Measurement of Relative Metamnemonic Accuracy

Aaron S. Benjamin and Michael Diaz

Introduction

Evaluating metamnemonic accuracy is an inherently difficult enterprise as the theorist must contend with all of the usual variability inherent to normal memory behavior and additionally consider other sources that are relevant only to the metamnemonic aspects of the task. This chapter reviews the arguments motivating the use of the Goodman-Kruskal gamma coefficient γ in assessing metamnemonic accuracy and pits that statistic against a distance-based metric d_a derived from signal detection theory (Green & Swets, 1966). We evaluate the question of which potential measures of metamnemonic accuracy have the most desirable measurement characteristics and which measures support the types of inference that researchers commonly wish to draw. In doing so, we attempt to make general arguments without providing a detailed account of the underlying mathematics or statistics, but we do place appropriate references should those interested desire a more technical treatment of the issues that arise.

T. O. Nelson was a pioneer of methodologies in the field and a consistent devotee of increasing analytical sophistication and rigorous measurement (see, e.g., Gonzalez & Nelson, 1996; Nelson, 1984). Although not all of the conclusions reached in this chapter are the same as those reached in Nelson's (1984) classic article, we would hope that the work nonetheless is considered a testament to Nelson's legacy of meticulous attention to the quantitative foundations of metacognitive research.

Metamemory Experiments

To begin, let us briefly review the basic substance of metamemory experiments, the data table, and the traditional analytic approaches. Be forewarned that the field is diverse and complicated, and any general portrayal of a metamemory experiment is bound to be a caricature at best. We do not mean to trivialize the many varieties of experiment that do not fit into the mold, but many, if not most, experiments share certain common characteristics:

1. *A manipulation of study or judgment conditions.* Many experiments evaluate metamemory in the context of a manipulation of memory. This manipulation may consist of an orienting instruction (e.g., generating vs. reading; Begg, Vinski, Frankovich, & Holgate, 1991); an ecological (e.g., altitude; Nelson et al., 1990) or pharmacological (e.g., benzodiazepines; Mintzer & Griffiths, 2005) intervention; use of item repetition (Koriat, Sheffer, & Ma'ayan, 2002), list position (e.g., recency vs. primacy; Benjamin, Bjork, & Schwartz, 1998), interference (Diaz & Benjamin, 2008; Maki, 1999; Metcalfe, Schwartz, & Joaquim, 1993), or scheduling (e.g., spacing between repetitions; Benjamin & Bird, 2006; Dunlosky & Nelson, 1992; Simon & Bjork, 2001; Son, 2004); or varying item characteristics (e.g., high- versus low-frequency words; Benjamin, 2003). The intent is to induce a difference in performance between conditions (although this is not necessarily the case), in order to evaluate the degree to which metamnemonic judgments reflect that difference. In other cases, populations of subjects (e.g., older and younger [Hertzog, Kidder, Powell-Moman, & Dunlosky, 2002]; memory impaired and memory intact [Janowsky, Shimamura, & Squire, 1989]), rather than items are compared. Alternatively, the study conditions may be held constant but the conditions of the metacognitive evaluation may be manipulated. Such manipulations might vary, for example, the timing (Nelson & Dunlosky, 1991) or the speed (Benjamin, 2005; Reder, 1987) of the judgment. Note that this aspect of the procedure is often, but not always, experimental: Items are randomly assigned to conditions, and the full force of experimental paradigms can be brought to bear on this part of the design.
2. *A measure of metamemory.* At some point prior to (Underwood, 1966), during, or after study (Arbuckle & Cuddy, 1969; Groninger, 1979), or even after testing (as in, e.g., feelings of knowing [Hart, 1965] or confidence in answers [Chandler, 1994]), subjects are asked to make a deliberate judgment about their memory performance. Mostly, those judgments are made on an item-by-item basis, but they may be for a group of items or for the entire set of items in the experiment. Alternatively, subjects may be asked to make a decision about restudying items (Benjamin & Bird, 2006; Son, 2004; Thiede & Dunlosky, 1999), and it is presumed that such decisions implicitly reflect their judgments of memory (Finn & Metcalfe, 2006). These judgments may take place within a context that allows an interrogation of memory, such as when only the cue term of a cue–target pair is used to elicit the judgment (Dunlosky & Nelson, 1992), or one in which such interrogation is difficult (e.g., if the entire cue–target pair is presented or if responses are speeded; Benjamin, 2005; Reder, 1987).
3. *A test of memory.* After some delay following the judgment procedure, memory is queried. It is rare (cf. Nelson, Gerler, & Narens, 1984) to employ an experimental manipulation at this point because it is uninformative to examine the effects of a manipulation on judgments that precede that manipulation. However, aspects of the test, particularly its relative difficulty, may play a role in evaluating metamnemonic accuracy.

Evaluating Metamemory Accuracy

Now, consider the fundamental question of metamemory experiments: How well does metamemory reflect memory? Metamemory is considered to be accurate when subjects show some sort of a calibrated assessment of their memory's failings and successes. Bear in mind that a useful measure of metamnemonic accuracy should be independent of actual levels of memory performance.

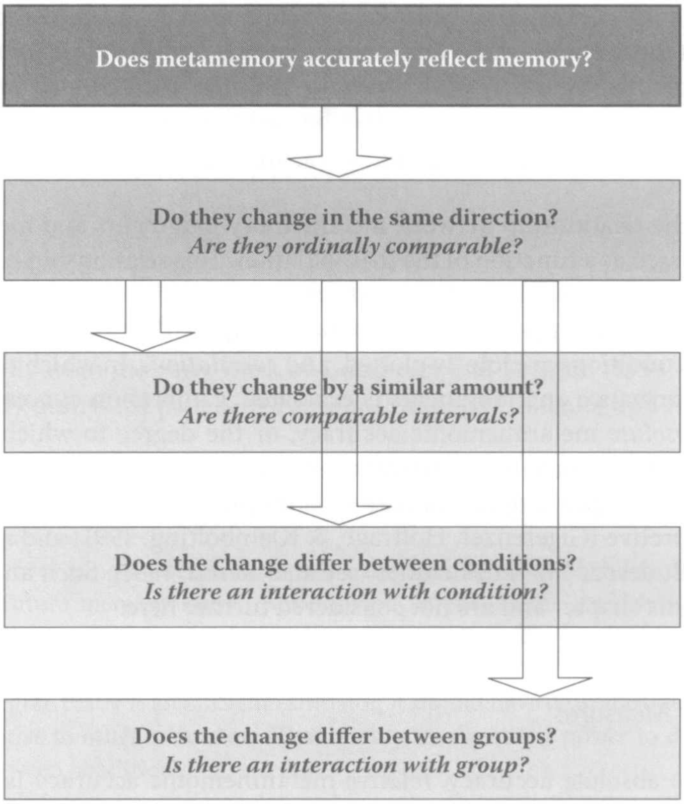

Does metamemory accurately reflect memory?
Do they change in the same direction?
Are they ordinally comparable?
Do they change by a similar amount?
Are there comparable intervals?
Does the change differ between conditions?
Is there an interaction with condition?
Does the change differ between groups?
Is there an interaction with group?

Figure 1 A taxonomy of questions about metamnemonic accuracy.

Figure 1 relates this fundamental question to the typical paradigm used to study metamemory and provides a rough taxonomy of questions ranked in order of measurement complexity. In rare circumstances, it might be informative to assess metamemory with reference to an absolute standard — for example, to evaluate whether a patient group reveals above-chance metamnemonic accuracy — but, more commonly, metamemory is tracked as a function of an experimental manipulation.

Ordinal Evaluation of the Experimental Factor

One straightforward analytic option is to jointly evaluate the effect of that manipulation on average memory performance and average metamemory judgments. Such paradigms are particularly powerful demonstrations when the effects of the variable are opposite for memory and metamemory (e.g., Benjamin, 2003; Benjamin et al., 1998; Diaz & Benjamin, 2008; Kelley & Lindsay, 1993; Metcalfe et al., 1993) but are limited by the inability to make interval-level comparisons between metamnemonic and mnemonic measures. This question is portrayed on the first sublevel of possible research questions in the hierarchy in Figure 1 to emphasize the minimal sophistication it requires on the part of the measurement scales: All that must be assumed is that higher scores indicate superior memory performance and a prediction of superior memory performance compared to lower scores. More complex demands are placed on those scales by the three questions that lie below this level.

Relationships Between Judgments and Performance

More often, the relationship between metamemory judgments and memory performance is assessed as a function of the manipulation. This relationship can be summarized in numerous ways, but the two most commonly used approaches are *calibration curves*, in which mean performance and mean judgments collapsed across a subset of items and conditions are jointly plotted, and *correlations*, in which the association between performance and judgments is evaluated. Calibration curves are used as a metric for *absolute* metamnemonic accuracy, or the degree to which mean rating values accurately estimate mean performance. Consequently, such analyses are only possible when ratings are made on scales isomorphic to probability scales and have certain interpretive (Gigerenzer, Hoffrage, & Kleinbolting, 1991) and analytic (Erev, Wallsten, & Budescu, 1994) difficulties (see also Keren, 1991). Such analyses are not the focus of this chapter and are not considered further here.

Correlational Measures

In contrast to absolute accuracy, *relative* metamnemonic accuracy is measured by the within-subject correlation of performance and predictions. Again, this assessment is usually made across conditions of a manipulation of memory. A good example is the delayed-judgment-of-learning effect (Nelson & Dunlosky, 1991), which is arguably the most robust and important effect in the metamemory literature. Nelson and Dunlosky (1991) showed that judgments about future recallability were much more highly correlated with later performance when a filled interval was interposed between study and judgments.

The consensual analytic tool for such paradigms is γ (Goodman & Kruskal, 1954, 1959), owing mainly to an influential article by Nelson (1984; see also Gonzalez & Nelson, 1996), in which γ was shown to be superior to a number of other measures of association, as well as to scores based on conditional probabilities and differences thereof (Hart, 1965), in terms of permitting a particular probabilistic interpretation of scores: What is the probability that Item X is remembered and Item Y is not given that Item X received a higher metacognitive judgment than Y?[1] Here, we reconsider that conclusion from the perspective of the three research questions at the bottom of Figure 1. For these cases, it is necessary to be in possession of data with relatively advanced metric qualities. To claim, for example, that a manipulation affects memory more than metamemory or that two groups who differ in baseline metamemory skills gain a differential amount from an intervention requires a measure that affords interval-level interpretation. The remainder of this chapter evaluates several candidate statistics for such qualities and reviews a solution based on the isosensitivity function of signal detection theory (SDT; e.g., Green & Swets, 1966; Peterson, Birdsall, & Fox, 1954; Swets, 1986a, 1986b). Nelson (1986, 1987) considered this alternative and rejected it, but we take a closer look at the debate, provide some supportive data for the SDT view with reanalyses of recent work, and demonstrate its metric qualities with simulated data sets. In addition, we show that a relatively simple transformation of γ improves its metric qualities and makes it comparable in certain ways to the measure derived from SDT.

Gamma and Its Use in Metamemory Research

Here, five major arguments in support of the use of γ are considered. These arguments derive primarily from the early work of Goodman and Kruskal (1959) as well as the psychologically motivated papers by Nelson (1984) and Gonzalez and Nelson (1996).

1. γ is easily generalized from the 2 × 2 case (in which it is equivalent to *Q*; Yule, 1912) to the *n* × *m* case. Thus, γ is appropriate when there are greater than two choices on the judgment scale.
2. Because there is no evidence concerning the form of the probability distributions relating future memory status (remembered or not) to the underlying judgment dimension, the machinery of SDT is unwarranted, and a purely nonparametric measure such as γ is preferred.
3. To the degree that γ is an efficient estimator, it should have desirably low error variance relative to other estimators. That quality increases the power to detect differences between conditions.
4. The γ coefficient bears a linear relationship to the probabilistic construal mentioned and thus has a transparent psychological interpretation in terms of subject performance (Nelson, 1984).
5. The γ coefficient is independent of criterion test performance, unlike other measures.

We shall consider each of these claims and revisit the adequacy of γ in light of the questions posed in Figure 1. Bear in mind that Nelson (1984) formulated these claims in the context of a search for a superior measure of feeling-of-knowing accuracy; here, we are more concerned with measuring metamemory more generally, and the prototype case we have in mind is in fact more like a typical judgment-of-learning (JOL) paradigm. It is not evident that this difference matters much.

Generalizability Across Experimental Designs

It is true that many alternative measures of association, such as phi, do not generalize coherently beyond the 2 × 2 case, and that such a limitation is undesirable for measuring metamnemonic accuracy. The γ coefficient is easily generalized to tables of arbitrary size, which makes it clearly superior in experiments in which predictions are more finely grained than binary ones. However, it is not clear that it is much of an advantage to be able to deal with more than two levels of the outcome variable; indeed, only the rare metamemory experiment has a memory outcome with more detail than "remembered" or "not remembered." In any case, the advantage of a measure that handles designs of $n \times m$ ($n,m \geq 2$) over one that effectively treats $2 \times m$ ($m \geq 2$) designs is likely minimal and may be offset by other relevant factors.

Signal Detection Theory Is Unsupported as an Analytic Tool

Unfortunately, it is not possible to do justice to the application of SDT to psychology in the limited space here (for further technical discussions, see Macmillan & Creelman, 2005; Wickens, 2001). Fundamentally, SDT relates performance in choice tasks to probability distributions of evidence conditionalized on the to-be-discriminated factor and decision criteria that partition that space into responses. Given the incredibly wide applicability of SDT to psychological tasks of detection and discrimination in perception (Swets, Tanner, & Birdsall, 1955), memory (Banks, 1970; Egan, 1958), and forecasting (Mason, 1982) and the impressive consistency of support across that wide array of tasks (Swets, 1986a), it certainly deserves a closer look in the case of metamemory. We do so and consider anew the unsupported assumptions pointed out by Nelson (1984, 1987).

Efficiency and Consistency

Measures derived from SDT have either lower error variance or usually lower error variance (that is, lower through a wide range of possible values) than does γ (Swets, 1986b, pp. 113–114). In addition, it has been noted that γ reveals disturbingly low levels of stability across alternative test forms, test halves, and even odd- and even-numbered items (Thompson & Mason, 1996; see also Nelson, 1988). Such low reliability calls into question experiments that fail to find differences between conditions, of which there are many.

A related question is whether γ is a consistent estimator — that is, whether the rate at which it approaches its asymptotic value with increasing sample size is as high as possible. Although we do not consider this property in detail, it is worth making note of one critical property of γ that is likely to influence consistency. As noted by Schwartz and Metcalfe (1994, Table 5.2), the fact that γ treats data purely ordinally — in terms of pairwise ranks — leads to both its desirable properties and perhaps some undesirable ones. A subject who assigns two item ratings of 5% and 95% probability of future recall is likely not making the same claim if the individual assigns those item ratings of 49% and 50%; yet, γ treats the cases equivalently. This property of γ is desirable only insofar as the prediction data are unlikely to have interval-level properties. Yet it discards vast amounts of information in treating them as purely ordinal. We will show that this treatment is overly conservative, and that relaxing that assumption only slightly affords the use of measures that may be more efficient and more consistent.

Psychological Interpretability

It is on the issue of psychological interpretability that much of our discussion centers. Nelson's (1984) argument about the clear relation between γ and the conditional judgment probability mentioned is a strong one, and we have no contention with the claim. However, we do question whether such a probabilistic interpretation affords the types of research questions and interpretations listed as the bottom three in Figure 1. That is, does the use of γ support interval-level analyses and conclusions? The answer is almost certainly no. At the very least, γ belongs to a class of measures (along with probability and other correlation measures) that are bounded on both ends. Measurement error leads to skewed sampling distributions at the margins of bounded scales and renders interpretation of intervals, and consequently interactions, difficult[2] (Nesselroade, Stigler, & Baltes, 1980; Willett, 1988). Schwartz and Metcalfe (1994) noted this problem in the context of between-group comparisons.

To be sure, this criticism is appropriately directed at a very wide range of analyses in the psychological literature (Cronbach & Furby, 1970), and we do not wish to imply any particular fault of researchers in metacognition. The important point is that equal intervals across a scale should not be assumed when treating psychological data, a point emphasized by Tom Nelson throughout much of his work. It is the burden of the theorizer to support such a claim prior to employing analyses that presume such measurement characteristics. To preview, it is on this very point that the application of SDT is most desirable. Measures of accuracy derived from SDT have interpretations rooted in geometry and are straightforwardly defensible as having interval characteristics.

Invariance With Criterion Test Performance

Nelson (1984, Figure 1) illustrated that γ, in contrast with a difference of conditional probabilities (Hart, 1965), was invariant with criterion test performance. However, Schwartz and Metcalfe (1994) noted that γ was not independent of the number of test alternatives in forced-choice recognition. Although we shall not consider the issue further here, it should be noted that γ may, under some conditions, vary with aspects of the task irrelevant to measurement of metamemory.

Signal Detection Theory and Metamemory Tasks

SDT provides an alternative solution to the question of how to summarize performance in contingency tables. The statistics of SDT are derived from a simple model of decision making under stimulus uncertainly, characterized by four basic assumptions (adopted from Benjamin, Diaz, & Wee, 2008):

1. Events are individual enumerable trials on which a signal is presented or not.
2. A strength value characterizes the evidence for the presence of the signal on a given trial.

3. Random variables characterize the probability distributions of strength values for signal-present and signal-absent events.
4. A scalar criterion serves to map the continuous strength variable onto a binary (or *n*-ary) decision variable.

For a metamemory task, it is assumed that stimuli that are later to be remembered (TBR) enjoy greater values of memory strength than stimuli that are later to be forgotten (TBF). The "memory strength" variable is really a variable by proxy; in fact, one of the great benefits of SDT is that, although an evidence axis needs to be postulated, it need not be identified. It simply reflects the evidence that can be gleaned from a stimulus regarding its memorability or, in this case, its perceived memorability.

To the degree that subjects can perform such a discrimination accurately — that is, if they can claim which items they will remember and which they will not at a rate greater than chance — then the distribution for TBR items must have generally higher values of memory strength than the distribution for TBF items. This is shown in the top panel of Figure 2. Evidence values (e_1 and e_2) are experienced by the subject and compared to a criterion *C*; in the case illustrated in Figure 2, the subject would reject the item yielding e_1 evidence and endorse the item yielding e_2 evidence.

SDT has been used primarily as a tool to aid in the separation of decision components of choice tasks from the actual sensitivity of the judgment. Sensitivity is a function of the overlap of the inferred probability distributions, and the placement of decision criterion (or criteria) represents the decision aspect of the task. As a theoretical device, *isosensitivity* functions can be plotted that relate the probability of a metacognitive hit (claiming that I will remember an item that will in fact be remembered later) to the probability of a metacognitive false alarm (claiming that I will remember an item that will not be remembered later). This function is a plot of how those values vary jointly as the criterion moves from a lenient position to a conservative one (or vice-versa). The bottom left panel for Figure 2 shows the isosensitivity function corresponding to the distributions in the top part of the figure in probability coordinates; the bottom right panel shows that same function in normal-deviate coordinates.

Empirical isosensitivity functions are useful in part because they allow one to evaluate whether the assumptions about the shapes of the probability distributions are valid. Specifically, normal probability distributions yield perfectly linear isosensitivity contours in normal-deviate coordinates, as shown in the bottom right panel of Figure 2 (Green & Swets, 1966). It has been claimed that the linearity of such functions is not a strong test of those assumptions because many different probability functions yield approximately linear forms (Lockhart & Murdock, 1970; Nelson, 1987). This is only partially true. Because the isosensitivity function is constrained to be monotonically increasing, there are many distributional forms that yield functions for which a large proportion of the variance (even above 95% in some cases) is linear. However, all forms except the normal distribution will lead to a nonlinear component as well. Consequently, an appropriate test is whether the addition of a nonlinear component to a linear regression model increases the quality of the fit. We present such a test and show that, contrary to the admonitions of Nelson (1987), SDT provides a viable model of the information representation and decision-making

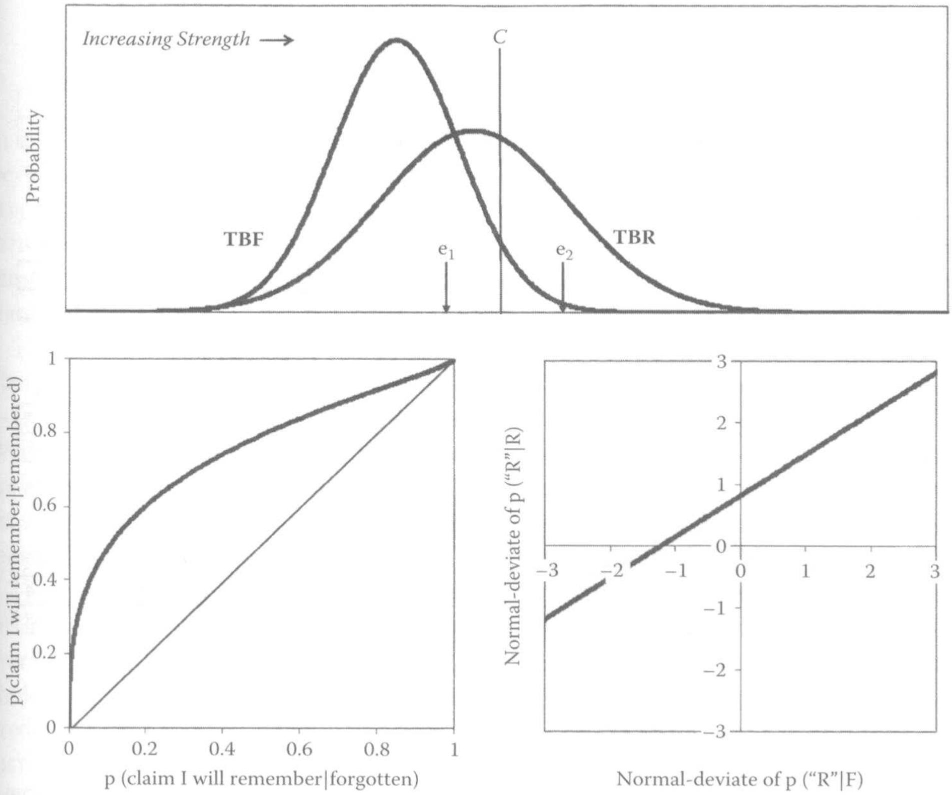
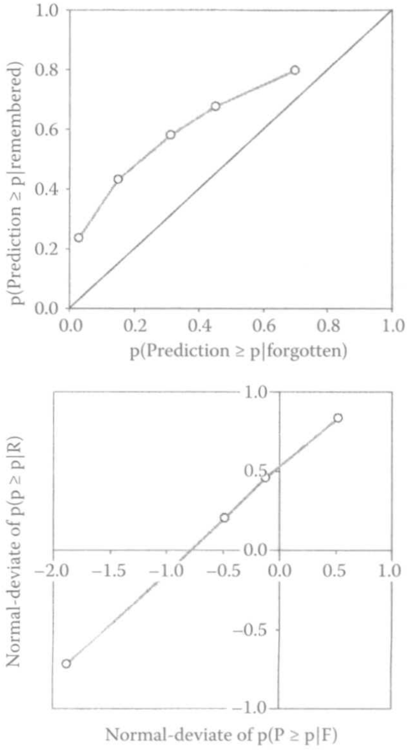


Figure 2 The signal detection theoretic framework and the isosensitivity function. Top panel: Normal probability distributions of strength for eventually forgotten (left) and remembered (right) items. e1 and e2 indicate possible values of experienced strength, or evidence, for future memorability. C indicates the location of a decision criterion. Bottom panels: Isosensitivity functions corresponding to the distributions shown in the top panel in probability coordinates (left) and normal-deviate coordinates (right).

process underlying metacognitive judgments. Let us first turn to the nitty-gritty of computing an isosensitivity function for metamemory data.

The Detection-Theoretic Analysis of a Metamemory Task

SDT analysis requires that our data be tabulated in the form of a contingency table. This requirement is straightforward in the case of a metamemory task, in large part because such a formulation is consistent with the computation of γ. Such a table is shown in the top right of Figure 3. Note that the data must be in a $2 \times m$ table in which there are m rating classes and two potential outcomes — presumably, remembered and forgotten. In the present example, there are six rating classes, with 1 indicating that the subject is very confident that they will *not* remember the stimulus and 6 indicating that they are very confident that they *will* remember it.



Frequency table

	1	2	3	4	5	6
Remembered	30	18	14	22	29	35
Forgotten	50	41	23	27	20	5
	80	59	37	49	49	40

Proportions

	1	2	3	4	5	6
Remembered	0.20	0.12	0.09	0.15	0.20	0.24
Forgotten	0.30	0.25	0.14	0.16	0.12	0.03

Cumulative Proportions

	≥1	≥2	≥3	≥4	≥5	≥6
Remembered	1.00	0.80	0.68	0.58	0.43	0.24
Forgotten	1.00	0.70	0.45	0.31	0.15	0.03

Normal-deviate scores

	≥1	≥2	≥3	≥4	≥5	≥6
Remembered	∞	0.83	0.46	0.20	−0.17	−0.72
Forgotten	∞	0.52	−0.12	−0.49	−1.03	−1.88

Figure 3 An example of how to estimate the isosensitivity function from data from a metamemory experiment.

Several additional transformations are necessary and are shown vertically on the right of Figure 3. First, frequencies are converted to proportions of each outcome class (shown in the second table on the right side of Figure 3). Those proportions are cumulated from right to left across the rating scale, such that the sixth cell in a row contains the proportion of 6 responses, the fifth cell in a row contains the proportion of a 5 or a 6 response, and so on. These cumulative proportions are treated as increasingly liberal response criteria, and a joint plot of those values yields the isosensitivity function shown in the top left of Figure 3. Note that the most liberal point is always going to be (1,1) since it reflects the cumulative probability of *any* response. The final data table shows the cumulative proportions after an inverse-cumulative normal transformation (i.e., changing from proportions to z scores) and yields the normal-deviate isosensitivity plot shown in the bottom left.

The sensitivity of the ratings can be understood as either the degree to which the theoretical distributions overlap, as mentioned, or as the distance of the isosensitivity function from chance performance, indicated in the top function as the major diagonal and in the bottom function as an unshown linear contour passing through the scale origin. We introduce one measure d_a that corresponds to the shortest possible distance from the origin (scaled by $\sqrt{2}$) to the isosensitivity function in the bottom plot. That value can be easily computed:

$$d_a = \frac{\sqrt{2}y_0}{\sqrt{1+m^2}}$$

in which y_0 and m represent the y-intercept and slope, respectively, of the normal-deviate isosensitivity function. The d_a can be conceptualized in terms of the geometry of the isosensitivity function, as defined above, or in terms of the distributional formulation in the top part of Figure 2; in that case, d_a is the distance between the means of the normal distributions divided by the root-mean-square average of their standard deviations.

Using d_a to measure metamemory accuracy is a novel suggestion to our knowledge. There was some consideration of whether *d′* — a similar but not equivalent measure – is an appropriate score to measure metamnemonic accuracy (Nelson, 1984, 1987; Wellman, 1977). The *d′* measures the distance between the probability distributions scaled by a common standard deviation. The assumption of common variance has proven incorrect in most substantive domains (Swets, 1986a) but is nonetheless commonly used because it can be computed on the ubiquitous 2 × 2 data table. At least a 2 × 3 table is required for d_a, and its fit is only testable with a minimum of four columns. Such a characteristic is hardly a limitation in metamemory research, however; it simply implies that subjects' rating scale must contain more than two discrete choices. In fact, it is more commonly necessary to construct judgment quantiles from prediction data to *reduce* the number of points in isosensitivity space (and thus also increase the precision of the estimates). In the next section, we directly address the question of whether the SDT model of metamnemonic judgment is an accurate one.

Analyses of Metamemory Tasks

Nelson (1984) wrote, "Unfortunately, there is no evidence in the feeling-of-knowing literature … to justify the assumption that the underlying distributions are normal" (p. 121). In this section, we present such evidence. We consider two data sets. The first is from our recent work (Diaz & Benjamin, 2008), for which the prediction task is on a scale of 0 to 100, and the criterion task is cued recall. For the second data set (Benjamin, 2003), the prediction is on a 1-to-9 scale, and the criterion tasks are both recognition and free recall. We have deliberately chosen tasks that differ substantively in order to demonstrate the robustness of the analysis.

Analysis of Diaz and Benjamin (2008)

These experiments involved multiple study–test trials with paired-associate terms, over which proactive interference was introduced by reusing cue terms. One condition is reported here in which there were 20 items per studied list (henceforth, the difficult condition), and another condition is reported in which there were 10 or 16 items per list (the easy condition).[3]

TABLE 1 An Example of How to Compute Quantile Frequencies Under Conditions With Tied Boundary Scores

Data Table

JOL	0	20	40	40	40
Recall	0	1	0	1	1

Frequency Table

	Q1	Q2	Total
Remembered	**1** + 0.5(2/3) = 1.33	2.5(2/3) = 1.67	3
Forgotten	**1** + 0.5(1/3) = 1.17	2.5(1/3) = 0.83	2
Total	2.5	2.5	5

Because the prediction data were on a 0-to-100 scale, the first step was to convert those data to quantile form. To get a reasonable estimate of the isosensitivity function, there should be a sufficient number of bins to estimate the shape of the function adequately (at least four and ideally five or more) and a sufficient number of observations to avoid very low frequencies in any particular bin. A good rule of thumb is to have subjects try to distribute their judgments more or less evenly across the rating scale and to try to have no fewer than 20 of each rating. In this case, the number of discrete ratings was actually greater than the number of observations, so it was necessary to convert the data to quantiles.

For each subject, individual matrices of performance and JOLs were sorted by JOL magnitude and divided into six bins. The goal was to have each bin contain an equal number of items and to partition those items by whether they were eventually recalled (or recognized). Because the total number of items was not always divisible by six, the column totals were not always integers. In addition, because of numerous ties on the JOL variable, some interpolation was necessary. Table 1 gives a simple example of how this was done. In this example, there are five total items to be divided into two bins. Thus, the marginal total for each (column) quantile bin must be 2.5. Because there are three remembered and two forgotten items, the row totals are also fixed.

In the first quantile, there is one item that is remembered, one that is forgotten (those values are in bold in the table) and half of an item remaining with a value that must be interpolated from the remaining tied scores. Because only one of those three tied scores represents a forgotten item, one third of the remaining half item is allocated to the forgotten bin and two thirds are allocated to the remembered bin. Similarly, for the second quantile, all of the members are tied and lie on the bin boundary. Thus, of the 2.5 total items, one third is allocated to the forgotten bin and two thirds to the remembered bin.

Parameters for the SDT model were estimated individually for each subject using maximum likelihood estimation (Ogilvie & Creelman, 1968). Linear regression accounted for a mean of 97.2% and 96.4% of the individual subject's data in the easy and difficult conditions, respectively. The addition of a quadratic term increased the mean variance accounted for to 99.1% and 98.7%, respectively; this increase was

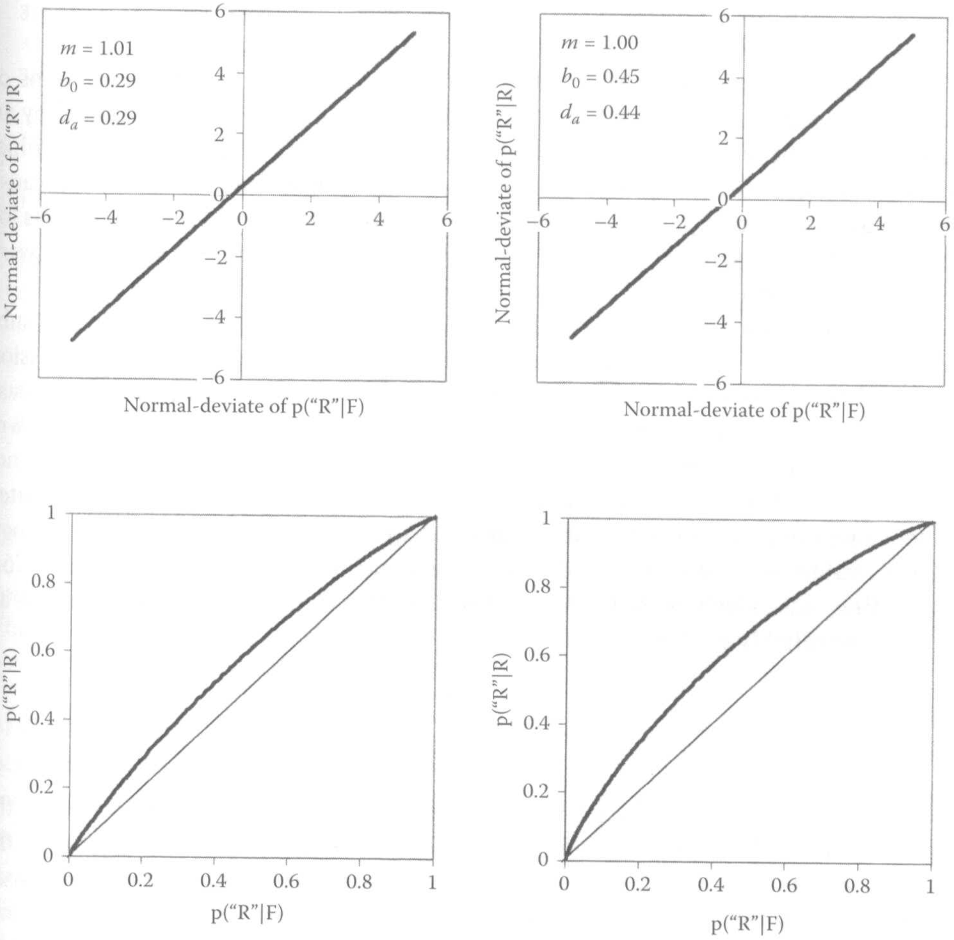


Figure 4 Isosensitivity functions in probability (top) and normal-deviate (bottom) coordinates for the difficult (left) and easy (right) conditions drawn from Diaz and Benjamin (2008).

reliable in only 2% of the subjects in each condition.[4] This value is lower than the chance probability of 5%. In addition, the mean value of the quadratic term in the full model was not reliably different from 0 in either condition. These findings suggest that the assumption of normally distributed evidence holds in these data.

Average isosensitivity functions based on the mean parameters of the linear model across subjects are shown in Figure 4. These data reveal that metamemory performance is in fact superior in the easy condition. The d_a values shown in Figure 4 are for the average functions shown in the figure; mean d_a values based on individual subject performance were similar but revealed an even larger difference (d_a [easy] = 0.51, d_a [difficult] = 0.25). The difference between conditions was reliable (t [169] = 4.23) and confirmed a similar result obtained using γ ($\gamma_{\text{easy}} = 0.32$, $\gamma_{\text{difficult}} = 0.19$; t [169] = 3.49), but with a larger effect size.

Analysis of Benjamin (2003)

In this experiment (Benjamin, 2003, Experiment 3), subjects made predictions of recognition performance on a 1-to-9 scale, took a test of recognition followed by an additional prediction phase for a test of recall, and then took the recall test. Unlike the case just described, frequencies did not need to be interpolated. However, because performance was so high on the recognition test, there were a number of subjects for whom the fit of isosensitivity functions could not be evaluated; those subjects were dropped from the analysis of the shape of the function.

Linear regression accounted for a mean of 84.7% and 85.8% of the individual subject's data in the recognition and recall conditions, respectively. Quadratic regression increased the mean fit to 89.3% and 93.6%, respectively. Despite the larger increase than in the previous analysis, the magnitude of the increase was reliable in only 3% of the cases. As before, the mean value of the quadratic term in the full model was not reliably different from 0 in either condition. The assumption of normally distributed evidence was thus supported in this data set as well.

Mean values of d_a were 0.44 and 0.51 for recognition and recall, respectively. Corresponding values of γ were 0.29 and 0.38. Neither difference was reliable, but all values were reliably different from 0.

Scale Characteristics of d_a and γ

The analyses reported in the previous section indicate that the application of the machinery of SDT to the traditional metamemory task is valid and thus permits the use of d_a as a measure of metamemory performance. Because d_a is rooted firmly in the geometry of the isosensitivity function, it has interpretive value as a measure of distance and all of the advantages that such an interpretation affords: equal intervals across the scale range and a meaningful 0. Like actual distance, d_a is bounded only at 0 and ∞.[5]

Let us now return to the question of the metric qualities of γ. We claimed that γ could not have interval-level properties because of its inherent boundaries. In the next section, we simulate data based on the confirmed assumptions that were tested and evaluate exactly how well γ performs and whether simple transformations are possible that increase its metric qualities. The strategy we use to evaluate γ and other measures is to generate data based on a population profile with a known metric space and then test the ability of γ, d_a, and other measures to recover that metric space. We use the assumption of normal probability distributions to generate simulated metamemory strengths for recalled and unrecalled items and apply different measures of metamemory accuracy to assess performance in those simulated data.

Simulations

For each of 1,000 sim-subjects, memory performance on 100 test trials was simulated by randomly sampling profiles from a normal distribution with a mean of 50 and

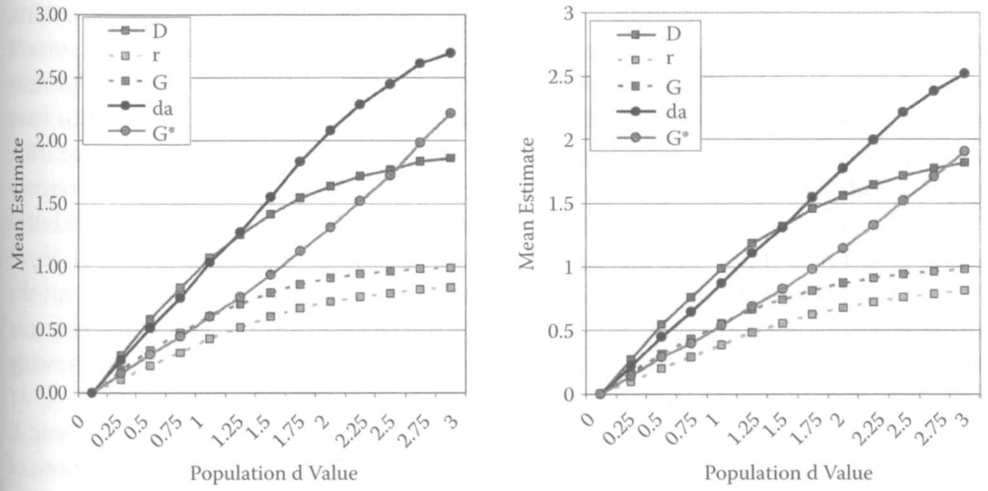


Figure 5 Estimates of r (the Pearson correlation coefficient), D (the Hart difference score), γ (the Goodman-Kruskal gamma correlation), and d_a (a distance measure based on signal detection theory) as a function of the distance between generating distributions. The degree of linearity of the function reveals the potential of the statistic for use in drawing interval-level inferences on data. Left panel: Signal variability = 1. Right panel: Signal variability = 1.5.

variance of 10. The profile represented the number of items recalled out of 100 for each sim-subject. Then, for each unremembered item, an evidence score was drawn from a normal distribution with mean 0 and variance 1, and for each remembered item an evidence score was drawn from a normal distribution with mean d and variance s. These scores were transformed into confidence ratings by relation to three criteria that were set for most simulations to lie at the mean of the noise distribution, the mean of the signal distribution, and halfway between the two. This transformation produced a matrix of memory scores (0 or 1) and confidence ratings (1, 2, 3, or 4) that was used to estimate the values of several candidate metamemory statistics, including γ, d_a, r (the Pearson correlation coefficient), and D (the difference in mean judgments between recalled and unrecalled items; Hart, 1965).

Results

The first important set of results can be seen in Figure 5, in which each statistic is plotted as a function of d (with $s = 1$ in the left panel and $s = 1.5$ in the right panel). The major diagonal indicates perfect recovery of the parameter d. Several general patterns are evident. First, the correlation measures suffer, as expected, near the boundary of the scale and exhibit a decided nonlinearity. Second, differential variability in the strength distributions (shown in the right portion of the figure) decreases the overall fit of all measures and results in estimates that are biased to be low. Because our estimates of the variability of the signal distribution were in fact quite close to 1, we consider more closely here the case in the left panel.

Because the two correlation statistics r and γ have probabilistic interpretations, they should not be expected to fall on the major diagonal. However, the important aspect of the failure of these measures is the clear nonlinearity. If a statistic is a linear transformation of the population value, then the estimator can be claimed to have interval-level properties. As noted, the boundary on r and γ introduce nonlinearity; consequently, a linear fit accounts for only 91% and 85% of those functions, respectively. The much-maligned Hart difference score statistic D fares better than γ but is also limited by a functional asymptote due to the judgment scale range (89%). However, it performs admirably over a limited range of performance. D_a outperforms the other statistics substantially at 98% linearity, and its failures lie only at the extreme end of the performance scale. D_a is thus the most promising candidate for drawing interval-level inferences from metamemory data.

The correlation measures suffer on this test because of the boundaries at –1 and 1. Thus, to test those measures more fairly, we additionally consider transformations of r and γ that remove the compromising effects of those boundaries. One commonly used function that serves this purpose is the logit, or log odds, which is defined as

$$Logit\ X = \log\left(\frac{X}{1-X}\right)$$

This function only operates validly on positive values; thus, rather than use G, we use the transformation of γ that Nelson (1984) called V and is presented in our footnote 1. Here, we define G^* as the logit of that value. It is related to γ as follows:

$$G^* = \log\left(\frac{\gamma+1}{1-\gamma}\right).$$

The linearity of the relationship between the candidate measures G^* and r^* (the equivalently transformed Pearson correlation coefficient) and the population value from which the data were generated was assessed. This transformation increased the fit of a linear relationship from below 95% to over 99% for both measures under both simulation conditions. It thus appears as though G^* (and r^*, for that matter) is a promising candidate for evaluation of interval-level hypotheses. However, several characteristics are noteworthy. First, G^* is $-\infty$ when $\gamma = -1$ and ∞ at $\gamma = 1$ (i.e., when performance is perfect), which means that it is quite unstable at the margins of performance. The untransformed measure γ does not have this unfortunate property, but this is the price that is paid by the conversion to a more valid measurement scale. Second, it allows for no obvious and immediate interpretation in terms of behavior or theory, although this disadvantage is mitigated by its easy translation to and from γ.

Several other conditions were simulated to assess the robustness of these effects. When the criteria are placed in either nonoptimally conservative or lenient locations, the fit of d_a is decreased by an order of magnitude smaller amount ($\Delta R^2 = 0.003$) than is γ ($\Delta R^2 = 0.03$), but both d_a and G^* are equally linear (~99%). Adding variance to the signal distribution increases linearity slightly; this general effect likely reflects the well-known advantage of rendering the frequency distribution of ratings more uniform. In all cases, d_a, G^*, and r^* all provide excellent fits (~99%). When the numbers of items and subjects are reduced to more validly approximate conditions of a typical experiment on metamemory (20 items for 20 subjects, with a mean performance of 10 and variance of 3), all fits suffer, but r^* outperforms all others (~97%) with G^* not far behind (~95%). Under conditions of relatively low or high mean memory performance (mean of 20 or 80 items remembered out of 100), none of the statistics (d_a, G^*, or r^*) shows an appreciable drop in fit.

The bottom line of these simulations is that the greater linearity of d_a extends over a great variety of conditions, and that a logit transformation of V improves its linearity significantly. The superiority of d_a should not be surprising given that the data were generated using assumptions that are built into signal detection theory. However, the robustness of the effect, as well as the poor performance of γ and quite impressive performance of G^*, should be surprising. It would appear that γ is a poor choice of a statistic for use in interval-level comparisons, such as those indicated in the bottom three lines of Figure 1. Either G^* or d_a should be used in experimental designs that invite interval-level comparison.

Turning to the question of measurement variance, γ fares much better. In fact, across all of the simulated conditions described above, the coefficient of variation (COV; a ratio of the standard deviation to the mean) was consistently lowest for γ. This is especially true at high levels of metamemory performance ($d > 2$). There are three important caveats to this finding. First, it is difficult to know to what extent the boundary at 1 on γ influences this effect. However, this concern has limited practical implications. More worrisome, there is a marked heteroskedasticity in estimates of γ as a function of d, and this effect has the potential to lead to analytic complications. In addition, it appears that at least some of that variability may be legitimate individual-difference variability that is lost by γ: Reducing memory variance in the simulations to 0 reduces (but does not eliminate) the advantage of γ over d_a in terms of COV. It does thus appear that the types of noise introduced in the simulations described here lead to greater variability in estimates of d_a than γ. This finding merited a closer look at empirical comparisons of the two measures.

Empirical Comparisons of Coefficient of Variation

The smaller COV in γ than d_a could reflect an oversimplification in the simulation or an empirical regularity. If it is in fact an empirical regularity, then it might temper our enthusiasm for d_a somewhat. We reexamined the data from Diaz and Benjamin (2008) and Benjamin (2003) and estimated the COV across both experiments. For the Diaz and Benjamin (2008) data, the estimates were equivalent (COV = 0.98). For the Benjamin (2003) data, COV for recognition was lower using d_a (1.25) than γ (1.56), but slightly higher for d_a (0.77) than γ (0.72) on the recall test. This result confirmed the claim that the superiority of γ in the simulations was a combination of devaluing individual-difference variability and the marked simplification of the generating process yielding rating data. Overall, the measures appear to be more or less equivalent in terms of COV.

Summary

Here, we have taken a closer look at the question of what types of measures might best support the types of inferences researchers wish to draw using metamemory data. In doing so, we have taken advantage of the theoretical framework of signal detection theory (Green & Swets, 1966; Peterson et al., 1954) and evaluated whether data from two metamemory experiments (Benjamin, 2003; Diaz & Benjamin, 2008) were consistent with the assumptions of that framework. Because those assumptions were strongly supported, we have advised that d_a and measures like it (MacMillan & Creelman, 2005; Wickens, 2001) can profitably be used as measures of metamemory. Using SDT, we have made our assumptions about the process of making metamemory judgments as explicit as possible. Using data simulated on the basis of those confirmed assumptions, we have shown that γ is unlikely to have those desirable interval-level characteristics, and we thus advise against its use when interactions, between-group comparisons, and across-scale comparisons are used. An alternative is to use G^*, which is a simple monotonic transformation of γ (or r^*, which is the equivalent transformation of Pearson's r), which appears to have superior measurement characteristics. However, these statistics suffer from certain characteristics as well: They are highly variable at their extremes, and they do not have an obvious or transparent interpretation in terms of subject behavior (like γ) or psychological theory (like d_a). Nonetheless, one possibility is to use γ except in analyses that require interval-level data and use G^* for such analyses. The disadvantages of such an approach relative to the use of d_a and signal detection theory are minimized.

With these recommendations, there are a few important details to keep in mind when estimating the isosensitivity function from metamemory data. First, there must be a reasonably large number of both remembered and unremembered items. When there is not, the probability of empty cells in the frequency table is undesirably high, and the isosensitivity function may be underdetermined. This recommendation should be familiar as γ is also notably unstable when there are not sufficient numbers of remembered and unremembered items. Ideal performance is at 50%.

Second, it is important that subjects use the full range of the judgment scale. This recommendation is much more important for the isosensitivity function than for γ because estimating that function takes advantage of the ordering of judgments (i.e., that $1 < 2 < 3 < 4$), whereas γ evaluates judgments only on a pairwise basis. Subjects should specifically be instructed to use the full range of the rating scale if the isosensitivity function is to be estimated.

Third, the rating scale should have at least four options. Bear in mind that m options lead to a curve with $m - 1$ points, and that subjects who perform particularly well or particularly poorly may yield fewer than $m - 1$ usable points. In addition, if the assumption of normal probability distribution functions is to be tested as part of the analysis, then there must be sufficient points to fit and test a quadratic function (i.e., > 3). In that case, the rating scale should have at least five options. We recommend the use of a semicontinuous scale, like the subjective probability scale described in Diaz and Benjamin (2008) and the quantile estimation procedure developed in this chapter and depicted in Table 1. This technique deals well with individual differences in scale use that are more difficult to rectify with a scale with fewer options.

For researchers who wish to evaluate the differential effectiveness of a manipulation on metamnemonic accuracy, either within or between groups, it is critical to have in hand a dependent measure that can be defended as having interval-level properties. The measure reviewed here, d_a, has such qualities to a much greater degree than does the commonly used γ, and we hope that the review provided here helps researchers better evaluate their measurement options and use d_a fruitfully in appropriate cases or use an appropriate transformation of γ under the necessary conditions.

Notes

1. Nelson called the value associated with this interpretation V, and it is related to γ by the following relationship: $V = 0.5\gamma + 0.5$.
2. Remember that "crossover" interactions, which require only an ordinal interpretation, are not subject to such a concern, as noted here.
3. The difficult condition corresponds to Experiment 1 in Diaz and Benjamin (2006) and the easy condition to Experiment 2. Both data sets reported here include additional versions of the experiments not reported in that article.
4. Model fit was tested as,

$$F = \left(\frac{\Delta R^2}{1 - R^2_{full}} \right) \left(\frac{N - K_{full} - 1}{K_{full} - K_{reduced}} \right)$$

in which N represents the number of data points (the number of points on the isosensitivity function) and K the number of parameters in each model (in this case, three in the full model and two in the reduced model). There were five points on the isosensitivity function for all but 6 subjects who had false alarm rates of 0 or hit rates of 1 for one rating range. Those subjects were omitted from this analysis because the F ratio was indeterminate. The test distribution was thus F (1, 1) with $\alpha = .05$, two tailed.

5. Strictly speaking, d_a is bounded at $-\infty$ and ∞ because the mean of the signal distribution can theoretically lie to the left of the mean of the noise distribution. However, values less than 0 reveal below-chance performance and thus should only arise because of measurement noise or perverse subject behavior.

References

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81,* 126–131.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74,* 81–99.

Begg, I., Vinski, E., Frankovich, L., & Holgate, B. (1991). Generating makes words memorable, but so does effective reading. *Memory & Cognition, 19,* 487–497.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition, 31,* 297–305.

Benjamin, A. S. (2005). Response speeding mediates the contribution of cue familiarity and target retrievability to metamnemonic judgments. *Psychonomic Bulletin & Review, 12,* 874–879.

Benjamin, A. S., & Bird, R. D. (2006). Metacognitive control of the spacing of study repetitions. *Journal of Memory and Language, 55*, 126–137.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*, 55–68.

Benjamin, A. S., Diaz, M., & Wee, S. (2008). *Signal detection with criterial variability: Applications to recognition memory.* Manuscript under review.

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition, 22*, 273–280.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74*, 68–80.

Diaz, M., & Benjamin, A. S. (2008). *The effects of proactive interference (PI) and release from PI on judgments of learning.* Manuscript under review.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20*, 374–380.

Egan, J. P. (1958). *Recognition memory and the operating characteristic.* USAF Operational Applications Laboratory Technical Note No. 58–51.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*, 519–527.

Finn, B., & Metcalfe, J. (2006). *Judgments of learning are causally related to study choice.* Manuscript in preparation.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin, 119*, 159–165.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Associations, 49*, 732–764.

Goodman, L. A., & Kruskal, W. H. (1959). Measures of association for cross classifications: II. Further discussions and references. *Journal of the American Statistical Association, 54*, 123–163.

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

Groninger, L. D. (1979). Predicting recall: The "feeling-that-I-know" phenomenon. *American Journal of Psychology, 92*, 45–58.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56*, 208–216.

Hertzog, C., Kidder, D. P., Powell-Moman, A., & Dunlosky, J. (2002). Aging and monitoring associative learning: Is monitoring accuracy spared or impaired? *Psychology and Aging, 17*, 209–225.

Janowsky, J. S., Shimamura, A. P., & Squire, L. R. (1989). Memory and metamemory: Comparisons between patients with frontal lobe lesions and amnesic patients. *Psychobiology, 17*, 3–11.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*, 1–24.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77*, 217–173.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*, 147–162.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74*, 100–109.

Macmillan, N. A., & Creelman, C.D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Maki, R. H. (1999). The role of competition, target accessibility, and cue familiarity in metamemory for word pairs. *Journal of Psychology: Learning, Memory, and Cognition, 25*, 1011–1023.

Mason, I. (1982). *On scores for yes/no forecasts.* Paper presented at the Ninth Conference on Weather Forecasting and Analysis, Australian Meteorological Society (pp. 169–174), Seattle, WA.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 851–864.

Mintzer, M. Z., & Griffiths, R. R. (2005). Drugs, memory, and metamemory: A dose-effect study with lorazepam and scopolamine. *Experimental and Clinical Psychopharmacology, 13*, 336–347.

Nelson, T. O. (1984). A comparison of current measures of feeling-of-knowing accuracy. *Psychological Bulletin, 95*, 109–133.

Nelson, T. O. (1986). ROC curves and measures of discrimination accuracy: A reply to Swets. *Psychological Bulletin, 100*, 128–132.

Nelson, T. O. (1987). The Goodman-Kruskal gamma coefficient as an alternative to signal-detection theory's measures of absolute-judgment accuracy. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 299–306). New York: Elsevier Science.

Nelson, T. O. (1988). Predictive accuracy of the feeling of knowing across different criterion tasks and across different subject populations and individuals. In M. M. Gruneberg (Ed.), *Practical aspects of memory: Current research and issues* (pp. 190–196). New York: Wiley.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science, 2*, 267–270.

Nelson, T. O., Dunlosky, J., White, D. M., Steinberg, J., Townes, B. D., & Anderson, D. (1990). Cognition and metacognition at extreme altitudes on Mount Everest. *Journal of Experimental Psychology: General, 119*, 367–374.

Nelson, T. O., Gerler, D., & Narens, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General, 113*, 282–300.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin, 88*, 622–637.

Ogilvie, J. C., & Creelman, C. D. (1968). Maximum-likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology, 5*, 377–391.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory, PGIT-4*, 171–212.

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology, 19*, 90–138.

Schwartz, B. L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 93–113). Cambridge, MA: MIT Press.

Simon, D. A., & Bjork, R. A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 907–912.

Son, L. K. (2004). Spacing one's study: Evidence for a metacognitive control strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 601–604.

Swets, J. A. (1986a). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99*, 181–198.

Swets, J. A. (1986b). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99*, 100–117.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1955). *The evidence for a decision making theory of visual detection.* Technical Report No. 40, University of Michigan, Electronic Defense Group.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1024–1037.

Thompson, W. B., & Mason, S. E. (1996). Instability of individual differences in the association between confidence judgments and memory performance. *Memory & Cognition, 24*, 226–234.

Underwood, B. J. (1966). Individual and group predictions of item difficulty for free learning. *Journal of Experimental Psychology, 71*, 673–679.

Wellman, H. M. (1977). Tip of the tongue and feeling of knowing experiences: A developmental study of memory monitoring. *Child Development, 48*, 13–21.

Wickens, T. D. (2001). *Elementary signal detection theory.* London: Oxford University Press.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, pp. 345–422). Washington, DC: American Education Research Association.

Yule, G. U. (1912). On the methods of measuring the association between two attributes. *Journal of the Royal Statistical Society, 75*, 579–652.