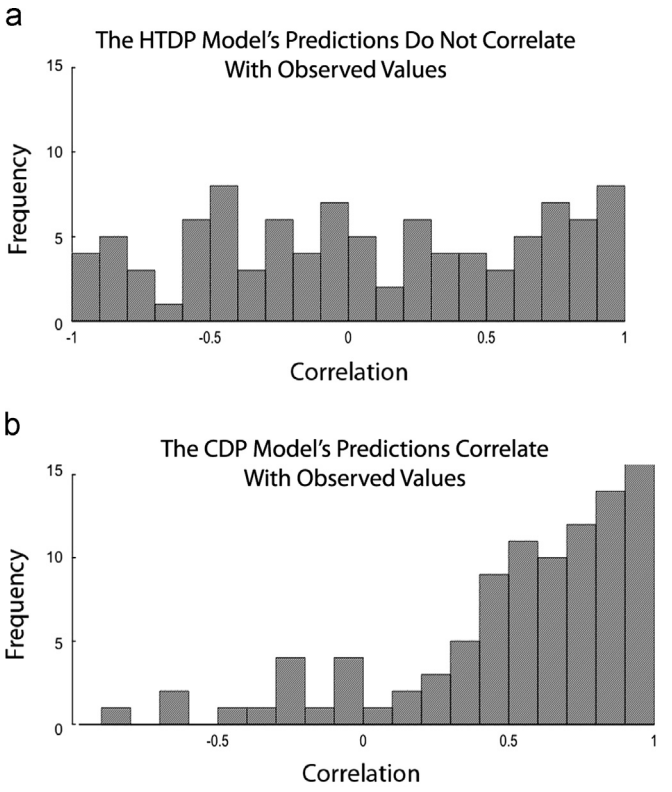


Fig. 3. Distribution of correlation values (one value for each of 97 participants) between the errors generated by fitting one model to the data and the errors predicted by the other model. (a) The CDP model's predictions of error are well correlated with the errors generated by the HTDP model. (b) The HTDP model's predictions of error are not well correlated with the errors generated by the CDP model.

simulated data, there will be a certain pattern of systematic errors in the fit of model B. This pattern will reflect the differences between models A and B and can be thought of as the pattern of predicted errors in the fit of model B. Most importantly, the predicted error pattern in the fit of model B is conditional on model A producing the data. Turning to the fitting of real data, it is unknown which model best approximates the phenomenon under study, but if the predicted pattern of error in the fit of model B is similar for real data and for data simulated by model A then model A likely reflects the phenomenon that produced the real data. This entire process and logic can be reversed to provide predictions of the errors in the fit of model A when model B produces the data.

For each participant, we correlated the pattern of error that was generated when each model was fit to the data with the pattern of error that was generated when each model was fit to data simulated by the other model. We also correlated the average pattern of error that was generated when each model was fit to a set of 500 non-parametric bootstrapped samples with the pattern of error that was generated when each model was fit to data simulated by the other model. This analysis resulted in two sets (one based on fits to raw data and one based on fits to bootstrapped data) of 97 correlations for each model (65 participants, 32 of whom were tested in two different conditions), based on the frequency of ratings at each confidence level (1–6). Fig. 3a shows the distribution of correlations (one correlation for each participant) between the errors generated by fitting the HTDP model to the data and the CDP model's prediction of errors. The average correlation was .44, a value greater than zero ($t(96) = 9.5$, $p < .001$). When this analysis was based on bootstrapped error patterns, which should be less susceptible to noise, the average correlation value increased to .54 ($t(96) = 12.9$, $p < .001$). Thus, the CDP model predicted the (systematic) errors made by the HTDP model when the HTDP model was fit to individual data. Fig. 3b shows the corresponding distribution of correlations between the errors generated by fitting the CDP model to the data and the HTDP model's prediction of errors. The average correlation was .02, which was not different from zero ($t(96) = .3$, $p = .75$). When this analysis was based on bootstrapped error patterns, the average correlation increased to .08 ($t(96) = 1.2$, $p = .22$), a smaller increase than was seen for the CDP model. Thus, the HTDP model did not predict the (nonsystematic) errors made by the CDP model when the CDP model was fit to individual data.

4. Discussion

Taking a novel approach to an old problem, we have found support for the CDP model and evidence against the HTDP model. In our first analysis (Fig. 2), the HTDP model exhibited systematic errors in its ability to predict the frequency of different confidence responses to target items (despite providing a good fit to the data, which is often taken as evidence of its validity). If the HTDP model accurately accounted for recognition memory, then errors in the predictions made by the best-fitting version of the model for each level of confidence should have been randomly distributed. Instead, the errors were systematic. These systematic errors imply that the HTDP model's assumptions about recognition memory are inaccurate.

By contrast, the CDP model did not exhibit systematic errors. Yet the absence of systematic error alone does not confirm the accuracy of the assumptions about recognition memory that

underlie the CDP model. Accordingly, we next asked whether the CDP model could predict the errors generated by the HTDP model. This analysis was performed at the individual level and demonstrated that the CDP model not only fits the data without systematic error but also predicts the systematic errors evident in the fits provided by the HTDP model (Fig. 3a). A second analysis (Fig. 3B) demonstrated that the HTDP model did not predict the (nonsystematic) errors evident in the fits provided by the CDP model. Taken together, these results suggest that the CDP model accurately accounts for recognition memory decision making and that the HTDP model does not.

There were two potential concerns about our analyses that are worth drawing attention to. First, in order to generate error predictions, the CDP model was always fit to data simulated by the HTDP model, and the HTDP model was always fit to data simulated by the CDP model. Our assumption was that neither model would predict errors in its own fit to the data (because those errors would be random). We tested this assumption by fitting the CDP model to data simulated by the CDP model and by fitting HTDP model to data simulated by the HTDP model. This analysis yielded no systematic errors, confirming our assumption.

Second, the CDP model is known to be slightly more flexible than the HTDP model, and it was unknown what effect this would have on our analyses of systematic error. We addressed this issue in an analysis presented in Section A.3 and found that model flexibility did not play a role in our results.

Within the discipline of cognitive psychology, reservations about the validity of the HTDP model have been expressed by many different researchers (e.g. Glanzer et al., 1999; Heathcote, 2003; Healy et al., 2005; Qin, Raye, Johson, & Mitchell, 2001; Qin, Raye, Johnson, & Mitchell, 2001; Rotello, Macmillan, Reeder, & Wong, 2005; Slotnick & Dodson, 2005; Starns, Rotello, & Ratcliff, 2012; Starns, Ratcliff, & McKoon, 2012). Thus, it is of interest to ask why the HTDP model has nevertheless held favor in guiding research into the neural substrates of recognition memory. An important consideration is that both the HTDP and CDP models virtually always fit experimental data well. Further, investigators often interpret a good fit to imply that a model is valid even though that is not a safe assumption (Roberts & Pashler, 2000). But if one does assume that both models are valid because they fit the data well, there is a choice to be made. On the one hand, the CDP model does not provide a simple way to differentiate between familiarity and recollection on the basis of old/new decisions alone. On the other hand, the HTDP model does. Assuming that a good fit implies a good model, and if the goal is to identify neural substrates of recollection and familiarity, the choice is straightforward: the HTDP model is the one to use.

Yet, considering the fundamentally different ideas about recollection inherent in the HTDP and CDP models, it should be clear that both models cannot be correct. The analyses presented here demonstrate that the CDP model is viable, but that the HTDP model is not (despite the fact that the HTDP model fits the data well). In light of the evidence presented here against the HTDP model, it would make sense to use the CDP model to guide studies of recognition memory (at least when words are used as stimuli, as they often are). It would also make sense to reconsider conclusions about the neuroanatomy of recognition memory that depend on the validity of the HTDP model.

Studies of recognition memory have commonly used fMRI and lesion studies to identify structures important for recollection and familiarity. Many of these studies have relied upon the assumptions of the HTDP model for interpreting the data (e.g. Aggleton et al., 2005; Ranganath et al., 2004; Yonelinas et al., 2002; Yonelinas, Otten, Shaw, & Rugg, 2005; for review see Eichenbaum, Yonelinas, & Ranganath, 2007). These studies have led to the idea that the hippocampus is important for recollection and that the surrounding medial temporal lobe (MTL) cortices are important for familiarity. To reach this conclusion, researchers have had to separate test trials based on recollection from test trials based on familiarity (e.g., in order to compare hippocampal activity for recollection-based vs. familiarity-based decisions), and the assumptions of the HTDP model have been relied upon for this purpose. Studies using the Remember-Know procedure, source memory procedures, and/or confidence rating procedures have all been implicitly or explicitly guided by the HTDP view of recollection. However, if the CDP model is correct (and the HTDP model is incorrect), then all of these studies share a common flaw in that trials assumed to differ only in whether recollection is present or absent also differ in memory strength (strong versus weak; Slotnick & Dodson, 2005; Wixted, 2007; Squire, Wixted, & Clark, 2007).

Unlike the HTDP model, the CDP model does not guide inquiry into the neural basis of recognition memory by providing quantitative estimates of recollection and familiarity. Instead, it suggests novel experimental designs that can be used to test whether (for example) the hippocampus plays a role in recollection and familiarity. A key idea suggested by this model is that it is important to control for memory strength because decisions thought to be based on recollection (e.g., Remember judgments) are typically made with higher confidence and higher accuracy than decisions thought to be based on familiarity (e.g., Know judgments). A difference in memory is not the essence of the theoretical difference between recollection and familiarity. Indeed, recollection can be weak and familiarity can be strong (Ingram, Mickes, & Wixted, 2012). Thus, memory strength is an experimental confound that should be controlled when comparing the two processes. For example, in one study that controlled for memory strength, the hippocampus was active when responses were based on recollection as well as when responses were based on familiarity (Smith, Wixted, & Squire, 2011). This result does not mean that the hippocampus and surrounding MTL structures provide only an undifferentiated signal of strength. Despite not being informed by the distinction between recollection and familiarity, the different structures of the MTL likely play different roles (Wixted & Squire, 2011a,b,c). Indeed, a recent study used state-trace analysis, combined with intracranial depth electrode recording, to demonstrate that the hippocampus and perirhinal cortex perform fundamentally different computations (Staresina, Fell, Dunn, Axmacher, & Henson, 2013). For further discussion concerning this issue, see Diana and Ranganath (2011), Montaldi and Mayes (2011) and Wixted and Squire (2011a,b,c).

In summary, we have found that the HTDP model does not accurately characterize recognition memory. Although the HTDP model can fit recognition memory data reasonably well, the relatively small errors it makes are systematic in nature. By contrast, the CDP model did not make systematic errors and also accurately predicted the systematic errors generated by the HTDP model. These findings suggest that the assumptions about recollection and familiarity that are underlie the CDP model (e.g., the assumption that recollection is a continuous process) are more accurate than the assumptions that underlie the HTDP model (i.e., e.g., the assumption that recollection is a threshold process). The key implication of these results is that the search for the neuroanatomical basis of recollection and familiarity should not be wedded to theoretical assumptions that are inconsistent with the empirical evidence.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.neuropsychologia.2013.10.012.

References

Aggleton, J. P., Vann, S. D., Denby, C., Dix, S., Mayes, A. R., Roberts, N., & Yonelinas, A. P. (2005). Sparing of the familiarity component of recognition memory in a patient with hippocampal pathology. *Neuropsychologia*, *43*, 1810–1823.

Dede, A. J. O., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2013). Hippocampal damage impairs recognition memory broadly, affecting both parameters in two prominent models of memory. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 6577–6582.

Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review*, *13*, 1–21.

Diana, R.A, & Ranganath, C. (2011). Recollection, familiarity and memory strength: Confusion about confounds. *Trends in Cognitive Science*, *15*, 337–338.

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, *30*, 123–152.

Glanzer, M., Hilford, A., & Kim, K. (2004). Six regularities of source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1176–1195.

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 500–513.

Healy, M. R., Light, L. L., & Chung, C. (2005). Dual-process models of associative recognition in young and older adults: Evidence from receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 768–788.

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

Howard, M. W., Bessette-Symons, B., Zhang, Y. F., & Hoyer, W. J. (2006). Aging selectively impairs recollection in recognition memory for pictures: Evidence from modeling and receiver operating characteristic curves. *Psychology and Aging*, *21*, 96–106.

Ingram, K. M., Mickes, L., & Wixted, J. T. (2012). Recollection can be weak and familiarity can be strong. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 325–339.

Kelley, R., & Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 701–722.

Koen, J. D., & Yonelinas, A. P. (2010). Memory variability is due to the contribution of recollection and familiarity, not to encoding variability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1536–1542.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guid*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Montaldi, D., & Mayes, A. R. (2011). Familiarity, recollection and medial temporal lobe function: an unresolved issue. *Trends in Cognitive Science*, *15*, 339–340.

Qin, J. J., Raye, C. L., Johson, M. K., & Mitchell, K. J. (2001). Source ROCs are (typically) curvilinear: Comment on Yonelinas (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1110–1115.

Ranganath, C., Yonelinas, A. P., Cohen, M. X., Dy, C. J., Tom, S. M., & D'Esposito, M. (2004). Dissociable correlates of recollection and familiarity within the medial temporal lobes. *Neuropsychologia*, *42*, 2–13.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). The remember response: subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review*, *12*, 865–873.

Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition*, *33*, 151–170.

Smith, C. N., Wixted, J. T., & Squire, L. R. (2011). The hippocampus supports both recollection and familiarity when memories are strong. *Journal of Neuroscience*, *31*, 15693–15702.

Squire, L. R., Wixted, J. T., & Clark, R. E. (2007). Recognition memory and the medial temporal lobe: a new perspective. *Nature Reviews Neuroscience*, *8*, 872–883.

Staresina, B. P., Fell, J., Dunn, J. C., Axmacher, N., & Henson, R. N. (2013). Using state-trace analysis to dissociate the functions of the human hippocampus and perirhinal cortex in recognition memory. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 3119–3124.

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology*, *64*, 1–34.

Starns, J. J., Rotello, C. M., & Ratcliff, R. (2012). Mixing strong and weak targets provides no evidence against the unequal-variance explanation of zROC slope: A comment on Koen and Yonelinas (2010). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 793–801.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 582–600.

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28–50.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*, 152–176.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review*, *117*, 1025–1054.

Wixted, J. T., & Squire, L. R. (2011a). The medial temporal lobe and the attributes of memory. *Trends in Cognitive Science*, *15*, 210–217.

Wixted, J. T., & Squire, L. R. (2011b). Confusion abounds about confounds: Response to Diana and Ranganath. *Trends in Cognitive Science*, *15*, 338–339.

Wixted, J. T., & Squire, L. R. (2011c). The familiarity/recollection distinction does not illuminate medial temporal lobe function: response to Montaldi and Mayes. *Trends in Cognitive Science*, *15*, 340–341.

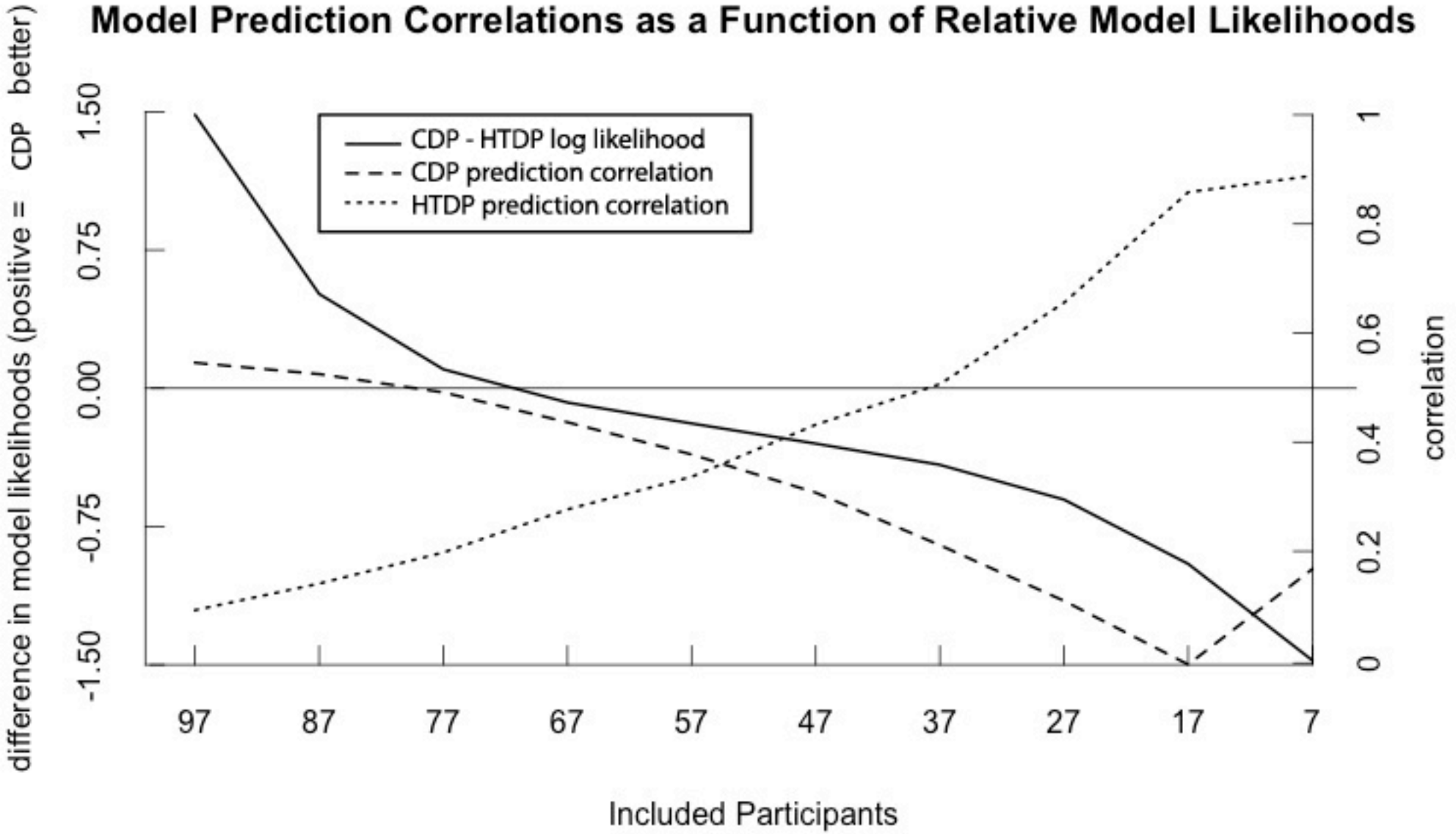
Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1341–1354.

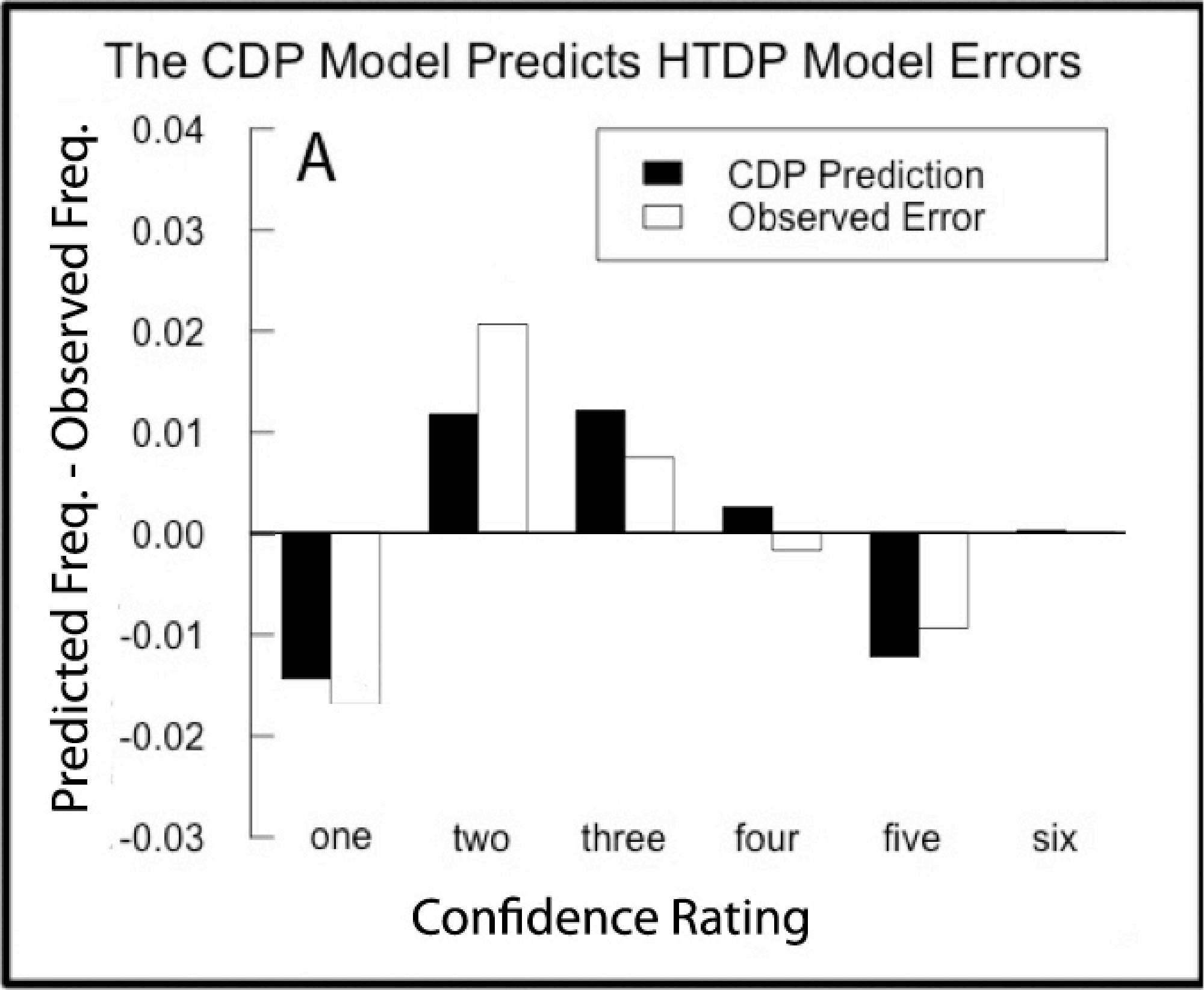
Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1415–1534.

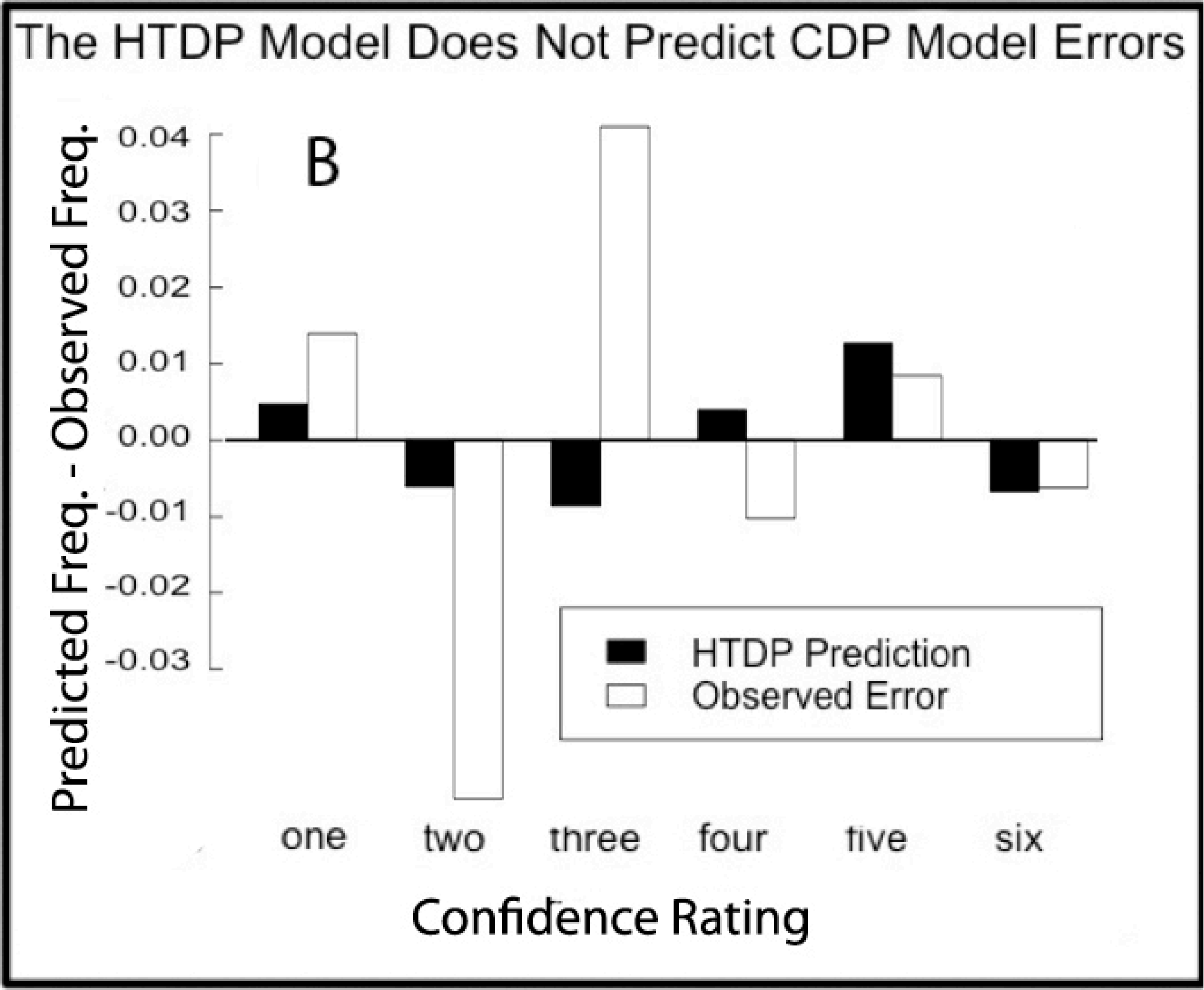
Yonelinas, A. P. (1999b). Recognition memory ROCs and the dual-process signal-detection model: Comment on Glanzer, Kim, Hilford, and Adams (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 514–521.

Yonelinas, A. P., Kroll, N. E. A., Quamme, J. R., Lazzara, M. M., Sauve, 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*, 1236–1241.

Yonelinas, A. P., Otten, L. J., Shaw, K. N., & Rugg, M. D. (2005). Separating the brain regions involved in recollection and familiarity in recognition memory. *Journal of Neuroscience*, *25*, 3002–3008.


Model Prediction Correlations as a Function of Relative Model Likelihoods
CDP - HTDP log likelihood
CDP prediction correlation
HTDP prediction correlation
difference in model likelihoods (positive = CDP better)
1.50
0.75
0.00
-0.75
-1.50
correlation
1
0.8
0.6
0.4
0.2
0
97
87
77
67
57
47
37
27
17
7
Included Participants


The CDP Model Predicts HTDP Model Errors
A
CDP Prediction
Observed Error
Predicted Freq. - Observed Freq.
0.04
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
one
two
three
four
five
six
Confidence Rating


The HTDP Model Does Not Predict CDP Model Errors
B
Predicted Freq. - Observed Freq.
0.04
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
HTDP Prediction
Observed Error
one
two
three
four
five
six
Confidence Rating

Appendix

A.1 Model Simulators

Simulator functions were written in R to create data for both the CDP and HTDP models. Email the first author for these functions if interested.

For the CDP model, a CDP_simulator(d, s, criteria) function was written. The function involved seven values: d, s, and a set of five criteria. d corresponded to the CDP d parameter and s corresponded to the CDP slope parameter. The CDP_simulator function sampled 150 times from a distribution with mean equal to d and standard deviation equal to 1/slope. These samples were stored as study items. Next the simulator sampled 150 times from a distribution with mean equal to zero and standard deviation equal to one. These samples were stored as new items. Using the five criteria points, the study items and the new items were assigned confidence ratings on a 1-6 scale (6 indicated that the item was very likely to be a study item and 1 indicated that the item was very likely to be a new item).

A similar HTDP_simulator(f, R, criteria) function was written for the HTDP model. The f value of this function corresponded to the familiarity parameter of the HTDP model, and R corresponded to the recollection parameter. Again, 150 study items and 150 new items were assembled. For each study item, a random number was generated between 0 and 1. If the random number was less than or equal to the recollection parameter, then the item was randomly assigned a value from a distribution having a mean equal to f plus 6 and a standard deviation equal to 1. In practice, this procedure resulted in recollection-based decisions always being assigned a confidence rating of 6. If the random number was greater than the recollection parameter, then the

sample was assigned a value from a distribution having a mean equal to f and standard deviation equal to 1. Then the samples of study items and new items were assigned confidence ratings in the same way as was done for the CDP model.

A.2 Individual Prediction Analysis

The goal of the individual prediction analysis (section 2.3) was to create predictions of model error that were tailored to the parameter space of each individual participant and to compare those predictions to each individual's observed error pattern. Here we present a detailed description of the analysis by summarizing the data analysis and illustrating the results for a representative participant.

Step 1: 500 non-parametric bootstrap samples were taken of the data. For each bootstrap sample, the original data were placed into two vectors of confidence ratings, one for studied items and one for distractors (as in Table A.1). Next, values were randomly drawn with replacement from each vector until new studied-item and distractor vectors were created containing observations that were equal in number to the number of observations in the original data. This process was repeated 500 times, which resulted in a list of 500 non-parametric bootstrap samples (i.e., 500 hypothetical data sets for a single participant) with each formatted in the same way as the original data. In the example here, this procedure resulted in a mean percent correct of .73 across bootstrap samples with a standard deviation of .01. The original data had a percent correct of .73.

Step 2: Each of the 500 non-parametric bootstrap samples was fit with both the CDP model and the HTDP model using maximum likelihood estimation. There were seven free parameters of each model. Five of these were criterion parameters, which were placed along the continuous strength of memory axis (Figure 1). The remaining two parameters were specific to each model. The CDP model's parameters were d and slope. d represented the distance between the mean of the studied item distribution and the mean of the distractor distribution standardized by the standard deviation of the distractor distribution. Slope represented the ratio of the standard deviation of the distractor distribution to the standard deviation of the studied item distribution. The HTDP model's parameters were f and R. f represented the distance between the mean of the familiarity distribution and the mean of the distractor distribution standardized by the standard deviation of the distractor distribution, which was defined to be equal to the standard deviation of the studied item distribution. R represented the probability that any individual studied item would be recollected. See Table A.2 for the results of this step in the present example. Table A.2 shows the parameter estimates for the two key parameters of each model when fit to the participant's data (first line) as well as the average parameter estimates for the two key parameters of each model when fit to the 500 bootstrapped samples drawn from that participant's data. Obviously, the parameter estimates from the real data and the average parameter estimates bootstrapped samples are the same. The average error pattern across the 500 bootstrapped samples was calculated for each model. This was used as a measure of the observed error that was more robust to noise than the observed error pattern generated when the models were fit to the raw data.

Step 3: Using the 500 sets of parameter estimates from the bootstrapped samples obtained in Step 2, 500 simulated data sets were created by both the HTDP and CDP model simulators described in section A.1. This step yielded 500 simulated data sets per participant for each model. In the present example the 500 data sets simulated by the CDP model had a percent correct of .73 (SD = .03), and those simulated by the HTDP model had a percent correct of .78 (SD = .02). Note that we could have produced one simulated HTDP data set and one simulated CDP data set for a given participant using the parameters from the fits to the real data (e.g., first line in Table A.1) instead of creating 500 simulated data sets for each model (using parameter estimates from the fits to the bootstrapped samples). The reason for computing 500 simulated data sets for each model is that this procedure more accurately characterizes the predicted performance of a particular participant because the results do not rely only on the data pattern produced on the one occasion the participant was tested.

Step 4. At this point, for each participant, we had 500 simulated data sets known to have been produced by the CDP model and 500 simulated data sets known to have been produced by the HTDP model. Next, for each participant, the HTDP model was fit to the 500 data sets simulated by the CDP model, and the CDP model was fit to the 500 data sets simulated by the HTDP model. That is, the wrong model was used to fit each data set to determine what systematic errors should be observed according to each model when the other (wrong) model is fit to the data. These fits were used to generate patterns of

systematic error by subtracting the observed frequencies of each confidence response from the corresponding model predictions of response frequency.

Step 5. The fits from Step 4 were used to derive an error prediction at each confidence level for each model. The error prediction was the mean error value for each confidence rating, as predicted by each model individually across the 500 simulated data sets. Figure A.1 shows the error predictions of the two models (solid bars) and the observed error (as measured using bootstrap; open bars) in each model for the present example.

Step 6. The predicted error values for each model's fit were correlated with the observed error values (based both on fits to raw data and to bootstrapped data). In the present example the correlation between the CDP model's predictions and the HTDP model's error was .93 (.94 bootstrapped), and the correlation between the HTDP model's predictions and the CDP model's error was .01 (.01 bootstrapped). The histograms of the correlation values across all participants were plotted for visual inspection (Figure 3), and the correlation distribution produced by each model was compared to 0 using one-sample t-tests (see section 3.2).

A.3 The Effect of Model Flexibility

In our analyses, and in many others, it has been shown that the CDP model will generally fit recognition memory data better than the HTDP model. This advantage of the CDP model might occur because the CDP model is slightly more flexible than the HTDP model, or perhaps because the CDP model is a more accurate representation of the underlying phenomena of recognition memory. The problem of differentiating between

truth and flexibility was one of the motivating reasons for the present analysis of systematic error. Note first, however, that no matter how little total error there is in a model's fit, an accurate model should never show systematicity in its error. By this logic, the HTDP model was proven incorrect in our initial analysis of systematic error (Section 3.1).

The next step of our analysis was to show that the CDP model is correct because it has the ability to predict the systematic errors of the HTDP model. Here, model flexibility may play a role. If a model is flexible, then to the extent that it is capable of mimicking the original data it will accurately predict the systematic errors of another model's fit to the original data. Thus, it was necessary to compare the HTDP and CDP models under conditions in which they mimicked the data equally well. Goodness-of-fit is effectively a measure of the extent to which a model can mimic the data to which it is fit. Thus, participants were eliminated one at a time from the individual prediction analysis (Sections 2.4, 3.3, and A.2) such that the participant with the greatest CDP model over HTDP model goodness-of-fit advantage was always eliminated. This process was repeated until, for the remaining participants, the two models fit equally well. Goodness-of-fit was measured with the log likelihood value for each model. After the elimination of each participant, the average log likelihood of each model was recalculated in the remaining group as were the average prediction correlation values for both the HTDP and CDP models. As shown in Figure A.2, the average log likelihood of the HTDP model was equal to the average log likelihood of the CDP model after 27/97 participants. That is, after 27 participants had been eliminated from the analysis, the two models were able to mimic the data for the remaining participants equally well. Thus,

model flexibility was unlikely to play a role in this subset. Despite fitting the remaining data equally well, the CDP model still performed far better than the HTDP model at predicting the systematic error of the competing model (r_{HTDP}=.21; r_{CDP}=.48). The CDP model continued to predict the errors of the HTDP model better than the HTDP model could predict the errors of the CDP model, even as we continued to eliminate participants who were better fit by the CDP model. It was not until 48/97 participants had been eliminated from the analysis that the HTDP model began to predict CDP model error better than the CDP model predicted HTDP model error.

The fact that there was such a large gap in this analysis between the point at which the HTDP model began to fit better and the point at which the HTDP model began to predict errors more accurately indicates that model flexibility did not influence our results. Highly flexible models are able to mimic any data, leading to high goodness-of-fit. This analysis eliminated the CDP model's goodness-of-fit advantage (whether that advantage was due to its theoretical accuracy or to its higher flexibility), and demonstrated that the CDP model was still capable of predicting the HTDP model's errors. Thus, we conclude that the advantage of the CDP model over the HTDP model is a reflection of the fact that the theoretical assumptions that underlie the CDP model provide a closer approximation to the truth than the theoretical assumptions that underlie the HTDP model.

Figure Captions.

Figure A.1. Predicted and observed error patterns for an individual participant. Predicted values are based on the procedure described in sections 2.3 and A.2. A. There was a strong correlation between the CDP model's prediction of the HTDP model's error and the observed values for the HTDP model's error. B. There was not a strong correlation between the HTDP model's prediction of the CDP model's error and the observed values for the CDP model's error.

Figure A.2. Model prediction correlations as a function of relative model log likelihoods. As participants with disproportionately good CDP compared to HTDP model fits were eliminated from the analysis, the average difference in log likelihood between the two models dropped down to zero and then become negative (indicating better average HTDP model fit). In parallel, the CDP model's ability to predict HTDP model error decreased and the HTDP model's ability to predict CDP model error increased. The cross-over point at which the HTDP model begins to fit better than the CDP model was far before the cross-over point at which the HTDP model began to make better predictions than the CDP model.

Table A.1. Example data

Confidence Rating	Studied Items	Distractor Items
6	185	5
5	158	38
4	173	106
3	106	131
2	110	296
1	68	224

Table A.2. Model parameters from original and bootstrapped data.

Data	d	slope	f	R
Original Data	1.41	.70	.92	.19
Bootstrap Data	1.41(.08)	.70(.04)	.92(.07)	.19(.02)

Note: Mean (Standard deviation)