Cognition 192 (2019) 103988

Contents lists available at ScienceDirect

Cognition

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

Original Articles

The language of accurate recognition memory

Ian G. Dobbins[a,*], Justin Kantner[b]

[a] Washington University in Saint Louis, United States
[b] California State University, Northridge, United States

ARTICLE INFO

Keywords:
Recognition memory
Language content analysis
Machine learning
Recollection
Receiver operating characterstics

ABSTRACT

The natural language accompanying recognition judgments is a largely untapped though potentially rich source of information about the kinds of processing that may support recognition memory. The current report illustrates a series of methods using machine learning and receiver operating characteristics (ROCs) to examine whether the language participants use to justify their 'old' and 'new' recognition decisions (viz., memory justifications) predicts accuracy. The findings demonstrate that the natural language of observers conveys the accuracy of 'old' (hits versus false alarms) but not 'new' (misses versus correct rejections) decisions. The classifier trained on this language was considerably more predictive of accuracy than the initial speed of the decisions, generalized to the justification language of two independent experiments using different procedures, and appeared sensitive to the presence versus absence of recollective experiences in the observer's reports. We conclude by considering extensions of the approach to several basic and applied areas, and, more broadly, to identifying the explicit bases (if any) of classification decisions in general.

1. Introduction

Introspection has a checkered past in Psychology, from its early use in examining sensory and perceptual phenomena to its more recent application to socioemotional cognition. Indeed, Nisbett and Wilson (1977) concluded that observers' verbal justifications of social behavior do not reflect access to critical content, but instead reflect implicit causal theories that are used to generate post hoc explanations of their actions or conclusions. Turning to the field of memory, research in eyewitness identification demonstrates that observers will not only increase their reported certainty, but will claim to have experienced more favorable encoding conditions if they receive suggestive feedback after an initial lineup identification is made (Wells & Bradfield, 1998). Thus, introspective reports about the causes of social behavior, and the veracity of personal memories are potentially sensitive to demand characteristics and contamination. However, while skepticism regarding the use of introspective reports in certain areas and with certain experimental designs is understandable, a general suspicion of subjective introspection now exists across most psychological domains. In the case of memory research, researchers almost never ask subjects to justify, in their own words, their memorial decisions or strategic approach to the testing procedures even when potential demand characteristics are unlikely and/or minimized by methodology. Nonetheless, as noted by Tulving (1985), 'One might think that memory should have something to do with remembering, and remembering is a conscious experience' (p. 1). Thus, despite the fact that episodic remembering is itself an introspective act, we know very little about the natural language observers use to describe their memorial experiences, or whether this language informs current models of memory and metacognition.

The current report begins to address this gap in several ways, focusing first on whether introspective language distinguishes accurate from inaccurate recognition decisions, and then on the psychological basis of this ability. We demonstrate a machine learning methodology for translating the language participants use in justifying memory decisions into an accuracy classifier, assessing classifier training efficiency, and conducting convergent and divergent validity tests of the language classifier. Additionally, we test whether the language classifier performs better at predicting recognition accuracy than a more traditional measure of decision uncertainty (reaction times). This experiment represents a critical test of the machine learning approach because it is the first design to minimize demand characteristics that might arise when recognition decisions are also accompanied by public statements of confidence, or by introspective judgments the researchers have extensively discussed with participants. It is also the first time a trained language classifier has been pitted against an ostensibly more objective measure of decision performance, and in the Discussion we consider why such classifiers might outperform traditional accuracy

* Corresponding author at: Department of Psychological & Brain Sciences, Washington University in Saint Louis, Saint Louis, MO 63130, United States.
E-mail address: idobbins@wustl.edu (I.G. Dobbins).

https://doi.org/10.1016/j.cognition.2019.05.025
Received 30 January 2019; Received in revised form 21 May 2019; Accepted 30 May 2019

indicators such as reaction time and confidence. Finally, we conclude by considering other domains in which this same basic approach should be useful. Before moving to the experiment proper we consider two prior studies that collected natural language during recognition memory decisions and were motivated by the Remember/Know procedure developed by Tulving (1985).

An important step towards examining introspective content supporting recognition decisions was the Remember/Know (R/K) procedure developed by Tulving (1985). During R/K paradigms, observers are asked whether items they identify as recognized from a prior study list were endorsed because of 'Remembering' or 'Knowing'. The former is held to reflect the recovery of first-person experiences of the prior encounter and is often termed recollection, whereas the latter is an acontextual sense of recent or frequent encounter tied to the item itself and often termed item familiarity or fluency. If R/K reports map onto different states of consciousness, supported by different retrieval processes, then they should be experimentally dissociable. Consistent with this, there is a large body of research demonstrating dissociations of remembering and knowing rates (for review see, Diana, Reder, Arndt, & Park, 2006; Yonelinas, 2002) leading to the development of decision models that assume separable underlying recollection and familiarity/fluency processes contribute to remember and know decisions (e.g., Mickes & Wixted, 2010; Yonelinas, 1994). These approaches, in contrast to models that assume a single source of undifferentiated evidence governs recognition decisions (e.g., Donaldson, 1996; Dunn, 2004), are compatible with the possibility that the introspective content of remember and know experiences may qualitatively differ. Critically however, the R/K procedure advanced by Tulving was not based on observed differences in the actual introspective language observers use during memory decisions but was instead inferred from a converging body of empirical and clinical findings suggesting that remembering reflects a distinct state of consciousness linked to the self as an agent within a specific prior episode. Despite the force of this argument, it remained the case that the subjective experiences of participants, as reflected by their natural language, was never directly tested to see if it was consistent with the R/K dichotomy.

To address this, Gardiner, Ramponi, and Richardson-Klavehn (1998) had participants complete a recognition test, following a one day retention interval, during which they indicated whether items endorsed as 'Old' were 'Remembered', 'Known', or simply 'Guessed'. Immediately following this test, a subset of probes previously receiving 'Old' reports was re-presented and participants were asked to explain what led them to initially recognize the materials. We refer to these explanations as *justifications*. After analyzing the justifications, the authors concluded that Remember responses "reflect the use of effortful strategies, associations, and imagery" (Gardiner et al., 1998, p. 5) and Know responses lacked "any indication that they involved recollection of any specific contextual details" (Gardiner et al., 1998, p. 7). Usefully, all of the justifications collected by the researchers were provided in an appendix.

Although the findings of Gardiner et al. (1998) were important and in agreement with the constructs of recollection and familiarity, the observers received extensive instructions about the purported content of remembering versus knowing. These consisted of written instructions, outlining the distinctions between Remembering, Knowing, and Guessing, followed by oral instructions that were tailored to the subject's apparent understanding of the distinctions, and both are contained in Appendix A. Aside from these instructions, when subjects later provided justifications, their prior endorsement of the probes as Remembered, Known or Guessed was visible. In combination, this raises the possibility that observers ensured their justifications supported the distinctions outlined in the instructions, inflating the convergence between natural language and the R/K distinction (see also, McCabe, Geraci, Boman, Sensenig, & Rhodes, 2011). Moreover, the scoring of the Gardiner et al. materials was largely informal, with no attempt to quantify the differences in language across Remember, Know and Guess justifications. Instead, the researchers examined specific characteristics within each class of responses to see if they were consistent with the R/K distinction. For example, for the Remember justifications, two raters coded for the presence of intra-list associations, extra-list associations, item-specific imagery, item physical features, and self-related content. These characteristics appeared to be prevalent in the justifications, with good agreement between the two raters for the coding (81%). For Know justifications however, the authors simply noted that through inspection it was clear they lacked recollection of specific details, and no attempt was made to quantify the language differences across Remember and Know justifications.

Finally, there were several design characteristics that potentially complicate the interpretation of the justifications provided in the Gardiner et al. (1998) report. First, the experimental design also manipulated word frequency (low versus high) and recognition decision biases. For the latter, some subjects were correctly led to believe 50% of the materials were studied whereas others were incorrectly led to believe only 30% were studied. However, neither factor is coded in the provided justifications, which therefore represent an unknown mix of word frequency and subject biases (although word frequency could be calculated since the recognition probe word is indicated alongside each justification). Of more concern, the collection of justifications was not restricted to hits, but to 'old' reports, which the experimenter scrolled through to try to find two instances of each introspective report (Remember, Know, or Guess). This means that the Remember, Know, and Guess justification categories contain an unknown mix of accurate and inaccurate 'old' judgments. Since errors are likely more prevalent in Know than Remember reports, accuracy is probably confounded across these justification categories.

To address these concerns, Selmeczy and Dobbins (2014) instead had observers provide justifications following two, pseudo-randomly selected, hit and correct rejections accompanied by simple low, medium, or high confidence ratings. Critically, this design made no mention of the putative phenomenology of recollection and familiarity, and was the first to use an objective machine learning approach to determine if the language of the justifications systematically differed for high and medium confidence recognized probes. If recollection tends to support high confidence recognition, and familiarity tends to support medium confidence recognition, then a machine classifier applied to the natural language would be able to distinguish these two reports, identifying words from the justifications that were consistent with the theoretical distinction between remembering and knowing (Selmeczy & Dobbins, 2014). It did so, such that a classifier trained to distinguish high and medium confidence hit justifications in one experiment generalized well when tested on the analogous justifications in a separate, independent experiment. Moreover, the classifier trained on 'old' justifications showed divergent validity in that it poorly distinguished justifications of high and medium confidence correct rejections, neither of which should contain language related to recollection. Thus, the classifier was not merely sensitive to language isolating high versus medium confidence broadly, but instead appeared to isolate recollection from familiarity-based language, a learned distinction that would not be useful in distinguishing high versus medium confidence correct rejections under dual process models that assume recollection is confined to 'old' recognition conclusions.

The findings of Selmeczy and Dobbins (2014) demonstrated that language differences isolating putative recollection and familiarity processes arise even in the absence of instructions informing subjects about these processes. Nonetheless, the design was susceptible to a subtler demand characteristic in which observers, having openly stated their confidence in their judgments, attempted to validate their confidence post hoc. For example, after indicating high instead of medium confidence an item was studied, and then being asked for a justification, subjects may have then re-engaged retrieval processes to provide further support, post hoc, for their stated high confidence. While this would result in genuine content differences for high versus medium

confidence 'old' justifications, these differences would have resulted from the confidence ratings instead of leading to them.

1.1. Experiment rationale

As noted above, providing subjects with instructions about the possible conscious correlates of recognition experiences, or requiring them to render metacognitive judgments prior to collecting introspective language justifications, may induce demand characteristics and/or differential retrieval efforts. To avoid these problems we conducted a recognition justification study but removed all metacognitive reporting requirements and avoided any instruction or discussion of the potential processes (e.g., recollection and familiarity) supporting recognition conclusions. Instead, we merely had subjects classify the items as old or new and then, on a random minority of trials, had them justify those decisions for (unbeknownst to them) accurate and inaccurate decisions. Under this design, any justification language differences following accurate versus inaccurate 'old' decisions, or following accurate versus inaccurate 'new' recognition decisions, necessarily reflects introspective content that differentiates accuracy. This content cannot be biased by reporting differences prior to the justifications, since the subjects make the same overt decisions within the 'old' and 'new' decision classes.[1] Aside from the key question of whether a machine learner could use introspective language to distinguish accurate from inaccurate recognition decisions (both old and new), we also test a specific interpretation for why our trained classifier successfully discriminated accurate from inaccurate 'old' decisions. Under our proposed recollection/non-recollection (R/NR) interpretation, the classifier distinguishes hit from false alarm justifications because it is sensitive to the presence versus absence of recollective experiences in the language. We test this hypothesis by examining the features the classifier selects during training, considering the individual justifications it 'believes' are the strongest illustrations of accurate and inaccurate judgments, and seeing if its predictions covary with individual differences in accuracy (i.e., face validity). We then examine the convergent and divergent validity of the R/NR interpretation by testing the trained classifier on the independent justifications obtained by Selmeczy and Dobbins (2014) and Gardiner et al. (1998) (i.e., construct validity). Finally, we consider whether the trained language classifier provides predictive information over and above reaction times (i.e., incremental validity), which are also known to discriminate accurate from inaccurate recognition decisions (Weidemann & Kahana, 2016), and are faster during initial, correct recognition judgments when followed by Remember then Know decisions (Dewhurst & Conway, 1994).

2. Method

2.1. Participants

Seventy-eight California State University, Northridge undergraduates participated in partial fulfillment of psychology course requirements.

2.2. Materials

Stimuli were 192 medium-to-high frequency nouns drawn from the MRC psycholinguistic database (http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm; Coltheart, 1981). All words contained between four and eight letters. Two 48-item study lists were populated by random selection from the 192-word pool. Two 96-item test lists contained the 48 words from the preceding study list mixed with 48 randomly selected new words. All study lists included 3 primacy and 3 recency buffers that were identical across participants and not tested. No words were repeated across the two study-test cycles. All lists were presented in a randomized order. The experiment was conducted with E-Prime software (Psychology Software Tools, Inc., Sharpsburg, PA).

2.3. Procedure

Data were collected from one to three participants at a time. Participants were instructed that they would be viewing a series of words and that they should try their best to memorize each one for a later test. Each study word appeared in the center of the screen for 1 s, followed by a blank 1 s interstimulus interval. Test instructions informed participants that they would be seeing a list composed of previously studied and non-studied items and that their task was to judge whether each item belonged to the former or latter category. Participants were also told that on a small number of trials they would be asked to type in an explanation for why they chose the 'old' or 'new' response. Each test word appeared in the center of the screen with the words "1 = Studied" and "0 = Not Studied" beneath. No feedback was provided. Upon pressing the appropriate key to register their old/new response, participants either proceeded to the following trial or, on a subset of trials, were first asked to "Please describe in as much detail as possible why you chose this response." On these trials, a box appeared into which participants could type up to 2000 characters. The test word remained on the screen while the justification was entered. Participants pressed the control key to indicate that they had completed their justification. Justifications were collected following one hit, false alarm, miss, and correct rejection in each half of the test (for a total of eight justifications across the 96-item test). Justification requests were subject to the constraints that no justifications could be collected within the first five trials of the test or within five trials of a previous justification following the same response outcome (i.e., hit, false alarm, correct rejection, or miss).

Neither recognition judgments nor justifications were speeded. A blank 500-ms interval separated each test trial. Following completion of the first study-test cycle, participants were engaged with word searches for 10 min. Participants then completed a second study-test cycle that was identical to the first. Across the two tests, a maximum of 16 justifications were collected, four for each of the four possible response outcomes, for each subject. This degree of sampling was chosen to avoid an excessive testing length and fatigue on the part of the participants, since each justification was hand typed by the participants. Although the majority of subjects provided four justifications for all four of the possible response outcomes (63/78: 81%), 15 did not. Of these, eight did not because of the sampling constraints noted above, which resulted in the program asking for one less error justification for seven participants, and two less error justifications for the remaining participant. Finally, the remaining seven participants provided either one (five participants) or more (two participants) blank justifications. Justification counts within this latter group were also reduced by the sampling constraints, eliciting seven (two participants) or six (two participants) error justifications. Overall, the sampling resulted in 308 justifications of hits, 308 of correct rejections, 300 of false alarms, and 301 of misses. Because there were 78 subjects, the maximum number of justifications possible for each decision outcome was 312, and hence the procedures achieved 96% or better collection rates with respect to the target goal. In total, only 8.3% of the observers' recognition decisions were queried for justification by the program.

To aid in exposition, we explain the methods of the machine learning analysis alongside the results below.

[1] Comparisons across 'old' and 'new' conclusions are not similarly protected from demand characteristics because participants have reached different overt conclusions ('old' versus 'new') and may have different beliefs about the kinds of content that should accompany and support these two conclusions.

3. Results

3.1. Behavior

Recognition accuracy was moderate, with an overall success rate of 67.5% arising from a hit rate of 65% and a correct rejection rate of 70%. Subjects were approximately unbiased, classifying 48% of the materials as studied.

3.2. Classifier training performance

The recognition justifications were not corrected for spelling, although contractions were spelled out for consistency. Five R language (Core, 2017) packages were used in the analyses and formatting; namely, quanteda (Welbers, Atteveldt, & Benoit, 2017), tidyverse (Wickham, 2017), pROC (Robin et al., 2011), gridExtra (Augie, 2017), and stargazer (Hlavac, 2018). Core scripts and data are provided at https://osf.io/465vw/. Following cleaning, the justifications were transformed into a document-feature-matrix (DFM) in which each row is a memory justification and each column a word present in the entire collection of justifications (the corpus). This is the so-called 'bag of words' approach because individual words are used as potentially predictive features without regard to syntax.[2] Within the DFM, each cell codes the tendency of each word to appear in each justification of each class (hit, false alarm, correct rejection, or miss). Here we use frequency coding in which the cell simply indicates the number of times the word appeared in the justification. A logistic regression model was then trained to use the resulting word predictors in learning the distinction between errors and correct responses separately for 'old' (i.e., hits versus false alarms) and 'new' (i.e., correct rejections versus misses) recognition decisions.

Here we use a logistic regression classifier under the LASSO (Least Absolute Shrinkage and Selection Operator) constraint (Tibshirani, 1996). The Lasso penalizes the model for reacting wildly to small changes in the predictors (viz., overfitting) by constraining the sum of the absolute values of the estimated regression coefficients. The appropriate severity of this constraint, λ, is determined via 10-fold cross validation across a range of potential λs in order to find one that jointly minimizes overfitting while preserving predictive ability. Remarkably, not only does the Lasso control overfitting, this same constraint also forces a large number of predictors (in this case, words within the justifications) entirely from the regression. This yields a final classifier in which relatively few words (compared to the possible candidates) are used to predict the memory classes of the justifications (for a review see James, Witten, Hastie, & Tibshirani, 2013). This is useful for natural language investigations because the number of potentially predictive words is typically much greater than the number of trials in these studies, and it yields sparse classifiers that are more easily interpretable.

Fig. 1 shows the relative training performance of classifiers applied to 'new' reports (the CRM [correct rejections versus misses] classifier) and 'old' reports (the HFA [hits versus false alarms] classifier). The y-axis indicates the mean deviance of the model across the 10 folds of cross-validation. During each fold, the classifier is trained on 9/10's of the data, and used to predict the held out 1/10th. The deviance during this particular fold reflects the difference in fit (operationalized as the likelihood of the observed data given the model) between the tested model and a completely saturated model that perfectly recovers the observed data. Thus, greater deviance reflects poorer model fit. The y-axis of Fig. 1 indicates the average tendency of the trained classifier to misfit the hold out data at each level of constraint on its coefficients (λ) across the ten folds of cross validation. The size of the points reflects the number of words retained in the model at each level of constraint.

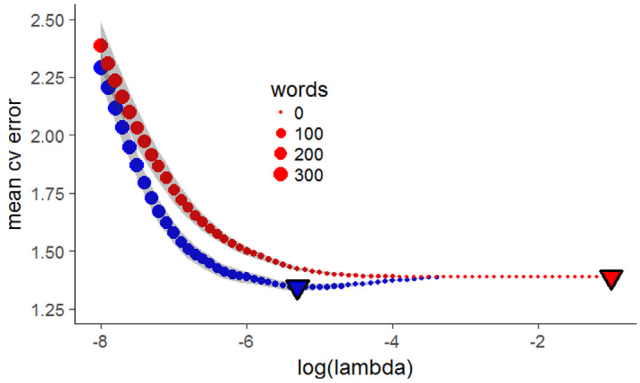

Fig. 1. Mean cross validation error during training for HFA (blue) and CRM (red) classifiers. x-axis shows the value of the constraint on logistic regression coefficients (λ) and y-axis shows the mean logistic deviance across 10 folds of cross validation. Size of points indicates the number of words remaining in model at each level of constraint. Inverted triangles indicate the minimum cross validation mean error. Shading reflects boostrapped 95% CI. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Regardless of which classifier one considers, the left-hand portion of the plot shows that as the constraint is increased the mean deviance decreases. This illustrates overfitting behavior. That is, the constraint increasingly prevents the model from chasing sampling error across folds, and as noted earlier, it also increasingly forces words from the model. Ideally, a minimum is reached, after which further imposing the constraint begins to harm generalization to the hold out samples, and so the mean deviance then begins to increase. This increase indicates that the constraint is now limiting the ability of the model to incorporate useful information for prediction. For the HFA model, this increase occurs at a log lambda of approximately -5, in which the model retains 19 predictive words (indicated by the inverted blue triangle). This is the retained model, because it jointly minimizes overfitting (initial decline of mean deviance) and information loss (subsequent rise in mean deviance). In contrast, the CRM model **never** reaches a minimum before all the words are forced from the model, which means that it is only capable of overfitting and therefore contains no useful predictive information.

These training data demonstrate, for the first time, that across individuals, introspective language more reliably distinguishes the accuracy of 'old' than 'new' recognition decisions. This is clear because at no point prior to all the words being forced from the model does the CRM classifier yield a reliably lower mean error than the HFA classifier, and because the minimum mean error of the HFA classifier is reliably lower than any mean error achieved by the CRM classifier. Moreover, since the accurate versus inaccurate trial counts were quite similar for 'new' (308 vs. 301) and 'old' (308 vs. 300) justifications, the relatively poorer performance of the CRM classifier is not the result of fewer cases upon which to train.

We next consider the content of the 'old' justifications that is predictive of recognition accuracy, a step referred to as 'feature inspection'. Because the features are words, this step provides the initial means of gleaning the psychological process or processes that render the language predictive of accuracy. Fig. 2 demonstrates the predictive words of the HFA classifier along with the value of their regression coefficients.

The HFA classifier appears to recover a distinction between a recollective versus non-recollective basis (R/NR) for the two outcomes.

[2] One can also use sequences of words (n-grams) as predictive features; for example, sequences of two or three words (bigrams and trigrams). However, this adversely increases the redundancy among the predictors. For example, the bigrams for the sentence 'i ate the apple' consists of i_ate, ate_the, and the_apple, which, along with the individual words, results in a highly correlated set of predictors, deleteriously affecting the stability of regression estimates.

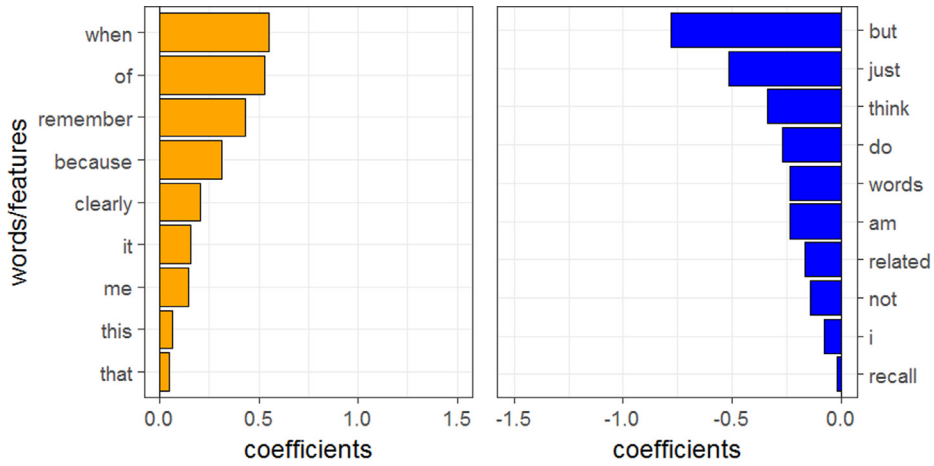

Fig. 2. HFA classifier logistic lasso solution. Left panel illustrates words predictive of hits whereas right panel illustrates words predictive of false alarms. In both, words are ordered from most predictive (top) to least predictive (bottom) as indicated by the coefficient values on the x-axis.

For example, hits are associated with the cognitive act 'remember', temporal information 'when', and self-reference ('me'). In contrast, false alarms are associated with equivocation ('but', 'think') or 'not' recovering supportive episodic information. They are also linked to present tense 'am', suggesting a greater focus on present tense may indicate greater reliance on reporting current feelings of fluency, familiarity or uncertainty versus describing recollection of prior episodic content. Consistent with this interpretation, Selmeczy and Dobbins (2014) found that 'am' was also predictive of medium versus high confidence hits.

Although the features of the HFA classifier seem to point towards a recollection/non-recollection (R/NR) distinction, further face validity for this interpretation arises from inspecting the raw justifications associated with high or low classifier output. For each trial, the classifier produces a numeric value indicating the estimated log-odds that the justification arose from the hit class. Thus, inspecting the justifications associated with the positive and negative extremes of log-odds estimates can help confirm the psychological interpretation of the selected features. For example, the following justifications received the highest and next highest log-odds (positive prediction values):

(1) ***When** thinking **of** a house, the first thing **that** popped into my head was the word key. In connecting the two words, **when** the word house appeared, I instantly thought **of** the word key which gave **me** a sense of familiarity **that** I had seen **this** word before.*
(2) *The reason why I chose **that** I have studied the word company is **because** I have a recollection of the word. I **remember** seeing the word and **it** simply stood out **when** I saw **it** the second time. GI haven* [error: given?] *the short amount of time I had to look at each word, I hope my mind is not playing tricks on **me**.*

And the lowest and next lowest log-odds (negative prediction values):

(3) *I think* I can recall seeing problem in the list, **but I am not** one hundred percent sure.
(4) *As **I am** studying the **words I am** viewing at the screen, **I am** remembering **words** that have a meaning relating to what **I am** doing or what I have done.*

Critically, these examples further converge on the conclusion the classifier isolates recollective from non-recollective bases for judgment. Moreover, the classifier correctly assigns justification (1) to the hit category even though it contains the phrase 'sense of familiarity' whereas it correctly assigns justification (3) to the false alarm category even though it contains the phrase 'i can recall'. This demonstrates that the 'bag of words' approach is able to achieve a fairly abstract distinction because it is sensitive to the collection of words present in the justification. Thus, even words that in isolation might suggest accuracy or failure do not dominate the classification when considered in context of the entire justification.

However, the raw texts suggest one potential concern in interpreting the classifier's behavior because the illustrative positive cases are clearly longer than the negative cases. This might suggest that the key distinction between hit and false alarm justifications is one of justification length rather than content per se. While length would still constitute an interesting predictive feature from introspective reports, it would perhaps require a different psychological interpretation of the classifier's behavior. We directly tested this possibility by comparing the lengths of the hit and false alarms justifications. The observed differences were not reliable ($t(6\,0\,6) = -1.2$, $p = .277$) despite the considerable sample size, and the mean lengths were highly similar for hits (M = 12.72) and false alarms (M = 12.02).

In an alternate test of the relevance of justification lengths, we used hierarchical logistic regressions to see if the addition of the justification lengths improved the predictive power of the logistic lasso estimates (produced by the HFA language classifier solution). In the initial model we confirmed that the log-odds estimates produced by the HFA classifier predicted the justifications classes (1 = hit, 0 = false alarm) ($z = 7.70$, $p < .001$). After this, we added the length of the justifications as an additional predictor. The classifier estimates remained similarly reliable ($z = 7.65$, $p < .001$) but the length predictor was not reliable ($z = -0.474$, $p = .635$). Thus, the length predictor does not provide a unique contribution, and it barely affects the performance of the language classifier estimates compared to when they were the sole predictor. When we directly compared the full and reduced models, they did not reliably differ ($\chi(1) = 0.225$, $p = .636$), confirming that the addition of justification lengths did not improve prediction. Thus, in total, there is no reason to assume that justification length per se carries additional information over and above the logistic lasso estimates, which instead are based on the presence of specific words and their combinations.

The next aspect of training performance we document is summarizing the overall accuracy of the retained HFA classifier. There are several ways to quantify discrimination accuracy, but here we use the receiver operating characteristic (ROC) of the classifier, with accuracy defined as the area under the curve (AUC) (Macmillan & Creelman, 2004). The ROC illustrates the discriminative value of the log-odds estimates produced by the HFA classifier by using increasingly lax log-odds values as the threshold for declaring that a justification arises from a hit. For example, the highest estimate produced by the trained HFA classifier was 2.82, with one hit justification and no false alarm justifications achieving this score. Thus, 1/308 or 0.0032 of the hit

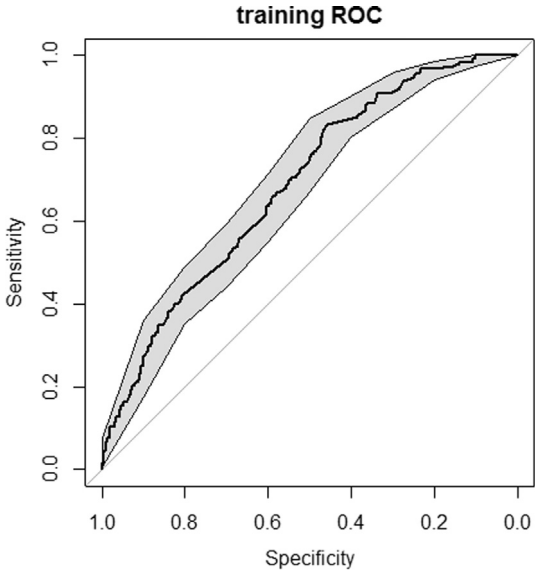

Fig. 3. Training ROC for HFA classifier. Shading illustrates a bootstrapped 95% CI around the ROC.

justifications are correctly identified, and none of the false alarm justifications (0/300) are incorrectly identified at this threshold. This is the first hit rate (y-axis) and false alarm rate (x-axis) point plotted on the ROC. The next log-odds estimate is 1.94 and at this threshold, two hit justifications are correctly identified (a new case and the prior case) and no false alarm justifications are incorrectly identified, and these proportions yield the next point on the ROC. This sequence proceeds until the entire ROC is traced, linking the cumulative hit and false alarm rates of the classifier as the log-odds criterion is relaxed. The points are then connected and the AUC is the area falling beneath this curve, with 0.5 representing chance discrimination. The ROC for the HFA classifier is shown in Fig. 3 along with its bootstrapped 95% CI (2000 iterations). The AUC was 0.69 with a 95% CI of 0.64 to 0.73. Thus, the classifier reliably discriminates the hit and false alarm classes.

The HFA classifier AUC, while reliable, may seem modest. However, this is consistent with the recollection/non-recollection interpretation. Under dual process recognition models, recollection will not occur for all hit justifications. Thus, if the classifier is distinguishing on the basis of recollection, then it will be limited by the frequency of recollection that occurs during hits which are randomly sampled for justifications on a small minority of trials. In generalization tests to experimental data drawn from paradigms in which recollection is well isolated and prevalent in one class of responses and not the other, the performance of the model markedly increases, as we demonstrate below.

The last aspect of training performance we consider is whether the HFA classifier's performance (i.e., log-odds estimates) covaries with individual differences in accuracy. To test this, we calculated the covariance between the log-odds model estimates and the categorical hit versus false alarm outcomes (hit = 1, fa = 0) for each subject, restricting the analysis to subjects providing the full four justifications for both hits and false alarms ($N = 66$). This covariance value is higher as the relationship between the log-odds estimates and the dichotomous outcomes increases for each subject. Additionally, we calculated the *d'* for each subject based on his or her total recognition data (i.e., including both justification and non-justification trials). These covariance and *d'* values were reliably correlated across subjects ($r = 0.46$, $t(64) = 4.10$, $p < .001$). Thus, the classifier's estimates covary more strongly with outcomes for subjects with high overall discrimination accuracy. Such an individual differences relationship makes sense, if one assumes that those who tend to have higher overall d' scores are those individuals with higher rates of recollective experiences.

3.3. Classifier transfer performance (Tests of Generalization)

The above analyses of the HFA classifier's training performance converge on the interpretation that it is sensitive to the relative presence versus absence of recollection in the justifications (R/NR interpretation). However, stronger tests of the validity of this claim require testing the trained classifier upon new data it has never encountered. These constitute true predictive (aka out-of-sample) tests because the model is fixed during training and then predicts classes of data to which it has never been exposed (Meehl, 1954). As we illustrate below, these predictions can demonstrate 'far transfer.' That is, a classifier trained to distinguish hit from false alarm justifications may also be able to distinguish other classes of memory justifications from new experiments with different procedures and requiring different judgments, provided the judgment classes in those experiments reflect the underlying psychological distinction captured by the model during its training. Thus, the patterns with which the HFA classifier does or does not generalize to new data speak to the convergent and divergent validity of the recollection/non-recollection interpretation (Campbell & Fiske, 1959).

We begin with convergent validity tests in which we applied the HFA classifier to the justifications of remember and know reports from Gardiner et al. (1998). If the classifier is sensitive to the recollection/non-recollection distinction, it should classify remember justifications as hits and know justifications as false alarms. To generate predictions, the new justifications are frequency coded for the presence of the words in Fig. 2, and the fixed weights in Fig. 2 are then applied to resulting counts, yielding a log-odds prediction for each encountered justification.

The resulting ROC from this generalization test is shown in Fig. 4a and yields an AUC of 0.92. Thus not only does the HFA classifier reliably distinguish remember from know justifications, it performs **better** than during its training, as revealed by the direct comparison of the ROC from this generalization test to the ROC from its training ($D = 8.13$, $p < .001$). Here, D indicates the observed difference in AUCs divided by a bootstrapped estimate of the standard error of these differences. The fact that the classifier generalizes provides convergent validity for the R/NR interpretation, but validity is further bolstered by the 'classifier gain'; that is, the increase in testing performance relative to training performance. To see why, it is important to note that under the recollection/non-recollection interpretation the classifier will be limited by the prevalence of recollection in one versus the other class it is attempting to distinguish. Its ability will thus be limited when discriminating hits from false alarms, since recollection will not occur for some proportion of hits that are instead based on familiarity/fluency (under dual process frameworks). Under this interpretation, out-of-sample testing on any recognition procedure that specifically isolates recollective from non-recollective trials, such as the Remember/Know paradigm, should yield a notable increase in performance because recollection is now more cleanly/exclusively isolated to the Remember versus Know classes, compared to the hit versus false alarm classes upon which it was trained. Thus, the fact that the HFA classifier generalizes to R/K justifications **and** performs better than during its training constitutes strong convergent validity for the R/NR interpretation.

The recollection/non-recollection (R/NR) interpretation also suggests a divergent validity test for Gardiner and colleagues' data. These researchers also collected justifications for guess reports. Under the R/NR account of the HFA classifier, its application to know and guess justifications should yield poor performance because neither know nor guess trials should be recollective. Fig. 4b confirms this prediction, demonstrating near chance discrimination of know and guess justifications (AUC = 0.58).

Critically, the transition in the HFA classifier's performance when applied to Remember versus Know justifications (Fig. 4a) versus Know versus Guess justifications (Fig. 4b) is essentially categorical. It excels at the former, but falls to near chance on the latter, which in turn suggests

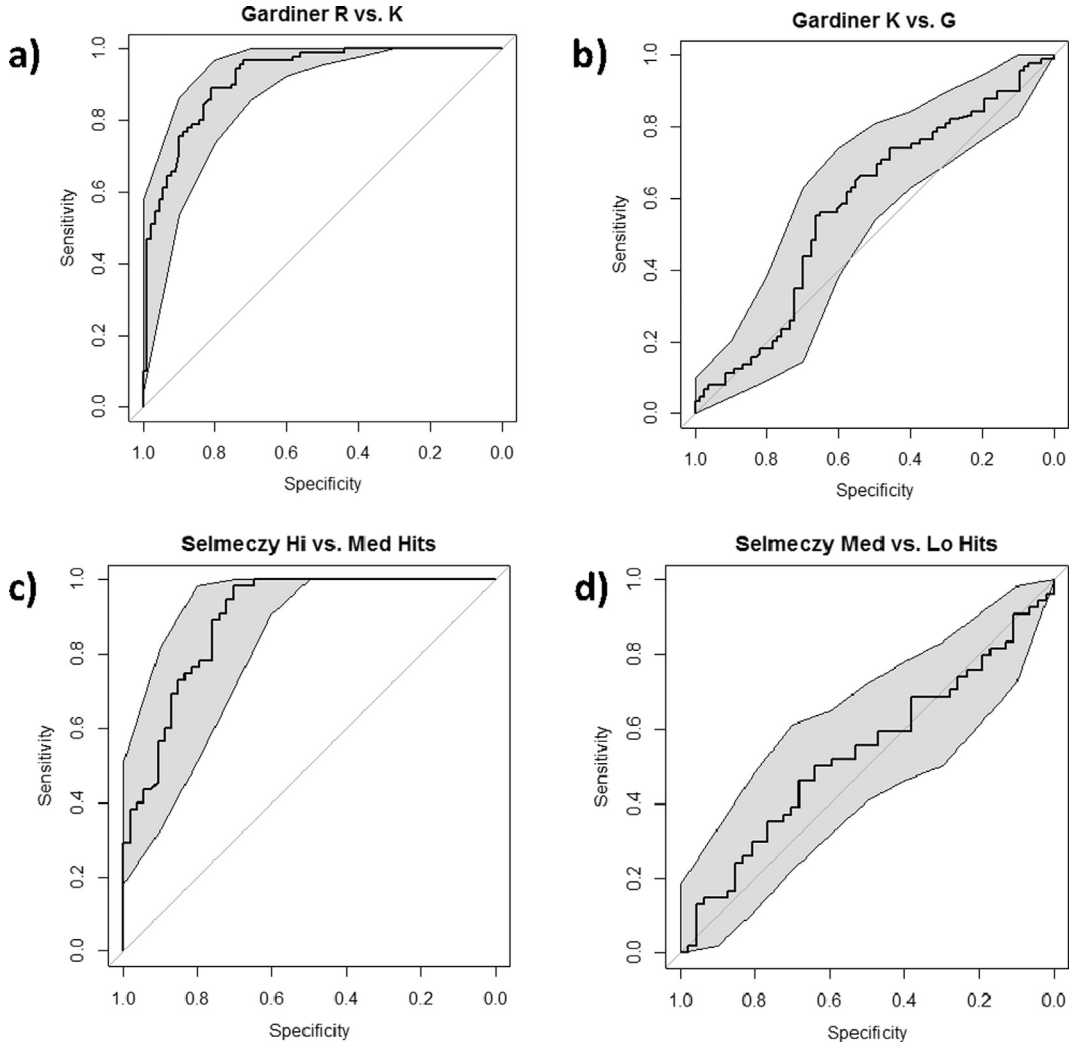

Fig. 4. Testing ROCs of HFA classifier applied to Remember, Know and Guess hit justifications of Gardiner et al. (1998) (panels a and b), or the High, Medium, and Low confidence hit justifications of Selmeczy and Dobbins (2014) (panels c and d). Panels on the left show the predicted high-test discrimination if the Recollection/Non-recollection (R/NR) interpretation of the trained HFA classifier is valid (convergent validity). Panels on the right show the predicted low-test discrimination if the R/NR interpretation is valid (divergent validity).

that a very specific distinction has been learned; namely, one that is useful for discriminating Remembering and Knowing justifications, but ineffective for discriminating Knowing and Guessing justifications. Unsurprisingly, the HFA classifier also well discriminates Remember and Guess justifications (AUC = 0.88, ROC not shown). In total, this demonstrates that the key information supporting its ability to discriminate lies solely within the Remember justifications, and this again supports the recollection/non-recollection interpretation of its behavior.

Finally, the HFA classifier demonstrates 'far transfer' of learning. The procedure on which it was trained did not entail Remember/Know reporting methods, instructions, or the other characteristics of the Gardiner and colleagues study noted in the Introduction, such as providing justifications in the presence of the investigators. In the current design, subjects were never asked to rank or classify the quality of their old or new conclusions in any manner before justifications were collected. From their perspective, some old and new conclusions are followed by justifications requests and most are not. They had no way of predicting which would be, and as noted in the methods, these requests follow a very small proportion of the total decisions (8.3%).

Repeating the logic of the convergent and divergent tests above, we next applied the HFA classifier to the high and medium confidence hit justifications of Selmeczy and Dobbins (2014). Under most dual process models it is assumed that recollection leads to high confidence during recognition (Dobbins, Kroll, Yonelinas, & Liu, 1998; Yonelinas, 2001). Thus, the HFA classifier under the R/NR interpretation should yield results similar to those above. Fig. 4c demonstrates that it does, producing an AUC of 0.89. Again, this is better than its training performance (D = 5.81, $p < .001$), demonstrating the classifier gain phenomenon and further supporting the R/NR interpretation.

Turning next to the justifications of medium versus low confidence hits from Selmeczy and Dobbins (2014), the R/NR interpretation again suggests the HFA classifier should fail, because neither medium nor low confidence responses should be associated with significant amounts of recollection. The classifier's performance confirms this prediction (Fig. 4d), yielding an AUC of 0.53. Finally, analogous to the R/K/G analysis above, when the HFA classifier was applied to high versus low confidence hit justifications, it performed well (AUC = 0.93; ROC not shown). This again demonstrates that the key recollection information to which the classifier is sensitive is selectively contained in one of the three classes of response, namely high confidence hit justifications.

In summary, the HFA classifier shows highly specific transfer of learning when applied to new justifications arising from Remember/Know and confidence ratings experiments. For High/Medium confidence hits, and Remember/Know hits, it discriminates better than during its original training, yielding classifier gain. This is noteworthy

because these are true predictive tests; there is no fitting taking place, just mechanical translation of word content into numerical log-odds predictions. Typically, these types of out-of-sample tests produce noticeably worse performance than observed during training, because models tend to overfit data during training (Copas, 1983). This 'shrinkage' in performance is the norm for all post-hoc model fitting approaches, including basic multiple regression analyses. In contrast, classifier gain reflects the exact opposite pattern, which is why it is both an unusual outcome and an important tool for construct validation. If the theory derived from the classifier's features and training performance correctly predicts that the learned distinction will be better isolated in the new testing data, it gains considerable support when classifier gain occurs. The recollection/non-recollection interpretation of the HFA classifier makes this prediction because recollection should be better isolated in Remember versus Know hits, and High versus Medium confidence hits, than in the comparison of overall hits and false alarms. Conversely, the R/NR distinction does not predict the ability to discriminate Know and Guess reports (since both lack recollection) or medium from low confidence hit reports (since both lack recollection). Thus, divergent validity was also demonstrated when the HFA classifier categorically failed to discriminate these classes of justification. Finally, the fact that the HFA classifier well discriminated Remember and Guess reports, and high and low confidence hits, considered in light of the other two tests, demonstrates that the key information upon which it relies is isolated to Remember and high confidence hit justifications. This also bolsters the recollection/non-recollection interpretation of the classifier.

3.4. Pitting reaction times against language

The findings above demonstrate that language can be objectively and reliably used to classify recognition memory outcomes, and that the HFA classifier finds a common language basis for distinguishing the justifications of (a) hits from false alarms, (b) high from medium confidence hits, and (c) remember from know hits. Traditionally, confidence or reaction times have been used as reliable indicators of response accuracy, and at first glance may seem more objective than the introspective reports gathered here. With respect to reaction time, Weidemann and Kahana (2016) recently demonstrated that above-chance ROCs can be created using reaction times as continuous predictors of accuracy because errors are generally slower than correct reports, and Dewhurst and Conway (1994) demonstrated that hits are faster when followed by Remember versus Know reports. We collected reaction times for the old/new judgments in conjunction with justifications, enabling us to test whether language-based classification is superior, redundant, or inferior to the classification of accuracy based on reaction time. We tested this in two manners.

First, we directly compared the log-odds estimates of the trained HFA classifier to those of a fitted reaction time model for the same reports.[3] We performed a simple (non-Lasso) logistic regression where reaction times were used to predict accuracy, yielding an RT model, and then we compared the output of this model to that of the HFA classifier to see if they differed in predictive accuracy. Both the HFA and the logistic regression RT classifiers output log-odds predictions that each trial reflects a hit, and their relative performance was contrasted via ROCs. Second, the log-odds predictions of each classifier were then directly pitted against one another in a multiple logistic regression to compare the degree to which they contain unique predictive abilities.

Fig. 5 shows the training ROC of the fitted, simple RT classifier, which yields an AUC of 0.57 and is just reliable with a 95% CI of 0.52 to 0.61. Statistically comparing this ROC with that of the HFA classifier shown in Fig. 3 demonstrates that the AUC is reliably higher for the language classifier ($D = 3.73$, $p < .001$). Table 1 shows the multiple logistic regression outcome when the log-odds estimates of both classifiers are jointly entered as predictors of the hit versus false alarm outcomes. The language predictions are clearly superior. For reaction time, a unit increase in the classifier's log-odds values yielded a 0.77 increase in the fitted log-odds of the justification arising from a hit. In contrast, for language, a unit increase in the classifier's estimate yielded a 1.57 increase. Since both predictors are scaled in the same units, the language estimates are more than twice as predictive as the reaction time estimates.

Interestingly, the fact that both the language and reaction time predictions were reliable in the multiple regression comparison indicates that they make unique contributions to the prediction of accuracy. Thus, the model in Table 1 is a kind of 'meta-classifier' that combines the predictions from classifiers trained using different domains of information. Since this meta-classifier itself yields log-odds estimates that each encountered instance is a hit, we can evaluate its ROC relative to the ROCs of the constituents. In particular, we can ask whether the addition of reaction time to language-based classification provides sufficient incremental gains in prediction to warrant its use. The ROC of the meta-classifier from Table 1 yielded an AUC of 0.69 (ROC not shown), in comparison to the HFA classifier in isolation, which had an AUC of 0.68. Direct comparison of these ROCs was not significant ($D = 1.52$, $p = .130$) which suggests limited utility for reaction times in the current data, but future work examining meta-classifiers incorporating reaction time as a separate source of information apart from language might nonetheless be useful.

One potential concern with the above approach is that it is entirely based on fitted data. The language predictions are based on the Lasso cross validation procedure, whereas the reaction time solution is based on a single predictor, using a form of least squares fitting without the need to impose a regularizing constraint. If the Lasso-based procedure were more prone to overfitting than the simple logistic regression then it might falsely demonstrate superior performance during training. We suspect this is unlikely, because cross-validation is specifically designed to limit overfitting, and because the generalization tests of the HFA classifier above demonstrate that it generalizes quite well to R/K and high and medium confidence hit justifications, a pattern that could not occur if it were grossly overfitting the training data.

Nonetheless, we addressed this concern by conducting a hold-out procedure using an approximately 2/3–1/3 split and pitting the trained language and RT classifiers against one another predicting the hold out sample. The language classifier demonstrated a greater testing AUC than the RT classifier, and when directly pitted against one another in predicting outcomes, only the language classifier was significant (see Supplementary Analyses).

As we further consider in the Discussion, there are several reasons why RT may be less predictive of accuracy than language. However, this does not mean that language is the sole predictor of accuracy; Table 1 illustrates reliable, unique contributions of language and reaction time for the entire training sample for 'old' decisions and raises the possibility that 'meta-classifiers', which combine language and non-language derived classifiers, may be useful in cognitive research.

4. Discussion

This is the first study to demonstrate that the introspective language of observers discriminates the accuracy of positive recognition conclusions, in a design that minimizes potential demand characteristics that could bias justification language across the considered classes. As noted in the Introduction, prior work examining introspective

[3] These constitute 'fair' tests. That is, reaction time and language models were developed and compared on the same trials and neither was altered as a function of any other information. For example, reaction times were not trimmed or normalized based on the subject's overall test performance since doing so would mean the reaction time model was based on more training information than the language-based model, which is necessarily limited to trials on which justifications were collected.

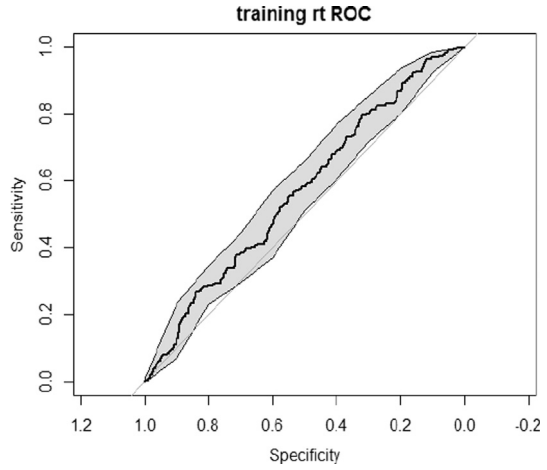

Fig. 5. Training ROC of logistic regression model using reaction times to predict accuracy. The model was trained on the same trials in which subsequent language justifications were collected.

Table 1
Direct comparison of Language and RT predictions using multiple logistic regression.

Reaction Time vs Language Classifiers	
	Dependent variable:
	Accuracy
Constant	−0.043 (−0.216, 0.130)
	p = 0.625
language	1.568 (1.145, 1.991)
	p = 0.000
rt	0.765 (0.149, 1.380)
	p = 0.015
Observations	599
Log Likelihood	−375.551

Note: values in parentheses are 95% CIs. rt = reaction time.

recognition language was potentially confounded by either providing extensive instructions about the assumed nature of conscious experiences during recognition and/or requiring metacognitive assessments before language introspections were collected (Gardiner et al., 1998; Selmeczy & Dobbins, 2014). This could lead observers to tailor their language or engage in differential retrieval strategies during the justification period in order to align their justifications to their public metacognitive statements or the researcher's instructions (e.g., 'Remember' or 'High Confidence'). If so, then the language would not capture the cause of the class distinction in question (e.g., high versus medium confidence hits) but would instead be the consequence of having already rendered the distinction.

Although asking for justifications, on even a small portion of trials, may generally alter the participants' approach to deciding recognition, this would not confound the key comparisons made in the current report because we restricted the classifier's training to discriminating accuracy outcomes having identical overt responses, namely 'old' or 'new'. Under this approach, any demand characteristics are fully matched across accurate versus inaccurate 'old' decisions, and accurate versus inaccurate 'new' decisions, and language could only differentiate these accuracy outcomes if the content of introspection systematically differs for accurate versus inaccurate memorial decisions.[1] Having minimized this serious potential confound, the HFA classifier in the current design demonstrates features consistent with a recollection/non-recollection (R/NR) distinction (Fig. 2 and illustrative texts 1–4).

Although suggestive (and intuitively appealing), feature inspection (viz., face validity) alone is an insufficient basis for understanding a classifier from a process or theory perspective. Additional support for the R/NR interpretation was gained by showing that the classifier's predictions improved with individual differences in recognition accuracy, and more importantly, through out-of-sample convergent and divergent validation tests that tested the construct validity of the R/NR interpretation. In the case of convergent validity, the HFA classifier discriminated Remember and Know hit justifications and High and Medium confidence hit justifications remarkably well (Fig. 4a, c). Indeed, ROC comparison demonstrated that the HFA classifier discriminated better during its testing on these new data than it did during its training, a phenomenon we term 'classifier gain.' Such gains are rare, as fitted models almost always demonstrate the reverse (shrinkage) pattern (Copas, 1983) due to overfitting sampling noise during training. Instead, classifier gain necessarily means that information to which the classifier is sensitive is better isolated in the testing than the training classes. This is predicted by the R/NR interpretation of the HFA classifier's performance because, under dual process models (e.g., Yonelinas, 2001), recollection should be better isolated to Remember and High Confidence hits than to hits generally. Thus the R/NR interpretation not only forecast that the HFA classifier should generalize to these data sets despite their different procedures (viz., far transfer), it predicted that the classifier should fare better than during its training.

The second, often overlooked aspect of construct validation is divergent validity (Campbell & Fiske, 1959). Divergent validity was established when the HFA classifier categorically failed to distinguish Know and Guess hit justifications, and likewise failed to distinguish Medium and Low confidence hit justifications (Fig. 4b, d). Since under most dual process models one would not expect recollection to be diagnostic across these classes, the discrimination failure makes sense, and because the HFA classifier discriminated extremely well in the convergent tests, it is clear that the divergent test null outcomes do not simply reflect low power of the classifier in general. Finally, the HFA classifier well discriminated Remember from Guess hit justifications and High from Low confidence hit justifications. The importance of this finding lies in the fact that when considered with the convergent and divergent tests mentioned above, it necessarily means that the key information upon which the HFA classifier depends resides in the High confidence and Remember hit classes of the generalization samples, which again converges on the R/NR interpretation. That is, if a classifier discriminates A from B, and A from C, but fails to discriminate B from C, then the key information it is using is selective to class A.

In total, the generalization behavior of the HFA classifier supports our initial interpretation of its features, and it is this generalization behavior, not the intuitive interpretation of the selected features, that represents a genuine test of the interpretation of the classifier. We emphasize this point because in discussing these data with other researchers, they have often suggested that a classifier trained using bigrams or tri-grams would be more valid because the selected features would be more meaningful. However, this confuses the ease with which features might be intuited by the experimenter (viz., face validity) with the utility of the model for generalizing and predicting outside the training sample (viz., construct validity). To illustrate, Fig. 6 shows the features that are selected when tri-grams are used as features during the HFA classifier's training. These clearly also intuitively support the recollection/non-recollection interpretation, and perhaps render this interpretation more easily than when one considers the single word features shown in Fig. 2.

Unfortunately, this ease of intuitive interpretation comes at the cost of generalization, as shown in Fig. 7. The figure displays the test ROCs for the HFA classifier trained on singletons (blue), bigrams (green), and trigrams (red) when applied to the Remember and Know hit justifications of Gardiner and colleagues. The AUCs clearly decline as the feature lengths increase, demonstrating decreasing predictive accuracy and indicating that the features discovered during training are becoming increasingly specific to the training subjects and/or the training experimental design. Since it is out-of-sample generalization that is key to establishing construct validity of the classifier's interpretation, we recommend that single word features be used for the actual classifier.

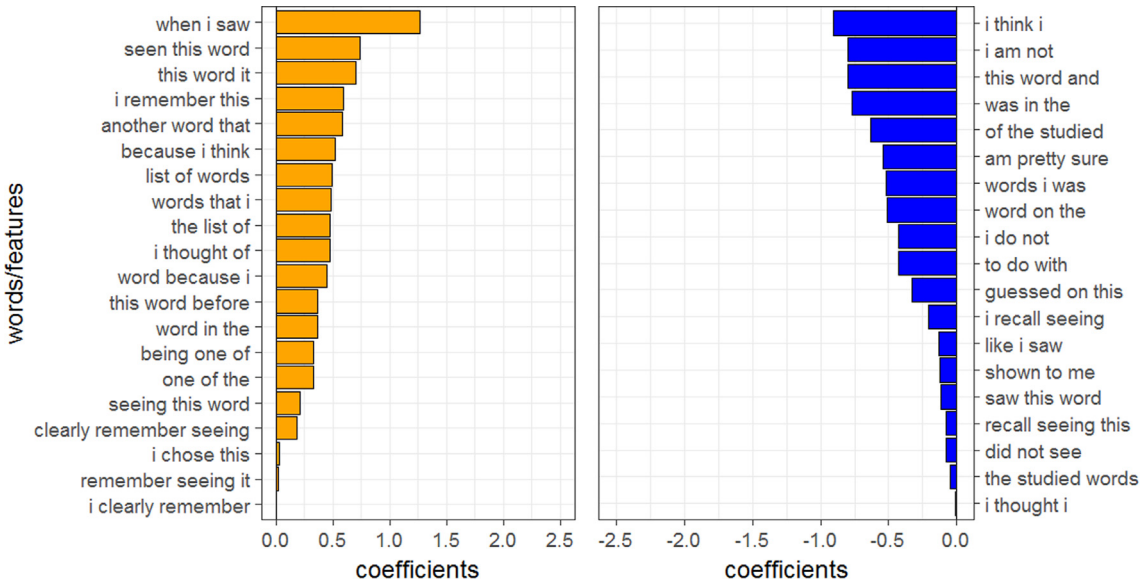

Fig. 6. HFA classifier logistic lasso solution and trigrams. Left panel illustrates words predictive of hits whereas right panel illustrates words predictive of false alarms. In both, trigrams are listed from most predictive (top) to least predictive (bottom) as indicated by the coefficient values on the x-axis.

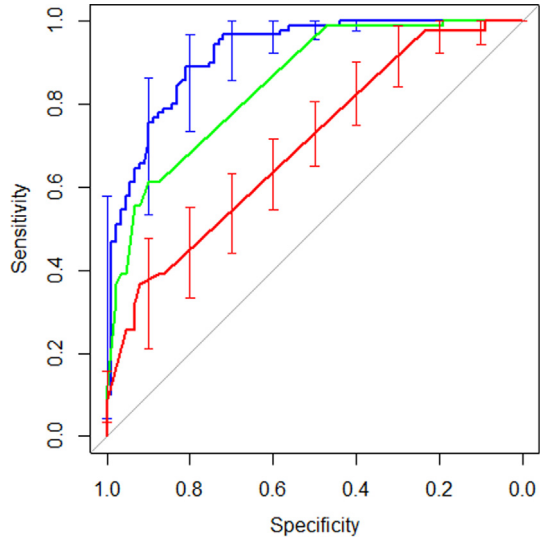

Fig. 7. Testing ROCs of HFA classifiers with different ngrams as features applied to the Remember and Know hits justifications of Gardiner and colleagues. Red is the ROC using trigrams, green bigrams, and blue singletons. Bars indicate bootstrapped 95% confidence intervals. CI for bigrams is omitted for clarity.

Nonetheless, bigram or trigram solutions, along with the examination of raw texts (e.g., justification examples 1–4 above), may serve as useful aids for arriving at the interpretation of the singleton classifier, which must then be tested using out of sample validation procedures.

4.1. Pitting language against other indices

To our knowledge, the current study is also the first to directly pit a language classifier against reaction times when classifying recognition accuracy. Under two different analyses, the data demonstrated that language was a superior basis for classifying the accuracy of 'old' reports. The direct comparison of language classifiers with other behavioral indices such as rated confidence and reaction time is an important new area for future research. Here, we briefly speculate as to why language may sometimes be a superior basis for classifying recognition accuracy, although future work is clearly needed in light of the preliminary nature of this finding. Our primary working hypothesis is that the description of recollective experiences is more universal and direct than reaction time (and perhaps other measures, such as confidence) because language is the primary means of conveying recovered memorial information. For example, reaction times are potentially subject to a host of non-episodic influences, such as item differences in lexical access, individual differences in motor speed and handedness, and trial-to-trial fluctuation in arousal. Many of these factors may have little to do with the type(s) of episodic information upon which the recognition decision will ultimately be made, and they are often unrelated to the experiences of the observer during encoding. Moreover, as shown in the Results when examining several raw texts, the multifeature 'bag of words' approach may itself confer considerable stability in that it is sensitive to the sum total of words within the text, such that any single word should not unduly influence the classification.

Speculation aside, it will be important in future work to pit other indices, such as numerical confidence ratings, against language. The current results (e.g., Table 1) suggest that in applied situations where the goal is maximizing classifier accuracy, the use of language, combined with other behavioral indices such as speed and confidence, may yield more effective 'meta-classifiers.' Moreover, the relative contributions of language and simple behavioral indices may change in theoretically important ways depending upon experimental design. For example, reducing the depth of processing or exposure duration at encoding may yield gains for the predictive value of reaction time relative to language. At retrieval, comparing two alternative forced choice to single item testing may also tip the predictive balance in favor of reaction times versus language in predicting decision accuracy. Also, the relative contributions of language, reaction time, or other behavioral indices may change with the types of materials tested. For example, whereas verbal materials may heavily encourage mental elaborations at encoding that are later introspectively reported (e.g., 'I remember thinking that…') during testing, pictorial information such as faces, may rely more on global assessments of familiarity at testing, rendering justification language perhaps less useful in comparison to reaction times.

4.2. Broader implications

In the current study we have focused on the relationship between language and recognition accuracy. However, the current techniques are relevant for testing any claim of differing explicit bases for

judgments across classes or contexts. For example, within the biases and heuristics framework originally developed by Kahneman and Tversky there is debate about the degree to which particular biases operate consciously or unconsciously (Gilovich & Griffin, 2002). This raises the interesting question of whether the heuristic mechanisms proposed in that framework might be detected in the natural language of observers when justifying their conclusions. These techniques may also be useful in the investigation of developmental trajectories, cognitive aging, and their relationship to the deployment of various assumed strategies during decision-making. Returning to memory, although the frequency of recollective experiences declines in healthy aging (e.g., McCabe, Roediger, McDaniel, & Balota, 2009), we do not know if the quality of justifications differs (provided recollection occurs). These procedures might also be useful in research on eyewitness accuracy, which poses a considerable challenge to prediction since each observer contributes just one observation. Given the current findings, it is possible that the language used in support of a lineup identification may convey predictive information over and above the speed and confidence of the selection. Note that the use of language and language classifiers need not be confirmatory, in that the failure to find reliable language differences across conditions theorized to depend on different explicit judgment strategies would call such theories into question.

4.3. Limitations

The current study strongly suggests that unguided introspective language identifies the contribution of recollection to the accuracy of 'old' recognition conclusions. However, the approach is clearly limited to discriminations that are supported by consciously apprehended content or strategies, and not applicable to discriminations arising from implicit processes. In this regard, the failure of the approach to discriminate the accuracy of 'new' recognition decisions suggests that the conscious bases for accurate and inaccurate conclusions of novelty are much more similar, and are perhaps indistinguishable (see also, Weber & Brewer, 2004). Regardless, the current study has several limitations that might be addressed with future methodological development.

For example, we restricted the collection of justifications to four instances of each potential outcome (hits, misses, correct rejections and false alarms) for each subject. This was done to reduce fatigue and experiment duration, since all justifications were hand-typed by the observers, but it necessarily weakens inferences at the individual differences level of analysis. However, the use of voice-recorded verbal justifications may enable a much larger percentage of justifications to be collected per participant. This approach appears feasible through the use of online transcription services or perhaps automatic speech-to-text translation via cloud services (Ziman, Heusser, Fitzpatrick, Field, & Manning, 2018). With a much larger proportion of justifications per subject it would be possible to tailor the classifier to each individual and, for example, pit his or her language against his or her reaction time or confidence in the prediction of outcomes. This in turn might lead to data suggesting that for some individual's language is most diagnostic whereas for others, confidence or reaction time is most diagnostic. Such findings would then raise the question of whether factors such as verbal intelligence are the mediators of such differences.

Additionally, the current justifications were collected from a single, undergraduate population, which raises interesting questions about the universality of the language used. The convergent validity tests clearly demonstrate that the language is not specific to a particular university setting, since all three data sets were collected at different universities over a range of 20 years, but there may be other cohorts or populations with different language to accuracy mappings. In addition, as noted above, one might expect introspective language to change across development, and research on aging memory may benefit by contrasting introspective language across younger and older populations to see if it differs in its relation to accuracy or confidence. Moreover, the machine learning approach illustrated here can be used to directly differentiate populations in terms of the language of justifying decisions, even if their language is similarly effective for predicting outcomes such as accuracy or confidence. In other words, there may be systematic language differences when justifying memory decisions that are not necessarily predictive of objective performance, but nonetheless is specific to particular cohorts or populations. Such language might reflect differences in beliefs about self-efficacy, for example, with older adults producing more qualifications or concerns during justifications even when they are equally accurate. For such questions, the classifiers would be trained to differentiate the populations directly, coding one population as 1 and the other as 0.

Finally, although we have demonstrated that the 'bag of words' classifier is capable of making a fairly abstract (and highly generalizable) distinction, there are presumably limits to this ability. For example, if researchers were specifically interested in intra-list reminding experiences (e.g., Jacoby & Wahlheim, 2013) then human raters would clearly be required. That said, having identified a large set of texts containing two types of particular experiences (with high interrater reliability), one could then potentially uses these texts to train a classifier that would automate the detection of such experiences in future research, eliminating the need for human coders.

5. Conclusion

The current study has demonstrated that the language of observers justifying recognition decisions distinguishes hits from false alarms, and does so because it reflects the psychological distinction between recollection and non-recollection. Since for both hits and false alarms observers make the same overt response (namely 'old'), the justification language distinguishing these outcomes is uncontaminated by different demand characteristics and hence directly conveys, at least in part, the causal basis of accuracy differences. The performance of such classifiers in different populations and conditions has considerable promise in testing the degree to which they differ in terms of explicit knowledge and strategy, and in clarifying the form of these contributions.

Acknowledgements

We would like to thank Falisha Kanji for help with data collection and text preprocessing.

Appendix A

Instructions taken from Gardiner et al. (1998)

Written Test Instructions:

In this test you will see a series of words, one word at a time. Some of the words are those that you saw yesterday. Others are not. For each word, click the YES button if you recognize the word as one you saw yesterday and click the NO button if you do not think the word was one you saw yesterday.

Recognition memory is associated with two different kinds of awareness. Quite often recognition brings back to mind something you recollect about what it is that you recognise, as when, for example, you recognize someone's face, and perhaps remember talking to this person at a party the previous night. At other times recognition brings nothing back to mind about what it is you recognise, as when, for example, you are confident that you recognise someone, and you know you recognise them, because of strong feelings of familiarity, but you have no recollection of seeing this person before. You do not remember anything about them.

The same kinds of awareness are associated with recognising the words you saw yesterday. Sometimes when you recognize a word as one you saw yesterday, recognition will bring back to mind something you remember thinking about when the word appeared then. You recollect something you consciously experienced at that time. But sometimes recognizing a word as one you saw yesterday will not bring back to

mind anything you remember about seeing it then. Instead, the word will seem familiar, so that you feel confident it was one you saw yesterday, even though you don't recollect anything you experienced when you saw it then.

For each word that you recognize, after you have clicked the YES button, please then click the REMEMBER button, if recognition is accompanied by some recollective experience, or the KNOW button, if recognition is accompanied by strong feelings of familiarity in the absence of any recollective experience. There will also be times when you do not remember the word, nor does it seem familiar, but you might want to guess that it was one of the words you saw yesterday. Feel free to do this, but if your YES response is really just a guess, please then click the GUESS button.

Lastly, when you are doing this test you may find it helpful to bear in mind that 30%/50% of the words are actually words that you saw yesterday.

Supplementary oral instructions

As usual in experiments of this kind, the written test instructions are followed by an oral briefing, in which the differences between the responses are explained again and then explanations and examples of them are elicited from the subjects, to check their understanding of them. The extent of this oral briefing is quite variable, depending on the particular individual. The following paragraphs exemplify what subjects were told at this time:

When you see these words today, if a word triggers something that you experienced when you saw it previously, like, for example, something about its appearance on the screen or the way it was spelled, or the order in which the word came in, I would like you to indicate this kind of recognition, by clicking the REMEMBER button. In other instances the word may remind you of something you thought about when you saw it previously, like an association that you made to the word, or an image that you formed when you saw the word, or something of personal significance that you associated with the word; again if you can recollect any of these aspects of when the word was first presented I would like you to click the REMEMBER button.

Instead, at other times you will see a word and you will recognize it as one you saw yesterday, but the word will not bring back to mind anything you remember about seeing it then, the word will just seem extremely familiar. When you feel confident that you saw the word yesterday, even though you do not recollect anything you experienced when you saw it, I would like you to indicate this kind of recognition, by clicking the KNOW button.

With know responses you are sure about seeing the word yesterday but cannot remember the circumstances in which the word was presented, or the thoughts elicited when the word was presented. With a guess response, you think it possible that the word was presented but you are not sure that it was. For some reason, you think there was a chance that the word was presented. Some people say "it looks like one of those words that could have possibly have been there." When you think your response was really just a guess, I would like you to click the GUESS button.

Supplementary material

The data and R processing scripts used in the current manuscript are available at https://osf.io/465vw/.

Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2019.05.025.

References

Augie, B. (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R Package Version 2.3. Retrieved from https://CRAN.R-project.org/package=gridExtra.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A, 33*(4), 497–505.

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological), 45*(3), 311–354.

Dewhurst, S. A., & Conway, M. A. (1994). Pictures, images, and recollective experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1088.

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), 1–21. https://doi.org/10.3758/BF03193807.

Dobbins, I. G., Kroll, N. E. A., Yonelinas, A. P., & Liu, Q. (1998). Distinctiveness in recognition and free recall: The role of recollection in the rejection of the familiar. *Journal of Memory and Language, 38*(4), 381–400. https://doi.org/10.1006/jmla.1997.2554.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24*(4), 523–533. https://doi.org/10.3758/BF03200940.

Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review, 111*(2), 524–542. https://doi.org/10.1037/0033-295X.111.2.524.

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (1998). Experiences of remembering, knowing, and guessing. *Consciousness and Cognition, 7*(1), 1–26.

Gilovich, T., & Griffin, D. (2002). Introduction-heuristics and biases: Then and now. *Heuristics and biases: The psychology of intuitive judgment* (pp. 1–18). Cambridge, UK: Cambridge University Press.

Hlavac, M. (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R Package Version 5.2.2. Retrieved from https://CRAN.R-project.org/package=stargazer.

Jacoby, L. L., & Wahlheim, C. N. (2013). On the importance of looking back: The role of recursive remindings in recency judgments and cued recall. *Memory & Cognition, 41*(5), 625–637. https://doi.org/10.3758/s13421-013-0298-5.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Retrieved from http://www.springer.com/us/book/9781461471370.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide.* New York: Taylor Francis.

McCabe, Geraci, L. D., Boman, J. K., Sensenig, A. E., & Rhodes, M. G. (2011). On the validity of remember–know judgments: Evidence from think aloud protocols. *Consciousness and Cognition, 20*(4), 1625–1633. https://doi.org/10.1016/j.concog.2011.08.012.

McCabe, D. P., Roediger, H. L., III, McDaniel, M. A., & Balota, D. A. (2009). Aging reduces veridical remembering but increases false remembering: Neuropsychological test correlates of remember–know judgments. *Neuropsychologia, 47*(11), 2164–2173.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.

Mickes, L., & Wixted, J. T. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117*(4), 1025.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231.

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Retrieved from https://www.R-project.org/.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 77. https://doi.org/10.1186/1471-2105-12-77.

Selmeczy, D., & Dobbins, I. G. (2014). Relating the content and confidence of recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 66.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267–288.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne, 26*(1), 1–12. https://doi.org/10.1037/h0080017.

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied, 10*(3), 156–172. https://doi.org/10.1037/1076-898X.10.3.156.

Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Open Science, 3*(4), 150670. https://doi.org/10.1098/rsos.150670.

Welbers, K., Atteveldt, W. V., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures, 11*(4), 245–265. https://doi.org/10.1080/19312458.2017.1387238.

Wells, G. L., & Bradfield, A. L. (1998). " Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*(3), 360.

Wickham, H. (2017). tidyverse: Easily Install and Load the'Tidyverse'. R Package Version 1.2.1, 1(1). Retrieved from https://CRAN.R-project.org/package=tidyverse.

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.

Yonelinas, A. P. (2001). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General, 130*(3), 361.

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. https://doi.org/10.1006/jmla.2002.2864.

Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., & Manning, J. R. (2018). Is automatic speech-to-text transcription ready for use in psychological experiments? *Behavior Research Methods, 50*(6), 2597–2605. https://doi.org/10.3758/s13428-018-1037-4.