To appear in: P.A. Higham and J.P. Leboe (Eds.) *Constructions of Remembering and Metacognition: Essays in Honor of Bruce Whittlesea*. Basingstoke: Palgrave-MacMillan.

Quantity-Accuracy Profiles or Type-2 Signal Detection Measures?

Similar Methods Toward a Common Goal

Morris Goldsmith

University of Haifa

Contact: Morris Goldsmith

Department of Psychology

University of Haifa

Haifa, Israel

Email: mgold@research.haifa.ac.il

FAX: 972-4-8249431 Phone: 972-4-8249745

I. Introduction

In his chapter, Phil Higham (this volume) follows up on previous work showing how methods based on type-2 signal detection theory (SDT) can be used to study the strategic regulation of memory performance, and compares some of the advantages and disadvantages of this approach to Koriat and Goldsmith's (1996b) Quantity-Accuracy Profile (QAP) methodology. I am glad to have the opportunity to comment on some of the points made in that chapter. I am also glad to be able to participate in this volume honoring Bruce Whittlesea, who among his many significant contributions to the study of memory, has done much to emphasize the critical role played by post-retrieval evaluation and decision processes in remembering (e.g., Whittlesea, 2002; Whittlesea & Williams, 2001a, 2001b).

That emphasis is shared by the type-2 SDT framework described by Higham in his chapter, and by the metacognitive framework that Asher Koriat and I originally developed (Koriat & Goldsmith, 1996b) and subsequently extended (Goldsmith & Koriat, 2008) for studying the strategic regulation of memory reporting, and more generally, for conceptualizing and assessing the contributions of metacognitive monitoring and control processes to memory accuracy and quantity performance. Thus, before addressing some of the specific differences between the two approaches, I would first like to stress the common view of memory that is fundamental to both frameworks. The basic assumption is that in the process of remembering, people do not simply spill out all of the information that comes to mind. Rather, between the retrieval of information on the one hand, and overt memory performance on the other, lie metacognitive monitoring and control processes that are used to strategically regulate the accuracy and quantity of the information that is reported. Hence, memory performance under free-report conditions—conditions typical of real-life remembering, in which one has the option to respond "don't know'—depends not only on the ability to access and retrieve the solicited information, but also on the ability to effectively monitor the likely correctness of that information and choose an appropriate control policy (report criterion) based on competing incentives for accuracy and informativeness. These metacognitive contributions to free-report memory performance have been demonstrated in simulation analyses (Higham, this volume; Koriat & Goldsmith, 1996b) and in empirical studies examining the effects of various manipulations and group (e.g., developmental) differences, and how these effects and differences are mediated (see Goldsmith & Koriat, 2008 for review).

Once the potential contributions of post-retrieval monitoring and control processes to memory performance are acknowledged, it is crucial to have a way of isolating and examining these contributions. The QAP method that we proposed (Koriat & Goldsmith, 1996a, 1996b) is based on a two-stage procedure including both forced and free responding: Participants are required to provide a best-candidate answer to each recall cue/question or recognition probe (forced report), together with a confidence rating reflecting the assessed probability that the answer is correct, but are also allowed to decide whether or not they want to volunteer the answer (free report), typically under an incentive structure in which points are gained for each correct volunteered answer, a penalty is paid for each incorrect volunteered answer, and there is no penalty—but also no gain—for withheld answers. Using this procedure, the various components contributing to free-report quantity and accuracy performance can be isolated (see Goldsmith & Koriat, 2008 for more details):

1. *Retrieval/retention* is evaluated in terms of forced-report proportion correct.
2. *Monitoring effectiveness* is evaluated in terms of:
    a. *calibration bias* (under- or overconfidence)—the difference between mean assessed-probability-correct and actual forced-report proportion correct;

b. *monitoring resolution*—the extent to which the confidence judgments distinguish between correct and incorrect best-candidate answers. This relationship that can be indexed by various measures, including but not limited to the within-participant Goodman-Kruskal gamma correlation (recommended by Nelson, 1984), the adjusted normalized discrimination index (ANDI; recommended by Yaniv, Yates, & Smith, 1991), and *d'* or d_a (recommended by Higham, 2002; Masson & Rotello, 2009).

3. *Control policy* (report criterion) is estimated in terms of P_{rc} (report criterion probability)—the assessed probability that best reflects the level above which answers are reported/volunteered, and below which they are withheld.
4. *Control sensitivity*—the extent to which the report/withhold decisions are in fact based on the monitoring judgments, indexed in terms of the within-participant (gamma) correlation between confidence and the volunteer/withhold decisions.

With respect to this list of component measures, I now address some points of agreement and contention between the QAP and type-2 SDT methodologies for assessing the memory and metamemory components contributing to free-report memory performance.

II. Retreival/Retention

Let us begin with a point of agreement: Both approaches use the forced-free paradigm just described, and then use forced-report percent correct to index the contribution of "memory" per se, untainted by the contributions of post-retrieval monitoring and report control. This use, however, deviates from the standard (type-1) SDT approach, in which two basic parameters are derived, one (*d'*, *A'*, etc.) reflecting "true memory" and the other (β, B''_D, etc.) reflecting the contribution of "extraneous" decision processes. Because three measures (at least) rather than two are needed to isolate the components contributing to free-report performance, the application of type-2 SDT in this context requires that some adjustments be made, both conceptually and methodologically, in line with the QAP approach.

III. Monitoring Effectiveness

With regard to the effectiveness of monitoring, the QAP approach distinguishes between the absolute aspect, captured in terms of calibration bias (over- or underconfidence), and the relative aspect, captured in terms of resolution or discrimination accuracy. This distinction has a long history within the literature on judgment and decision making (e.g., Lichtenstein, Fischhoff, & Phillips, 1982; Yates, 1982), and has been adopted in the metamemory literature as well (e.g., Nelson, 1996; Schraw, 2009). In contrast, the type-2 SDT approach conceptualizes and measures monitoring effectiveness solely in terms of its relative aspect—resolution or discrimination accuracy. With regard to this aspect, there has been a great deal of discussion recently regarding the advantages and disadvantages of different measures, some associated with the SDT framework (e.g., Benjamin & Diaz, 2008; Masson & Rotello, 2009; Rotello, Masson, & Verde, 2008), adding to previous discussions of similar issues (e.g., Liberman & Tversky, 1993; Nelson, 1984, 1996; Schraw, 1995; Yaniv, Yates, & Smith, 1991; Yates, 1982). In the context of comparing the QAP and type-2 SDT approaches, this appears to be a relatively minor issue. As noted above, the QAP method is not tied to any particular measure, and there is nothing to preclude the use of SDT-based measures to supplement or replace other measures (e.g., gamma), if indeed these turn out to be the better choice.

The different treatment of the absolute aspect of monitoring, however, appears to constitute a fundamental difference between the two approaches. In fact, by the type-2 SDT approach, over- or underconfidence is not treated as an aspect of monitoring, but rather, of "control," captured by differences in the confidence criteria used to assign explicit numeric or linguistic scale values to specific levels of subjective evidence ("true" confidence). Clearly, however, differences in elicited subjective probability values, and hence in over- or underconfidence, are not merely a matter of scaling differences (Wallsten & Budescu, 1983; Wallsten, Budescu, & Zwick, 1993). The source of the problem appears to be that in type-2 SDT, much of what is conceived as "monitoring" in the metacognition literature is hidden away from view in the presumed mapping function that translates psychophysical evidence (or "cues," e.g., retrieval latency and fluency, cue familiarity, accessibility of supporting and competing information, and so forth; see Brewer & Sampaio, 2006; Koriat, 2008; Wixted & Stretch, 2004) into subjective confidence. Thus, theoretically interesting differences in absolute monitoring are expressed as "distribution shifts," in which there is an overall increase or decrease in the true subjective confidence levels attached to incorrect and/or correct answers. Yet, distribution shifts are notoriously difficult to distinguish from actual changes in response criteria (e.g., Starns, Lane, Alonzo, & Roussel, 2007; Wixted & Stretch, 2000), as both of these are picked up as changes in response bias parameters, and therefore, differences in monitoring may mistakenly be attributed to changes in control (see further discussion below). Indeed, as of yet, the type-2 SDT framework offers no standard way of examining or measuring differences in absolute monitoring accuracy between conditions or populations (e.g., age differences). This limits the usefulness of the framework for addressing situations in which subjective confidence and actual performance dissociate (e.g., Busey, Tunnicliff, Loftus, & Loftus, 2000; Kelley & Lindsay, 1993; Weingardt, Leonesio, & Loftus, 1994).

IV. Control Policy (Report Criterion)

Another point of contention, which is the one emphasized in Higham's chapter, concerns the way in which report criterion is conceived and measured in the two frameworks. Higham rightly points out that the QAP measure of report criterion, P_{rc}, is calculated with respect to the numeric values of the subjective probability ratings elicited from the participants, and that this makes the P_{rc} susceptible to scaling issues. However, it is precisely because the various type-2 SDT measures of report criterion are *not* tied to scale values of subjective confidence that these measures become susceptible to the misinterpretation of evidence/confidence distribution shifts, in which the overall level of true subjective confidence relative to the actual correctness of the answers changes between conditions (reflecting differences in over- or underconfidence), without any change in the true report criterion. Because P_{rc} is calculated relative to the elicited confidence values, its value is unaffected by confidence distribution shifts, whereas all type-2 SDT measures of report criterion are affected.

Fortunately, both QAP and type-2 SDT offer additional diagnostics that can be used to help gauge whether confidence scaling or distribution shifts have occurred. First with regard to the risk that a confidence scaling shift may be influencing the QAP measure of report-criterion, P_{rc} (Higham's Figure 5, this volume), the threat of a pure scaling shift is signaled whenever a change in P_{rc} (e.g., toward liberality) is accompanied by a parallel and opposite change in calibration bias (e.g., toward underconfidence). Any change in P_{rc} that is not accompanied by a change in calibration bias, or which is accompanied by a change in the same direction (e.g., more liberal P_{rc} accompanied by greater overconfidence), cannot be due to a confidence scaling shift alone. However, even in the first case it is possible that a true shift in report criterion, rather than a scaling shift, has occurred. As Higham points out, the situation depicted in his Figure 5 would also result from a downward shift in the subjective confidence distributions, together with a parallel downward shift in the report criterion. Such a scenario would be reflected correctly in the QAP measures, as a change in calibration bias (less overconfidence) and in P_{rc} (more liberal report criterion), but wrongly in the type-2 SDT measures, as a shift in confidence criteria (confidence scaling shift) with no change in report criterion.

Which underlying scenario is more plausible—a pure scaling shift or a joint distribution and report-criterion shift? This would seem to be a judgment call, and must take into account not only "parsimony," but also the specific memory variables involved, plausible theoretical arguments, and common sense. In his chapter, Higham discusses a pattern of results taken from Higham (2007), comparing a group of participants given feedback about the correctness of their answers (to SAT-type questions) to a no-feedback group, suggesting that a confidence scaling shift occurred as a result of the feedback, and that this was wrongly picked up by P_{rc} as a shift in report criterion. Acknowledging that those results are ambiguous with regard to whether they reflect a confidence scaling shift or a joint distribution and report-criterion shift, Higham nevertheless concludes that the former scenario is more plausible than the latter. But is it? The scaling interpretation assumes that feedback changed the way in which the participants assigned numbers to their true levels of subjective confidence, without actually changing their subjective confidence (or the report criterion). Why should that occur? In contrast, the distribution-shift interpretation holds that feedback about the correctness of their answers actually made the participants less (over-)confident about the correctness of their answers (a confidence distribution shift), and that in order to continue to volunteer, rather than omit, a reasonable number of answers, the report criterion was relaxed by a corresponding amount. This interpretation seems quite plausible to me, though because the QAP approach was not used in that study, some of the data needed to evaluate it (e.g., mean confidence and overconfidence scores) are not available.

Turning now to the diagnostics that the type-2 SDT approach offers to signal and potentially avoid the threat of a confidence-distribution shift, these too involve examining the overall pattern of indices, and not just the measure of report criterion per se. Thus, although one might naturally misinterpret a change in a type-2 SDT bias measure as reflecting a shift in report criterion when actually it reflects a confidence distribution shift, an examination of the specific pattern of differences in the hit and false-alarm rates, and in the confidence criteria, may signal that a distribution shift has occurred, and that the report criterion measure should not be taken at face value. Here too, however, the overall pattern of effects may be complex, and difficult to interpret unambiguously.[1]

In this regard, I disagree with Higham's claim that the SDT and QAP methods are equally threatened by the problem of distribution shifts. When a distribution shift has occurred, the QAP measures, taken at face value, will reflect the correct underlying scenario: a change in calibration bias (overconfidence), reflecting a change in the relationship between subjective confidence and the actual correctness of one's answers, with no change in report criterion. In contrast, the type-2 SDT measures, taken at face value, will reflect the wrong scenario: a change in the report criterion, suggesting that participants were more (or less) willing to risk providing wrong answers, and a change in the confidence criteria, suggesting that there was also a confidence scaling shift. As Higham correctly observes, it is once again the possible scenario of a confidence scaling shift—now tied to a lockstep shift in report criterion in the same direction—that threatens the QAP measures, which taken at face value, would wrongly indicate the occurrence of a distribution shift (i.e., an overall shift in subjective confidence that is not a mere scaling effect).

So, it seems that in weighing the advantages and disadvantages of QAP versus type-2 SDT, one really has no choice but to consider the relative risks of confidence scaling shifts versus confidence distribution shifts. Of course this is not an easy task, and it may be necessary to gather further empirical data regarding the conditions that are likely to produce such shifts. Here I will just note that psychometric issues concerning the measurement of subjective confidence have been studied and discussed extensively in the judgment and decision-making literature, including issues of reliability and validity that are essentially the same as with any other self-report measure (e.g., Wallsten & Budescu, 1983; Wallsten et al., 1993). In particular, there has been no suggestion that comparisons of mean confidence or mean overconfidence between groups or conditions be avoided because of the risk that these may merely reflect differences in usage of the confidence scale, a suggestion which would essentially preclude the use of confidence judgments (and all other subjective-report measures) in psychological research. Thus, Higham's suggestion that the use of P_{rc} in the QAP methodology should be restricted to situations in which the variables of interest are manipulated only after the confidence judgments have been elicited seems a bit odd. In fact, by the same token, one would have to recommend that the use of the type-2 SDT measures be similarly restricted, since that is also the only way in which a confidence-distribution shift can be completely precluded.

In this vein, I should note that although it is rather difficult to think of common memory manipulations that are likely to systematically affect the way in which subjective confidence is translated into numbers (i.e., a pure scaling shift), it is quite easy to think of those that are likely to increase subjective confidence in one's answers relative to the actual likelihood that they are correct (i.e., a single- or double confidence-distribution shift). This in fact is what most of the memory manipulations commonly used to elicit memory errors are designed to do, including the DRM paradigm (Roediger & McDermott, 1995), the misinformation paradigm (Loftus, Miller, & Burns, 1978), imagination inflation (Goff & Roediger, 1998), and so forth. In each of these paradigms, a primary indication that the manipulation has "succeeded" is that wrong answers are now held with high confidence (see, e.g., Weingardt et al., 1994). Unless there is a corresponding reduction in the confidence with which correct answers are held in these paradigms (and to my knowledge there is not), a confidence distribution shift is implied (for discussion of this issue in the context of the DRM paradigm, see Starns, et al., 2007; Wixted & Stretch, 2000).

V. Control Sensitivity

Finally, turning to the issue of “control sensitivity” (the extent to which the report control decisions are in fact based on subjective confidence), whereas this variable is explicitly included and measured within the QAP framework, so far it has essentially been ignored in the application of the type-2 SDT methodology to memory reporting. This omission is perhaps not arbitrary, as the signal detection framework is founded on the assumption that control (e.g., report) decisions are made by placing response criteria on distributions of subjective evidence. In contrast, a great deal of work in metacognition has emphasized the potential rift between monitoring and control—the possibility that one may have metacognitive knowledge or information that is not actually used in controlling one’s cognitive behavior, or that is perhaps overridden by other considerations (e.g., Ackerman & Goldsmith, 2008; Dunlosky & Connor, 1997; Schneider & Pressley, 1997). Along these lines, among university undergraduate participants, control sensitivity is generally very high, with within-participant gamma correlations between confidence and volunteering often averaging .95 or higher (Koriat & Goldsmith, 1996b)! Yet, when turning to special populations, a reduction in control sensitivity is sometimes observed, offering insights into the nature of metacognitive control deficits ensuing from old age (Pansky, Koriat, Goldsmith, & Pearlman-Avnion, 2009), mental illness (Danion, Gokalsing, Robert, Massin-Krauss, & Bacon, 2001; Koren, Seidman, Goldsmith, & Harvey, 2006), and psychoactive drugs (Massin-Krauss, Bacon, & Danion, 2002).

VI. Conclusions

To sum up, Koriat and Goldsmith’s QAP framework and Higham’s type-2 SDT framework both share the common goal of studying and assessing the contributions of post-retrieval monitoring and control processes to free-report memory performance, and both have adopted very similar approaches to achieving that goal. Yet, there are disagreements about specific measures, and also about the ways in which some of the underlying theoretical components are conceptualized. Although it is true that the SDT framework has a long history of significant contribution to memory research, offering a rich arsenal of “tried and true” tools that can be drawn upon, these tools were established in the application of type-1 SDT to old/new recognition memory, in which “don’t know” answers are not allowed. In contrast, the application of type-2 SDT in memory research has been relatively rare and fraught with conceptual and methodological confusions (Healy & Jones, 1973). In his work, Phil Higham has done an outstanding job of clarifying the interpretation of type-2 SDT measures in memory, and adapting them to study the strategic regulation of memory performance in free-report situations. Nevertheless, along with the advantages of adopting a relatively familiar and well established all-purpose framework, there appear to be some disadvantages in being tied to concepts and assumptions that take on different meanings than in the more dominant, type-1 framework, and which may not capture the full complexity of the topic under investigation, in particular, the nature of subjective confidence (e.g., Busey et al., 2000; van Zandt, 2000).

The metacognitive framework and methodology developed by Koriat and Goldsmith (1996b) for studying the strategic regulation of memory reporting is still very much a work in progress. Although originally developed to examine the mechanisms and performance consequences of report option—the option to volunteer or withhold individual items of information—the framework has since been extended to encompass an additional type of control—control over the precision or coarseness of the information that is reported, control over “grain size” (Goldsmith, Koriat, & Weinberg-Eliezer, 2002; Goldsmith, Koriat, & Pansky, 2005). The mechanisms underlying the control over grain size were found to be similar to those involved in the control of report option, and in fact point to the possibility of a common integrated model that can account for the joint use of both types of control in memory reporting (see Ackerman & Goldsmith, 2008; Goldsmith & Koriat, 2008). In addition, an attempt is now being made to extend the framework and the QAP methodology to examine monitoring and control processes involved in the retrieval of information from memory, in addition to those involved in the evaluation and reporting of that information (Goldsmith, Jacoby, Halamish, & Wallheim, 2009; Koriat, Goldsmith, & Halamish, 2008). It is not clear whether the type-2 SDT framework is flexible enough to support such extensions as well.

It is now the case that researchers interested in studying the strategic regulation of memory performance have alternative guiding frameworks and tools to choose from—each with its own specific advantages and disadvantages. Hopefully, the discussion embodied in this and Phil Higham's chapter in this volume can offer some guidance to researchers who may choose to join the endeavor, as to which set of concepts and tools is most suited to their personal inclinations and research goals.

References

Ackerman, R. & Goldsmith, M. (2008). Control over grain size in memory reporting—with and without satisficing knowledge. *Journal of Experimental Psychology: Human Learning and Memory*, *34*, 1224-1245.

Benjamin, A.S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R.A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 73-94). New York: Psychology Press.

Brewer, W. F., & Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: A metamemory approach. *Memory, 14,* 540–552.

Danion, J. M., Gokalsing, E., Robert, P., Massin-Krauss, M., & Bacon, E. (2001). Defective relationship between subjective experience and behavior in schizophrenia. *American Journal of Psychiatry*, *158*, 2064-2066.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, *7*, 26–48.

Dunlosky, J., & Connor, L. T. (1997). Age differences in the allocation of study time account for age differences in memory performance. *Memory and Cognition*, *25*, 691-700.

Goff, L.M., & Roediger, H.L., III. (1998). Imagination inflation for action events: Repeated imaginings lead to illusory recollections. *Memory & Cognition*, *26*, 20-33.

Goldsmith, M., Jacoby, L. L., Halamish, V., & Wahlheim, C. N. (2009). *Metacognitively Guided Retrieval and Report (META-RAR): Quality control processes in recall.* Paper presented at the 50th Annual Meeting of the Psychonomic Society, Boston, MA.

Goldsmith, M. & Koriat, A. (2008). The strategic regulation of memory accuracy and informativeness. In A. Benjamin and B. Ross (Eds.), *Psychology of Learning and*

Motivation, Vol. 48: Memory use as skilled cognition (pp. 1-60). San Diego, CA: Elsevier.

Goldsmith, M., Koriat, A., & Pansky, A. (2005). Strategic regulation of grain size in memory reporting over time. *Journal of Memory and Language (Special Issue on Metamemory)*, *52*, 505-525.

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). The strategic regulation of grain size in memory reporting. *Journal of Experimental Psychology: General*, *131*, 73-95.

Healy, A. F., & Jones, C. (1973). Criterion shifts in recall. *PsychologicalBulletin*, 79, 335-340.

Higham, P. A. (2007). No special K! A signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General*, *136*, 1-22.

Higham, P. A. (2010). Accuracy discrimination and type-2 signal detection theory: Clarifications, extensions, and an analysis of bias. In … **THIS VOLUME**.

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.

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language, 48,* 704-721.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *34*, 945-959

Koren, D., Seidman, L. J., Goldsmith, M. & Harvey, P.D. (2006). Real-world cognitive—and metacognitive—dysfunction in schizophrenia: a new approach for measuring (and remediating) more "right stuff". *Schizophrenia Bulletin*, *32*, 310-326.

Koriat, A. & Goldsmith, M. (1996a). Memory metaphors and the real-life/laboratory controversy: Correspondence versus storehouse conceptions of memory. *Behavioral and Brain Sciences*, *19*, 167-188.

Koriat, A. & Goldsmith, M. (1996b). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490-517.

Koriat, A. Goldsmith, M., & Halamish, V. (2008). Control processes in voluntary remembering. In H. L. Roediger, III (Ed.), *Cognitive psychology of memory. Vol. 2 of Learning and memory: A comprehensive reference, 4 vols*. (J. Byrne, Editor) (pp. 307-324). Oxford, UK: Elsevier.

Liberman, V. and Tversky, A. (1993). On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin*, 114, 162-173.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). Cambridge, England: Cambridge University Press.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 19-31.

Massin-Krauss M., Bacon E., & Danion J-M. (2002). Effects of the benzodiazepine lorazepam on monitoring and control processes in semantic memory. *Conscioussness and Cognition, 11*, 123-137.

Masson, M.E.J, & Rotello, C.M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 509-527.

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

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. Comments on Schraw (1995). *Applied Cognitive Psychology*, *10*, 257-260.

Pansky, A., Koriat, A., Goldsmith, M., & Pearlman-Avnion, S. (2009). Memory accuracy and distortion in old age: Cognitive, metacognitive, and neurocognitive determinants. *European Journal of Cognitive Psychology, 21*, 303-329.

Roediger, H. L. & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory & Cognition, 21*, 803-14

Rotello, C.M., Masson, M.E.J., & Verde, M.F. (2008). Type 1 error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics, 70*, 389-401.

Schraw, G. (1995). Measures of feeling-of-knowing accuracy: a new look at an old problem. *Applied Cognitive Psychology,* 9, 321-332.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*, 33-45.

Schneider, W., & Pressley, M. (1997). *Memory development between two and twenty*. Mahwah, NJ: Lawrence Erlbaum Associates.

Starns, J. J., Lane, S. M., Alonzo, J. D., & Roussel, C. C. (2007). Metamnemonic control over the discriminability of memory evidence: A signal-detection analysis of warning effects in the associative list paradigm. *Journal of Memory and Language, 56*, 592–607.

Wallsten, T. S., & Budescu, D. V. (1983). Encoding subject probabilities: a psychological and psychometric review. *Management Science,* 29, 151-173.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, *39*, 176-190.

Weingardt, K. R., Leonesio, R. J., & Loftus, E. F., (1994). Viewing eyewitness research from a metacognitive perspective. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 157-184). Cambridge: MIT Press.

Whittlesea, B. W. A. (2002). Two routes to remembering (and another to remembering not). *Journal of Experimental Psychology: General*, *131*, 325-348.

Whittlesea, B. W. A., & Williams, L. D. (2001a). The discrepancy-attribution hypothesis: I. The heuristic basis of feelings and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 3-13.

Whittlesea, B. W. A., & Williams, L. D. (2001b). The discrepancy-attribution hypothesis: II. Expectation, uncertainty, surprise, and feelings of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 14-33.

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review*, *107*, 368-376.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, *11*, 616–641.

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, *110*, 611-617.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, *30*, 132-156.

Footnotes

[1] I should note that I do not accept Higham's distinction between measures that are potentially "misleading" and those that are merely "ambiguous." By this distinction, $P_{rc,}$ which purports to be an unambiguous measure of report criterion is "misleading," whereas the type-2 SDT measures such as β, C and B''_D, which do not purport to be unambiguous measures of report criterion, are merely ambiguous. I do not think it should be necessary to scour the literature in order to gauge how often SDT bias measures have been reported and taken at face value as measures of response criterion, regardless of the fact that they are indeed, ambiguous when used in this way. Thus, the common use of SDT bias measures as an index of response/report criterion is "potentially misleading." Once one is aware of this problem, one can and probably should take Higham's advice and not use the type-2 SDT bias measures to index report criterion (or confidence criteria, for that matter). But this solution essentially leaves the type-2 SDT approach without a measure of report criterion, because inferring criterion changes from differences in the patterns of hit- and false-alarm rates, or by mathematical modeling (e.g., Starns, et al., 2007) is unwieldy, and does not serve well as a dependent variable, for example, in examining the interaction between age and accuracy incentives on the report criterion setting (cf. Kelly & Sahakyan, 2003).