ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Journal of Mathematical Psychology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Journal of Mathematical Psychology

www.elsevier.com/locate/jmp

Problematic effects of aggregation in *z*ROC analysis and a hierarchical modeling solution

Richard D. Morey*, Michael S. Pratte, Jeffrey N. Rouder

University of Missouri-Columbia, United States

Received 14 September 2007; received in revised form 18 January 2008

Abstract

The confidence-rating recognition memory paradigm has been used extensively for testing theories of memory. Current methods for analyzing confidence-rating data require that data be aggregated over participants, items, or both. We show that this aggregation threatens claims of curvature in *z*ROCs and distorts estimates of signal-distribution variance—the very findings that are used to make fine-grained distinctions between competing models. We develop a hierarchical signal detection model that accounts for variability from the sampling of participants and items in addition to variability from the underlying psychological processes. The model provides for accurate signal detection parameter estimates. With accurate estimates, the validity of benchmark findings in recognition memory may be assessed.

The study of human memory has had a long and rich history in psychology. One influential paradigm for testing theories of memory is the confidence-rating recognition memory paradigm. Participants judge whether a presented word was previously studied or not, and then rate their confidence in this judgment. Researchers have used this paradigm to test various positions about representations in memory (e.g., Ratcliff, McKoon, & Tindall, 1994), whether there are one or multiple memory processes (e.g., Yonelinas, 1999; Yonelinas & Parks, 2007), and how recognition judgments are made (e.g., Glanzer & Adams, 1990). The most widely used model for drawing conclusions about mnemonic processes from the confidence-rating recognition-memory paradigm is the theory of signal detection (SDT; Green & Swets, 1966). In conventional designs and analyses, researchers aggregate results over items, participants, or both, to obtain estimates of hit and false alarm rates. We show how this aggregation leads to asymptotic distortion in estimates of SDT sensitivity (d') and variability (σ) parameters. This demonstration suggests that much of the theoretical insight gained from the confidence-rating paradigm is tenuous. To provide for accurate SDT analysis of the paradigm, we advocate a hierarchical SDT model in which participant and item variability, as well as variability in processing, can be accounted for simultaneously. The development here is similar to our recent development for the process dissociation paradigm (Rouder, Lu, Morey, Sun, & Speckman, in press). We showed that while aggregation across disparate items leads to asymptotic distortion in the process-dissociation parameters, a Bayesian hierarchical model provides for accurate analysis.

We start with the supposition that there is substantial variation across people and items in typical recognition memory tasks. It is highly plausible that participants vary in their mnemonic abilities and in their biases. Likewise, it is highly plausible that items vary in their memorability as well as the biases they induce in processing. Because both participants and items presumably vary, each participant-by-item combination has its own characteristic memorability and bias. It is therefore desirable to estimate parameters for each participant-by-item combination. Participant-by-item combinations, however, are not replicated in standard designs. As a consequence, researchers aggregate data over items or participants to estimate item-averaged or participant-averaged parameters, respectively. We will show how this aggregation leads to distortion which, in turn, threatens the validity of benchmark findings that have guided model development.

Our critique of aggregation and the need for the hierarchical model are both vested in the presumption of item variability.

* Corresponding address: University of Missouri-Columbia, Psychological Sciences, 210 McAlester Hall, 65211 Columbia, MO, United States.
E-mail address: moreyr@missouri.edu (R.D. Morey).

0022-2496/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2008.02.001

Please cite this article in press as: Morey, R. D., et al. Problematic effects of aggregation in *z*ROC analysis and a hierarchical modeling solution. Journal of Mathematical Psychology (2008), doi:10.1016/j.jmp.2008.02.001

We believe there is good evidence that items vary in important ways in memory experiments. One source of evidence comes from the well-known effect of word frequency on memorability. Participants remember low-frequency words better than high-frequency words, and the effect is often surprisingly large (e.g., Peters, 1936; Scarborough, Cortese, & Scarborough, 1977; Schwartz & Rouse, 1961; Shepard, 1967). Whereas most experimenters simply classify words as "high" or "low" frequency, there is often much unmodeled variation within each category. A second source of evidence comes from our recent work in process dissociation (Rouder et al., in press). We fit a model, not unlike the one presented here, to stem completion data to measure recollective and automatic processes. Not only did we find substantial item variability within an experiment, item scores were correlated at $r = .7$ across experiments with different encoding instructions and different participants. Items in our process dissociation experiments had divergent and stable recollection and automatic effects. There is no reason to doubt that the same holds for recognition memory.

1. The theory of signal detection in recognition memory

Although SDT serves as a model of memory in its own right (Kintsch, 1967), SDT analysis also often forms the link between data and theory for more complex analyses. In SDT, when an item is studied, its mnemonic strength or familiarity is increased. These strengths are distributed as normals, with the means and variances depending on whether an item was studied or not. Without any loss of generality, the mean and variance of the nonstudied-item distribution is set to 0.0 and 1.0, respectively, and these settings scale the memory-strength dimension. The mean and variance of the studied-item distribution are free parameters denoted d' and σ^2, respectively. An example of these distributions is provided in the top panel of Fig. 1. To produce a response, the participant sets criteria on memory strength. In the two-choice case in which participants decide whether an item was studied or not, there is a single criterion. A "studied-item" response is produced if strength is above this criterion; a "nonstudied-item" response is produced otherwise. In the confidence-rating task, multiple criteria are placed on memory strength. An example of the signal detection model for a confidence-rating task is displayed in the top panel of Fig. 1. The participant's response ("Sure Nonstudied", "Believe Nonstudied", "Believe Studied", or "Sure Studied") is determined by where the mnemonic strength lies relative to the criteria.

A conventional representation of data from the confidence rating task is the receiver-operating characteristic (ROC) and its z-score transform (zROC). To derive SDT model predictions for ROC and zROC plots, let H be the probability that a studied item evokes a familiarity greater than a criterion c, and let F denote the probability that an unstudied item evokes a familiarity greater than c. We define the criterion c as relative to 0, the mean of the nonstudied-item distribution. Consequently, H and F are

$$H = \Phi\left(\frac{d' - c}{\sigma}\right),$$

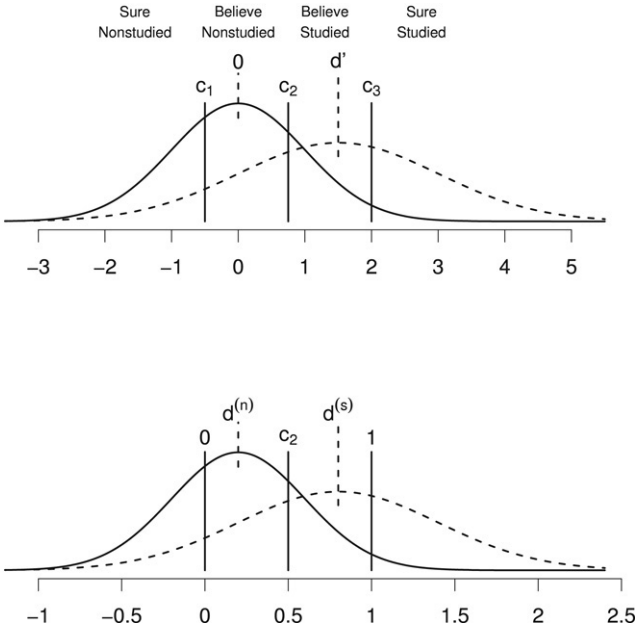

Fig. 1. The signal detection model. Top: The conventional signal detection model for an experiment with 4 response options. Bottom: The fixed-criteria parametrization in which the outer criteria are set to 0.0 and 1.0.

$$F = \Phi(-c),$$

where Φ is the cumulative distribution function of the standard normal distribution. Let h and f denote the respective z-transformed probabilities, i.e., $H = \Phi(h)$ and $F = \Phi(f)$. Then,

$$h = \frac{d' - c}{\sigma}, \tag{1}$$

$$f = -c. \tag{2}$$

The zROC plot is a plot of h as a function of f when c is varied and d' and σ are held constant. It is clear from (1) and (2) that the SDT model predicts linear zROCs with a slope of $1/\sigma$ and an intercept of d'/σ. The solid line in Fig. 2 is the zROC from a signal detection model with $d' = 2$ and $\sigma = 1$; the line has an intercept of 2, and a slope of 1.

2. Benchmark findings in the literature

The form of zROCs is used to constrain theories of recognition memory (cf. Malmberg, 2002). Yonelinas and Parks (2007) note that zROCs display several regularities in recognition memory experiments. We follow them in highlighting three benchmark findings:

1. Empirical zROC curves are approximately linear, but sometimes show curvature (Ratcliff et al., 1994; Yonelinas, 1999). This curvature provides evidence for Yonelinas' (1994) dual process theory of recognition memory, and against other models such as signal detection theory.
2. Almost all empirical zROCs have a slope of less than 1, indicating that the variance of the studied-item distribution is greater than that of the nonstudied-item distribution. This result, popularized by Ratcliff, Sheu, and Grondlund (1992),

Please cite this article in press as: Morey, R. D., et al. Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. Journal of Mathematical Psychology (2008), doi:10.1016/j.jmp.2008.02.001

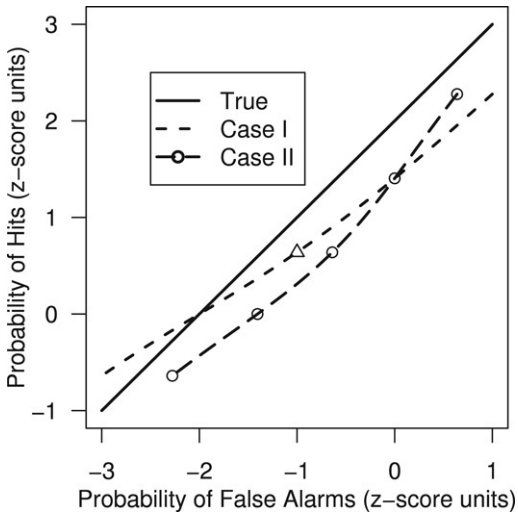

Fig. 2. The effects of aggregation on zROC curves. The straight line labeled "True" corresponds to a signal detection model with $d' = 2$ and $\sigma = 1$. The curve labeled "Case I" results from aggregating over items with different d' values. The curve labeled "Case II" results from aggregating over items with different criteria.

has been replicated many times. Glanzer, Kim, Hilford, and Adams (1999), for example, reported greater variance for studied items in 40 out of 41 experiments they reviewed. This result is concordant with theories that predict $\sigma > 1$ (e.g., MINERVA2, Hintzman, 1986) and is discordant with those that predict $\sigma = 1$ (e.g., TODAM, Murdock, 1982).

3. The slopes and intercepts of empirical zROC curves are often negatively correlated (Glanzer et al., 1999; Heathcote, 2003; cf. Ratcliff et al., 1994). Specific relations among the slope and intercept serve as evidence for and against various formal models (Heathcote, 2003; Ratcliff et al., 1992).

In the next section, we show how these benchmark findings using zROCs could simply be the result of aggregation. If the benchmark findings are artifactual, then theories have been built, supported, and rejected on the basis of aggregation artifacts.

3. The effects of aggregation

In this section, we document how aggregation distorts zROC plots and the resulting estimates of signal detection parameters. Before doing so, we draw a sharp distinction between variation in psychological process from variation at other levels, such as that from sampling items or participants. The distinction is best illustrated by a thought experiment. Imagine it were possible to produce independent and identically distributed replicates of item-by-participant combinations in a recognition memory paradigm. Clearly this thought experiment is impossible in reality; replicates would be contaminated by learning and priming effects, which would violate the independent-and-identically-distributed assumption. Nonetheless, if it were possible, we would observe variability across the replicates. This variability would not reflect variability in sampling items and participants, but would only reflect variability in the mnemonic processes.

Many researchers assume, at least implicitly, that the results of recognition memory experiments reflect the underlying variability solely in mnemonic processes. Though these researchers do not explicitly deny item variability, there is the tacit belief that this unmodeled variability does not affect ROC and zROC curves if enough data are collected and items are counterbalanced. In this view, aggregation is not problematic because in the limit, a true assessment of the mnemonic processes may be obtained. Indeed, when researchers use the benchmark findings of $\sigma > 1$ to revise mnemonic theories (e.g. Murdock, 1993), they are implicitly assuming this finding reflects mnemonic variation and not item variation. Likewise, if curvature of zROCs curves is interpreted as evidence for the presence of two processes, then, again, the interpretation is conditioned on the implicit assumption that zROCs provide an uncontaminated view of the mnemonic process. We call this implicit assumption the *optimist* view of aggregated recognition-memory data and show how it is flawed.

3.1. Threats to the optimist view

If sources of variability other than mnemonic process exist in recognition memory experiments, then the optimist view is incorrect. Malmberg and Xu (2006) describe the effects of participant variability on the shape of ROC curves. Aggregating data across variable participants affects estimates of the slopes of zROC curves, thus affecting inferences made about the mnemonic process by the optimist researcher. Malmberg and Xu's critique is also applicable to aggregation across items. We highlight the case of item variability because researchers most often aggregate over items in order to draw conclusions about psychological process. We consider two ways in which items can vary: In *Case* I items differ in their memorability; in *Case* II items induce different biases. We take these cases in turn.

Case I: *Items vary in memorability*. Consider a study list in which, unknown to the researcher, half of the items are easy and half are hard. To make the situation concrete, let d' for the easy and hard items be 3 and 1, respectively, and let $\sigma = 1$ for all items. Assume the researcher aggregates the results across all items. If aggregation has no ill effects, then estimates from aggregated data should yield the mean memorability of $d' = 2$ and a standard deviation of $\sigma = 1$ in the large-sample limit. Moreover, the zROC curves of aggregated data should be straight lines with a slope of 1 and an intercept of 2.

Unfortunately, aggregated data yield neither average parameter values nor the appropriate zROC curve, even in the large-sample limit. Consider, for example, the probabilities of hits and false alarms for a criterion of $c = 1$. The probability of a hit is .5 for difficult items and .977 for easy ones. The probability of a hit from aggregated data is therefore $(.5 + .977)/2 = .739$. The probability of a false alarm is .16 for both types of items, hence the probability of false alarm from aggregated data is .16. The z-transforms of these probabilities from aggregated data are .639 and -1.0 for hit and false alarms, respectively. The point defined by these values is denoted by the symbol "△" in Fig. 2. Unfortunately, this point does not fall on the line labeled "True". The line labeled "Case I" in

Please cite this article in press as: Morey, R. D., et al. Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. Journal of Mathematical Psychology (2008), doi:10.1016/j.jmp.2008.02.001

Fig. 2 is the zROC curve from aggregated data. The zROC from aggregation is not perfectly straight; it has a slight degree of curvature. The best-fitting line has a slope and an intercept of .72 and 1.46, respectively. These values correspond to signal detection parameters of $d' = 2.03$ and $\sigma = 1.39$. The resulting value of d' from aggregated data is very near the true average of $d' = 2$, but the resulting value of σ is far too high. Because we used true probabilities to construct zROCs, this upward bias is asymptotic; that is, it exists even in arbitrarily large data sets.

Case II: *Items vary in bias*. We consider the case that items vary in their baseline strength. Suppose an item on a memory test was recently encountered by a participant prior to the experiment. This recent encounter adds strength; for example, a nonstudied recently-encountered item may have strength distributed as a normal with a mean of .5 rather than 0. Likewise, if this recently-encountered item is studied, the strengths are distributed with a mean of $d' + .5$ rather than d'. In the conventional signal detection model, where the mean of the nonstudied distribution is fixed, these baseline differences in mnemonic strength are manifest as criteria shifts; that is, a recently-encountered item has the same sensitivity and variance as other items but a criterion that is lowered (in the above example, the criteria would be lowered by .5). Singer, Gagnon, and Richard (2002) found evidence for trial-by-trial criteria shifts in memory for elements of stories. We find it plausible that item-based effects exist in simpler recognition memory experiments as well.

To demonstrate how variability in criteria distorts zROCs, we once again consider two types of items. The recently-encountered items and not-recently-encountered items have $d' = 2$ and $\sigma = 1$, but the recently-encountered items induce a criterion 2 units lower than the not-recently-encountered ones. The data are aggregated across items, hence the probability of hits and false alarms are averages across different criteria. The resulting zROC curve from the aggregated data is shown by the line labeled "Case II" in Fig. 2. If aggregation across criteria has no ill effects, then the curve will lie on the straight line denoted "True". As seen in Fig. 2, the curve deviates substantially and has a fair degree of convexity. The resulting values of d' and σ obtained from this curve by a linear fit are 1.43 and 1.01, respectively. The value of σ is sufficiently accurate, but the value of d' is substantially lower than the true value of 2.0.

The above demonstrations are idealized in that item variability is assumed to occur in either criteria or memorability, but not in both. It is plausible, however, that items vary in both criteria and memorability. In this case, under the optimist view, a combination of distortions will occur: 1. estimates of σ will be too high; 2. the zROC curves will be concave; and 3. estimates of d' will be too low. The first two distortions are especially problematic from a theoretical perspective as they correspond to the first two benchmark findings. The implication is that these benchmark findings may not reflect the underlying mnemonic process, but instead, may reflect the influences of extraneous sources of variation, such as those from participants and items.

The above demonstration shows that ROC dynamics may critically reflect multiple sources of variation. This fact has been acknowledged by a minority of authors including Malmberg and Xu (2006), Ratcliff et al. (1994), and Wixted (2007). For instance, Wixted (2007) considers familiarity distributions to reflect both mnemonic and item variation. Under this view, the results of the above demonstrations are not surprising—whereas item variability contributed to the studied-item distribution alone, it is expected that the resulting σ estimate is greater that 1. We refer to this view, that aggregated data can only inform about the convolution of sources of variation, as the *realist view* of aggregated recognition memory data.

Although this view is realistic, it is also disheartening, as its basic tenet is that it is not possible to separate out different sources of variation. Consequently, the benchmark findings are ambiguous. For example, because it is unclear whether the finding of $\sigma > 1$ reflects item variation or process variation, this benchmark cannot be used to inform mnemonic theories. Likewise, if curvature in zROC may be due to either item or process variation, the curvature benchmark cannot be used to differentiate between process models.

It is critical to emphasize three points regarding what may be inferred from aggregated data: Under the optimist view, 1. The distortions are asymptotic and cannot be overcome by running large experiments or by replication. 2. Because SDT is a nonlinear model, parameter distortions occur in all estimation methods that utilize aggregated rates, including maximum likelihood (Dorfman & Alf, 1969; Heathcote, Raymond, & Dunn, 2006). 3. Under the realist view, mnemonic process may not be isolated from other sources of variance. It is these pragmatic difficulties that motivate the need for the following hierarchical process model.

4. A hierarchical confidence-rating signal detection model

In this section, we provide a hierarchical signal detection model for accurate estimation of mnemonic processes as distinct from other sources of variation. The key feature of the model is that it simultaneously accounts for variability on three levels: variability due to participants, variability due to items, and variability in mnemonic processing. The model is similar to the one proposed by Rouder and Lu (2005) and Rouder, Lu, Sun, Speckman, Morey, and Naveh-Benjamin (2007). These previous models were applicable to dichotomous studied/nonstudied judgments, but could not be applied to confidence-rating data. Moreover, these previous models were not designed to clarify whether the existing zROC benchmark findings reflect mnemonic variation or are artifacts of aggregation. The extension to the confidence-rating case with the inclusion of unequal variances is nontrivial and required substantial development. In this section, we specify the model. In the next section we provide algorithms for analysis. Following that, we show through simulation that the model provides accurate estimation even when aggregation-based estimation fails to accurately reflect mnemonic variation.

The conventional signal detection model is shown in the top panel of Fig. 1. We first discuss the case of a single participant observing a single item. Let y denote the response with $y = 1, \ldots, K$, where K denotes the number of response options.

Please cite this article in press as: Morey, R. D., et al. Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. Journal of Mathematical Psychology (2008), doi:10.1016/j.jmp.2008.02.001

We assume $K > 2$ for the confidence-rating task. In addition to sensitivity parameter d' and standard deviation σ, there are $K-1$ criteria denoted $c_1, \ldots, c_{K-1}$. The probability that $y = k$ is:

$$\Pr(y = k \mid \text{Not Studied}) = \Phi(-c_k) - \Phi(-c_{k-1}),$$

$$\Pr(y = k \mid \text{Studied}) = \Phi\left(\frac{d' - c_k}{\sigma}\right) - \Phi\left(\frac{d' - c_{k-1}}{\sigma}\right),$$

where $c_0 = -\infty$ and $c_K = \infty$.

The hierarchical model is most easily developed with a reparametrization of the signal detection model. In the conventional parametrization, the mean and variance of the nonstudied-item distribution is set to fixed values that scale the memory-strength space. In the reparametrization, the outer two criteria, c_1 and c_{K-1}, are set to fixed values that scale the memory-strength space and the mean and variance of the nonstudied-item distribution are free to vary. Consequently, we refer to this reparametrization as the *fixed-criteria* parametrization of SDT. Without any loss of generality, the fixed criteria may be set to $c_1 = 0$ and $c_{K-1} = 1$. The means and variances of the nonstudied-item and studied-item distributions are denoted by $d^{(n)}$, $d^{(s)}$, σ_n^2 and σ_s^2, respectively. Sensitivity is therefore given as $d' = \frac{d^{(s)} - d^{(n)}}{\sqrt{\sigma_n^2}}$. The bottom panel of Fig. 1 shows the fixed-criteria reparametrization. This reparametrization is discussed further in the General Discussion.

We extend the fixed-criteria parametrization to account for both participant and item effects. Let y_{ij} be the response of the ith participant to the jth item, $i = 1, \ldots, I$, $j = 1, \ldots, J$. The probabilities of making the kth response to nonstudied and studied items are:

$$\Pr(y_{ij} = k \mid \text{Not Studied}) = \Phi\left(\frac{d_{ij}^{(n)} - c_{ik}}{\sqrt{\sigma_n^2}}\right) - \Phi\left(\frac{d_{ij}^{(n)} - c_{i(k-1)}}{\sqrt{\sigma_n^2}}\right), \quad (3)$$

$$\Pr(y_{ij} = k \mid \text{Studied}) = \Phi\left(\frac{d_{ij}^{(s)} - c_{ik}}{\sqrt{\sigma_s^2}}\right) - \Phi\left(\frac{d_{ij}^{(s)} - c_{i(k-1)}}{\sqrt{\sigma_s^2}}\right), \quad (4)$$

for $k = 1, \ldots, K$. Parameters $d_{ij}^{(n)}$ and $d_{ij}^{(s)}$ are means of the nonstudied-item and studied-item distribution for the ijth participant-by-item combination, respectively. Parameters $c_{i\ell}$, $\ell = 1, \ldots, K-1$ are criteria for the ith participant. Throughout this article, parameters indexed by n are related to the nonstudied-item distribution, and those indexed by s are related to the studied-item distribution.

Clearly, it is not possible to estimate all of the distribution means ($d_{ij}^{(n)}$ and $d_{ij}^{(s)}$) without further constraint. This constraint is provided by treating participant and item effects as additive:

$$d_{ij}^{(n)} = \mu^{(n)} + \alpha_i^{(n)} + \beta_j^{(n)}, \quad (5)$$

$$d_{ij}^{(s)} = \mu^{(s)} + \alpha_i^{(s)} + \beta_j^{(s)}, \quad (6)$$

where $\mu^{(n)}$ and $\mu^{(s)}$ are grand means, $\alpha_i^{(n)}$ and $\alpha_i^{(s)}$ are zero-centered participant effects, and $\beta_j^{(n)}$ and $\beta_j^{(s)}$ are zero-centered item effects. We do not restrict $\mu^{(n)} < \mu^{(s)}$. Although the restriction is plausible, it would greatly complicate analysis (see Chen & Shao, 1998) and would be for little gain.

Participant and item effects are treated as random such that results may be generalized to novel people and items:

$$\alpha_i^{(n)} \overset{iid}{\sim} \text{Normal}(0, \sigma^2_{\alpha^{(n)}}), \quad (7)$$

$$\alpha_i^{(s)} \overset{iid}{\sim} \text{Normal}(0, \sigma^2_{\alpha^{(s)}}), \quad (8)$$

$$\beta_j^{(n)} \overset{iid}{\sim} \text{Normal}(0, \sigma^2_{\beta^{(n)}}), \quad (9)$$

$$\beta_j^{(s)} \overset{iid}{\sim} \text{Normal}(0, \sigma^2_{\beta^{(s)}}). \quad (10)$$

Overall sensitivity is given as

$$d' = \frac{\mu^{(s)} - \mu^{(n)}}{\sqrt{\sigma_n^2}}. \quad (11)$$

This expression for overall sensitivity is consistent with unequal-variance signal detection, where the unit for measuring discriminability is the standard deviation of the nonstudied distribution.

The assumption of additivity in Eqs. (5) and (6) is common and has been successful in previous models of signal detection (Rouder & Lu, 2005; Rouder, Lu et al., 2007).

Because criteria are not estimated separately for each item, it may appear as though the plausible variation in item baseline strength mentioned previously is not modeled. The model, however, readily accommodates these effects. For example, the top panel of Fig. 3 shows a baseline effect as a coordinated increase in the means of both $d^{(n)}$ and $d^{(s)}$. These baseline shifts are equivalent to effects on criteria: the effect shown in the top panel is equivalent to shifting criteria to increase the probability of a "studied" response regardless of whether the item was studied or not. In general, baseline effects (or equivalently, response biases) are reflected by positive correlations between $\beta_j^{(n)}$ and $\beta_j^{(s)}$ for items and positive correlations between $\alpha_i^{(n)}$ and $\alpha_i^{(s)}$ for participants.

The dependence of distribution means on both items and participants provides for a large degree of flexibility. This flexibility may be demonstrated by consideration of the mirror effect (Glanzer & Adams, 1990). The mirror effect is the phenomenon that conditions resulting in increased hit rates also result in decreased false-alarm rates. A mirror effect may be modeled as a negative correlation between $d^{(s)}$ and $d^{(n)}$ (see the bottom panel of Fig. 3). One advantage of the current parametrization is that mirror effects across conditions, participants, and items may be independently assessed.

The model outline above accounts for item and participant variability. It is plausible that other sources of variance exist in recognition memory experiments, such as study-test lag. If important covariates are left out, the nonprocess variance

Please cite this article in press as: Morey, R. D., et al. Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. Journal of Mathematical Psychology (2008), doi:10.1016/j.jmp.2008.02.001

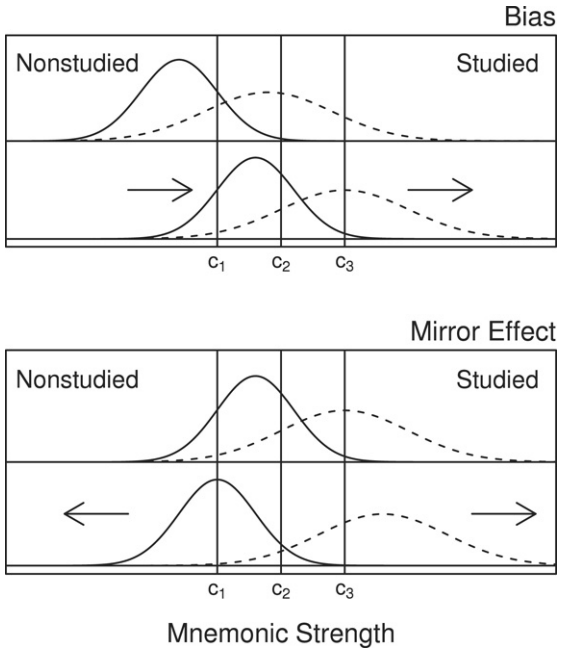

Fig. 3. Baseline (response bias) and mirror effects are accounted for by the inclusion of participant and item effects on distribution means. Top: Increasing the means of both distributions reflects a liberal response bias. Bottom: Mirror effect results from a decrease in the mean of the nonstudied-item distribution and an increase in the mean of the studied-item distribution.

will continue to contaminate estimates of process variance. Although we did not include other covariates in Eqs. (5) and (6), it is straightforward to include them in a model analysis. Parameters may be added to Eqs. (5) and (6), along with appropriate priors analogous to Eqs. (7)–(10). The design matrix is then expanded to accommodate the new parameters and known covariates. In this way, researchers may assess the relative importance of covariates of interest.

5. Model analysis

The hierarchical confidence-rating model of (3) through (10) is analyzed in the Bayesian framework with Gibbs sampling (Gelfand & Smith, 1990). Because the development is Bayesian, priors are needed for parameters $\mu^{(n)}$, $\mu^{(s)}$, σ_s^2, σ_n^2, $\sigma_{\alpha^{(n)}}^2$, $\sigma_{\alpha^{(s)}}^2$, $\sigma_{\beta^{(n)}}^2$, and $\sigma_{\beta^{(s)}}^2$. In this section, we provide and justify priors, present the full conditional posterior distributions of each parameter, and provide sampling algorithms for estimating the marginal posterior distribution of the parameters.

5.1. Priors

We place normal priors on the grand means of the nonstudied- and studied-item distributions:

$$\mu^{(n)}, \mu^{(s)} \overset{iid}{\sim} \text{Normal}(0, \sigma_\mu^2).$$

We place inverse gamma priors on variance parameters:

$$\sigma_n^2, \sigma_s^2 \overset{iid}{\sim} \text{IG}(a, b),$$
$$\sigma_{\alpha^{(n)}}^2, \sigma_{\beta^{(n)}}^2, \sigma_{\alpha^{(s)}}^2, \sigma_{\beta^{(s)}}^2 \overset{iid}{\sim} \text{IG}(e, f).$$

Values of σ_μ^2, a, b, e and f must be chosen before analysis. As long as the value of σ_μ^2 is sufficiently large, the prior is approximately noninformative. We use $\sigma_\mu^2 = 10{,}000$ though other large values will yield the same results. In reasonable sample sizes, we recommend values of $a = e = 2$ and $b = f = 1$. These values yield priors with a mean of 1.0 and an infinitely large variance. With these choices for (a, b, e, f), the priors are vaguely informative, but this information has a minimal effect on the posterior means for variance parameters.

5.2. Conditional posterior distributions

The following notation is used in deriving the conditional posterior distributions. Parameters in bold type indicate a vector of parameters or data; for example, $\mathbf{y}$ denotes the vector of all data. Let s_{ij} indicate whether the jth item for the ith participant was studied ($s_{ij} = 1$) or not ($s_{ij} = 0$). Derivation of conditional posterior distributions is greatly aided by introducing a set of latent variables. Following Albert and Chib (1993), let w_{ij} be related to y_{ij} as follows:

$$(y_{ij} = k) \Longleftrightarrow (c_{i(k-1)} \le w_{ij} < c_{ik}).$$

Random variables w_{ij} are distributed as normals:

$$w_{ij} \overset{\text{indep.}}{\sim} \begin{cases} \text{Normal}(d_{ij}^{(n)}, \sigma_n^2), & s_{ij} = 0, \\ \text{Normal}(d_{ij}^{(s)}, \sigma_s^2), & s_{ij} = 1. \end{cases}$$

With these definitions, it is obvious that

$$\begin{aligned} Pr(y_{ij} = k) &= Pr(c_{i(k-1)} \le w_{ij} < c_{ik}) \\ &= \begin{cases} \Phi\left(\dfrac{d_{ij}^{(n)} - c_{ik}}{\sqrt{\sigma_n^2}}\right) - \Phi\left(\dfrac{d_{ij}^{(n)} - c_{i(k-1)}}{\sqrt{\sigma_n^2}}\right), & s_{ij} = 0, \\ \Phi\left(\dfrac{d_{ij}^{(s)} - c_{ik}}{\sqrt{\sigma_s^2}}\right) - \Phi\left(\dfrac{d_{ij}^{(s)} - c_{i(k-1)}}{\sqrt{\sigma_s^2}}\right), & s_{ij} = 1, \end{cases} \end{aligned}$$

which matches Eqs. (3) and (4). It is more convenient to derive full conditional posterior distributions conditioned on $\mathbf{w}$ than on $\mathbf{y}$.

5.3. Conditional posterior distributions

The conditional posterior distributions are provided by the following facts. The proofs are presented in the Appendix. Throughout, let $\theta \mid \cdot$ denote the full conditional posterior of parameter θ.

Fact 1. The conditional posterior distribution of w_{ij} is

$$w_{ij} \mid \cdot \overset{\text{indep}}{\sim} \begin{cases} \text{TN}_{(c_{i(y_{ij}-1)}, c_{i(y_{ij})})}(\mu^{(n)} + \alpha_i^{(n)} + \beta_j^{(n)}, \sigma_n^2), & s_{ij} = 0, \\ \text{TN}_{(c_{i(y_{ij}-1)}, c_{i(y_{ij})})}(\mu^{(s)} + \alpha_i^{(s)} + \beta_j^{(s)}, \sigma_s^2), & s_{ij} = 1, \end{cases}$$

where $\text{TN}_{(a,b)}(\mu, \sigma^2)$ is a Normal(μ, σ^2) distribution truncated below at a and above at b.

Please cite this article in press as: Morey, R. D., et al. Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. Journal of Mathematical Psychology (2008), doi:10.1016/j.jmp.2008.02.001

Fact 2. Let $N_n = \sum_{i,j}(1 - s_{ij})$ be the number of total trials in which nonstudied items are tested. Let $\boldsymbol{\lambda}_n$ be the vector of grand mean, participant, and item effects for nonstudied trials: $\boldsymbol{\lambda}_n = [\mu^{(n)}, (\boldsymbol{\alpha}^{(n)})^T, (\boldsymbol{\beta}^{(n)})^T]^T$. Let $\mathbf{w}_n$ be the vector of w_{ij} for nonstudied trials. Let $\mathbf{X}_n$ be the $N_n \times (I + J + 1)$ design matrix such that $\mathrm{E}[\mathbf{w}_n] = \mathbf{X}_n\boldsymbol{\lambda}_n$. Let $\boldsymbol{\Sigma}_w^{(n)} = \sigma_n^2\mathbf{I}$ denote the covariance matrix of $\mathbf{w}_n$. Finally, let $\boldsymbol{\Sigma}_\lambda^{(n)} = \mathrm{diag}(\sigma_\mu^2, \sigma_{\alpha^{(n)}}^2, \ldots, \sigma_{\beta^{(n)}}^2, \ldots)$. Then the full conditional posterior distribution of $\boldsymbol{\lambda}_n$ is

$$\boldsymbol{\lambda}_n \mid \boldsymbol{\Sigma}_w^{(n)}, \boldsymbol{\Sigma}_\lambda^{(n)}, \mathbf{w}_n \sim \mathrm{MVNormal}_q\left(\frac{1}{\sigma_n^2}\mathbf{V_n}\mathbf{X_n^T}\mathbf{w_n}, \mathbf{V_n}\right),$$

where $q = I + J + 1$ and $\mathbf{V}_n = \left(\frac{1}{\sigma_n^2}\mathbf{X_n^T}\mathbf{X_n} + (\boldsymbol{\Sigma}_\lambda^{(n)})^{-1}\right)^{-1}$.

Fact 3. Let N_s, $\boldsymbol{\lambda}_s$, $\mathbf{w_s}$, $\mathbf{X_s}$, and $\boldsymbol{\Sigma}_\lambda^{(s)}$ be defined analogously to the comparable quantities in Fact 2. Then the full conditional posterior distribution of $\boldsymbol{\lambda}_s$ is

$$\boldsymbol{\lambda}_s \mid \boldsymbol{\Sigma}_w^{(s)}, \boldsymbol{\Sigma}_\lambda^{(s)}, \mathbf{w_s} \sim \mathrm{MVNormal}_q\left(\frac{1}{\sigma_s^2}\mathbf{V}_s\mathbf{X_s^T}\mathbf{w_s}, \mathbf{V_s}\right),$$

where $\mathbf{V_s} = \left(\frac{1}{\sigma_s^2}\mathbf{X_s^T}\mathbf{X_s} + (\boldsymbol{\Sigma}_\lambda^{(s)})^{-1}\right)^{-1}$.

Fact 4. The full conditional posterior distributions of $\sigma_n^2, \sigma_s^2, \sigma_{\alpha^{(n)}}^2, \sigma_{\alpha^{(s)}}^2, \sigma_{\beta^{(n)}}^2, \sigma_{\beta^{(s)}}^2$ are

$$\sigma_n^2 \mid w_{ij}, \boldsymbol{\lambda} \sim \mathrm{IG}\left(a + \frac{1}{2}N_n, b + \frac{1}{2}\sum_{i,j}(1 - s_{ij})(w_{ij} - \mu^{(n)} - \alpha_i^{(n)} - \beta_j^{(n)})^2\right),$$

$$\sigma_s^2 \mid w_{ij}, \boldsymbol{\lambda} \sim \mathrm{IG}\left(a + \frac{1}{2}N_s, b + \frac{1}{2}\sum_{i,j}s_{ij}(w_{ij} - \mu^{(s)} - \alpha_i^{(s)} - \beta_j^{(s)})^2\right),$$

$$\sigma_{\alpha^{(n)}}^2 \mid w_{ij}, \boldsymbol{\lambda} \sim \mathrm{IG}\left(e + \frac{I}{2}, f + \frac{1}{2}\sum_i(\alpha_i^{(n)})^2\right),$$

$$\sigma_{\alpha^{(s)}}^2 \mid w_{ij}, \boldsymbol{\lambda} \sim \mathrm{IG}\left(e + \frac{I}{2}, f + \frac{1}{2}\sum_i(\alpha_i^{(s)})^2\right),$$

$$\sigma_{\beta^{(n)}}^2 \mid w_{ij}, \boldsymbol{\lambda} \sim \mathrm{IG}\left(e + \frac{J}{2}, f + \frac{1}{2}\sum_j(\beta_j^{(n)})^2\right),$$

$$\sigma_{\beta^{(s)}}^2 \mid w_{ij}, \boldsymbol{\lambda} \sim \mathrm{IG}\left(e + \frac{J}{2}, f + \frac{1}{2}\sum_j(\beta_j^{(s)})^2\right).$$

Fact 5. The full conditional posterior distribution of $c_{i\ell}$ is

$$c_{i\ell} \mid \cdot \overset{\text{indep.}}{\sim} \mathrm{Unif}\left(\max_j[w_{ij} \text{ such that } y_{ij} = \ell], \min_j[w_{ij} \text{ such that } y_{ij} = \ell + 1]\right), \quad \ell = 2, \ldots, K - 2. \tag{12}$$

5.4. Sampling algorithms

In Gibbs sampling, it is necessary to sample from the conditional posterior distributions. Fortunately, all of the conditional posterior distributions in Facts 1–5 are easy to sample from and details are provided in Rouder, Lu et al. (2007). Although sampling is straightforward, the resulting MCMC chains show a large degree of autocorrelation, especially in the criteria (see Fig. 5, panels A and B). A high degree of autocorrelation is undesirable because obtaining adequate convergence requires a large number of Gibbs sampling iterations.

The source of the autocorrelation may be diagnosed. The distribution of $c_{i\ell}|\cdot$ in Eq. (12) is uniform between two values: the maximum w_{ij} in the response category below the criterion, and the minimum w_{ij} in the response category above the criterion. As shown in Fig. 4, from iteration to iteration, the maximum latent value in response category k and minimum latent value in category $k + 1$ will be close to one another, restricting the range of the full conditional posterior distribution. Because the range of the distribution is restricted, samples of $c_{i\ell}$ will change very little from iteration to iteration in the Gibbs sampler. In Fig. 4, each response category contains 20 responses; the problem gets worse with more responses in each category. The result is high autocorrelation in the criteria chains, as shown in Fig. 5, panels A and B.

In order to mitigate this autocorrelation, we can modify the sampling scheme. Instead of sampling from the full conditional posterior of $c_{i\ell}$ in Fact 5, we sample from the conditional distribution

$$\int_{\mathbf{w}}[\mathbf{c_i}, \mathbf{w} \mid \boldsymbol{\lambda}_n, \boldsymbol{\lambda}_s, \sigma_n^2, \sigma_s^2, \mathbf{y}]\,\mathrm{d}\mathbf{w} = [\mathbf{c_i} \mid \boldsymbol{\lambda}_n, \boldsymbol{\lambda}_s, \sigma_n^2, \sigma_s^2, \mathbf{y}] \tag{13}$$

on every iteration of the Gibbs sampler. Integrating over the latent variables leads to more efficient MCMC chains (Holmes & Held, 2006). Let $c_{i\ell}^{(t)}$ be a sample of the ℓth criterion for the ith participant, on the tth iteration of the MCMC chain. On each iteration t, the following Metropolis–Hastings step is implemented to sample from (13).

Step 1. For each participant i, sample $K - 3$ independent values $z_{i2}^{(t)}, \ldots, z_{i(K-2)}^{(t)}$ from a Normal$(0, \sigma_d^2)$ distribution. The value σ_d^2 is chosen before analysis.

Step 2. Let $c_{i\ell}^{*(t)} = c_{i\ell}^{(t-1)} + z_{i\ell}^{(t)}$ for $\ell = 2, \ldots, K - 2$. The parameter $c_{i\ell}^{*(t)}$ serves as a proposal for a new sample of $c_{i\ell}$.

Step 3. For each participant, check that all proposal criteria are in the correct order, i.e. $c_{i\ell}^{*(t)} < c_{i(\ell+1)}^{*(t)}$ for $\ell = 1, \ldots, K - 1$. If the criteria are out of order for participant i, set $b_i = 0$ and skip Step 4 for participant i.

Step 4. For each participant, compute the likelihood of the model given the proposal criteria $c_{i\ell}^{*(t)}$, and the likelihood of

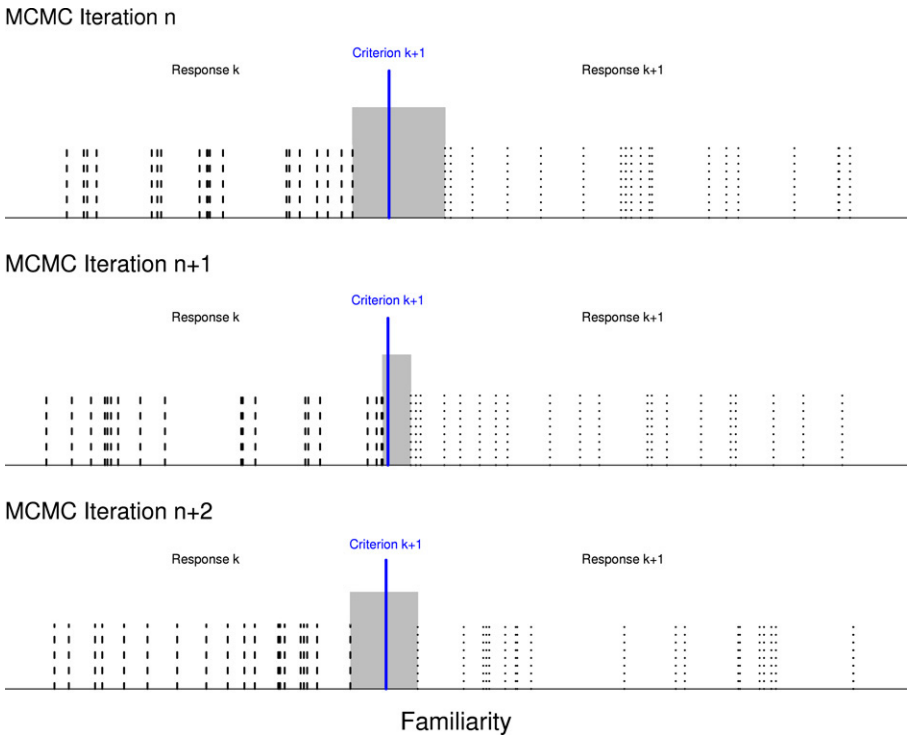

Fig. 4. Source of autocorrelation in criteria chains. Panels from top to bottom represent successive MCMC iterations. Dashed and dotted lines represent latent variables w_{ij} in categories k and $k+1$, respectively. The shaded box is a uniform distribution from which criterion $k+1$ is sampled.

the model given the samples $c_{i\ell}^{(t-1)}$, then compute its ratio, b_i:

$$b_i = \prod_{j=1}^{J} \left[\frac{\Phi\left(\frac{c^{*(t)}_{i(y_{ij})} - d^{(s)}_{ij}}{\sqrt{\sigma_s^2}}\right) - \Phi\left(\frac{c^{*(t)}_{i(y_{ij}-1)} - d^{(s)}_{ij}}{\sqrt{\sigma_s^2}}\right)}{\Phi\left(\frac{c^{(t-1)}_{i(y_{ij})} - d^{(s)}_{ij}}{\sqrt{\sigma_s^2}}\right) - \Phi\left(\frac{c^{(t-1)}_{i(y_{ij}-1)} - d^{(s)}_{ij}}{\sqrt{\sigma_s^2}}\right)} \right]^{s_{ij}} \times \prod_{j=1}^{J} \left[\frac{\Phi\left(\frac{c^{*(t)}_{i(y_{ij})} - d^{(n)}_{ij}}{\sqrt{\sigma_n^2}}\right) - \Phi\left(\frac{c^{*(t)}_{i(y_{ij}-1)} - d^{(n)}_{ij}}{\sqrt{\sigma_n^2}}\right)}{\Phi\left(\frac{c^{(t-1)}_{i(y_{ij})} - d^{(n)}_{ij}}{\sqrt{\sigma_n^2}}\right) - \Phi\left(\frac{c^{(t-1)}_{i(y_{ij}-1)} - d^{(n)}_{ij}}{\sqrt{\sigma_n^2}}\right)} \right]^{1-s_{ij}}. \quad (14)$$

Step 5. For each participant, accept the proposal criteria $\mathbf{c}^{*(t)}$ as the new sample for the criteria $\mathbf{c}^{(t)}$ with probability $\min(b_i, 1)$. Otherwise, let $\mathbf{c}^{(t)} = \mathbf{c}^{(t-1)}$.

Because the expression b_i is not dependent on latent data $\mathbf{w}$, this source of autocorrelation is eliminated. Fig. 5, panels C and D, show a sample of a chain for a selected criterion and its autocorrelation function. As can be seen, the new sampling scheme is highly effective.

The value of σ_d^2 must be chosen before the analysis. Some values of σ_d^2 are more effective than others. If the value is too high, values of $c_{i\ell}^{*(t)}$ are unreasonably large or small, and the candidate is rejected too often. If the value is too low, candidates do not deviate enough from the last sample and the chains will be highly autocorrelated. We recommend that researchers experiment with several values of σ_d^2 until an acceptance rate (Step 5) between 35% and 45% is achieved. In our experience this experimentation takes a matter of minutes even in large-sized samples.

Code in the C language and a precompiled Windows executable for analyzing data with the hierarchical model are available at www.missouri.edu/~pcl/code/.

5.5. Simulations

To test the performance of the model, we performed two simulation studies. The first simulation was designed to assess whether the model yields accurate estimates of signal detection parameters. The data were generated from and analyzed with the fixed-criteria hierarchical signal-detection model. A second simulation was designed to assess whether model analysis is robust to psychologically plausible misspecification. In the model for generating data, item and participant baseline effects were distributed as normals, but the mnemonic gain from study was distributed as an exponential distribution. Data generated this way violate the assumptions in (9) and (10).

In both simulations, hypothetical confidence rating data were choices among $K = 4$ options. In each hypothetical data set, 50 participants were tested on 100 items, half of which were studied and half of which were not. The design matrix was appropriately counterbalanced to ensure identifiability (see Christensen, 1996). For each simulation, the process of simulating data sets and estimating parameters was repeated 200 times.

We assessed how well the model recovered item and participant effect parameters, criteria, overall sensitivity (d' in Eq. (11)), and the ratio of standard deviations, denoted by $\eta = \sqrt{\sigma_s^2/\sigma_n^2}$. The ratio η corresponds to σ in the conventional

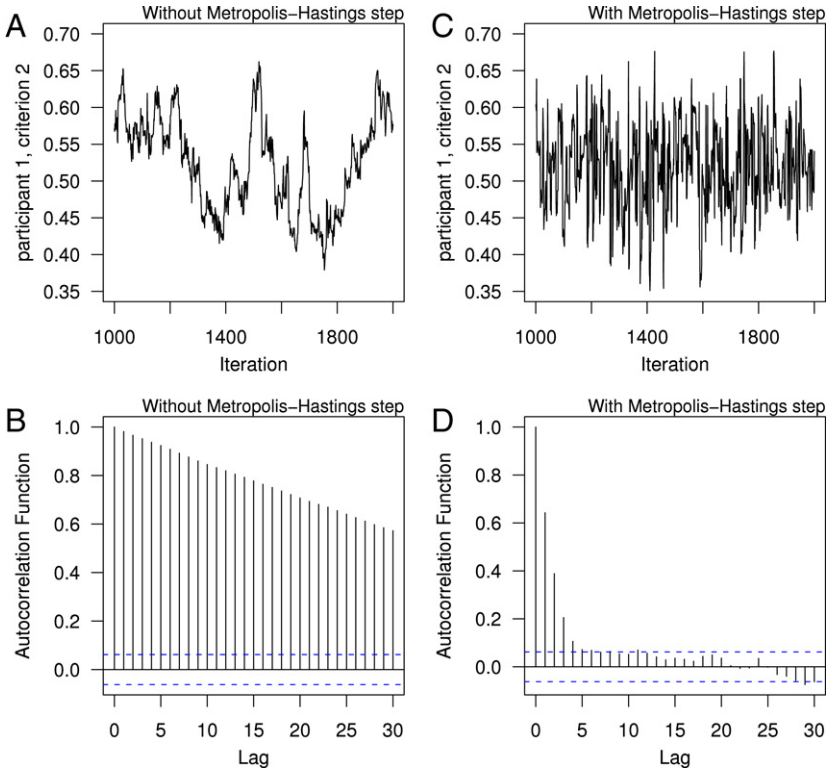

Fig. 5. A: Segment of the MCMC chain for a selected criterion with Gibbs sampling. B: Autocorrelation function for the chain with Gibbs sampling. C-D: Same as A-B, with Metropolis–Hastings step.

SDT parametrization and the question at hand is whether η is biased upward.

Simulation 1A. In Simulation 1A, data were generated from the model with the following true values: $d^{(n)} = -.5$, $d^{(s)} = 1.5$, $\sigma^2_{\alpha^{(n)}} = \sigma^2_{\alpha^{(s)}} = \sigma^2_{\beta^{(n)}} = \sigma^2_{\beta^{(s)}} = .04$, and $\sigma^2_n = \sigma^2_s = 1$. Each participant's sole free criterion c_2 was sampled from a beta distribution with parameters $\alpha = \beta = 10$. Most of the mass of this beta distribution is distributed between .3 and .7.

Simulation 1B. Simulation 1B was identical to Simulation 1A, except the true value of σ^2_s was set to 1.44. This choice is motivated by the benchmark finding in the field that the standard deviation of the studied-item distribution is often about 1.2 times that of the nonstudied-item distribution.

Figs. 6 and 7A–C show the results of Simulation 1A. Each panel in Fig. 6 shows estimates of participant or item effects as a function of true values. The points lie close to the diagonal, indicating that these random effects are accurately recovered. Fig. 7A shows estimates of criterion c_{i2}. The model is able to accurately recover criteria values. Fig. 7B shows how well overall sensitivity (d') is estimated. The thick center points denote posterior means of d' across the 200 replicates. These are ordered and plotted as a CDF function. For example, the median value of the posterior mean across the 200 replicates is 2.02. The plot also shows 95% credible intervals (from 2.5% to 97.5%) on d'. For the replicate corresponding to the median posterior mean, the credible interval is $(1.82, 2.22)$. The points to the left and right of the thick point at 2.02 denote the endpoints of this credible interval. Therefore, the plot shows how often the credible interval covers the true value (denoted by a vertical line). We refer to these plots as *CDF-coverage plots*. Two points are clear: (1) the posterior mean estimate of d' is centered around the true value, and (2) the 95% credible intervals cover the true values approximately 95% of the time. Therefore, not only is the estimate accurate, the model accurately recovers the degree of uncertainty in the estimate. Fig. 7C shows the CDF-coverage plot for estimates of η. Once again, the estimate is accurate and the coverage is good. Therefore, parameter estimation from the hierarchical model is highly satisfactory.

The results of Simulation 1B were equally good and the majority of the plots are omitted for brevity. Fig. 7D shows the CDF-coverage plot for estimates of η. In Simulation 1B, the true value of η was 1.2 and the recovery is good. A comparison of Fig. 7C and D reveal that even with moderate sample sizes, the model can easily distinguish between the equal-variance signal detection model and the inflated-variance model suggested by the benchmark result of increased studied-item variance. Therefore, the hierarchical model is well-suited for assessing the true ratio of variances.

Simulation 2. Simulation 2 was designed to assess the robustness of the model to a psychologically plausible misspecification. All parameters had the same true values as Simulation 1A with the following exception:

$$d^{(s)}_{ij} = d^{(n)}_{ij} + E^{(\alpha)}_i + E^{(\beta)}_j,$$

where $E^{(\alpha)}_i$ and $E^{(\beta)}_j$ are all independent and identically distributed standard exponential random variables. This data-generation model makes the realistic assumption that the

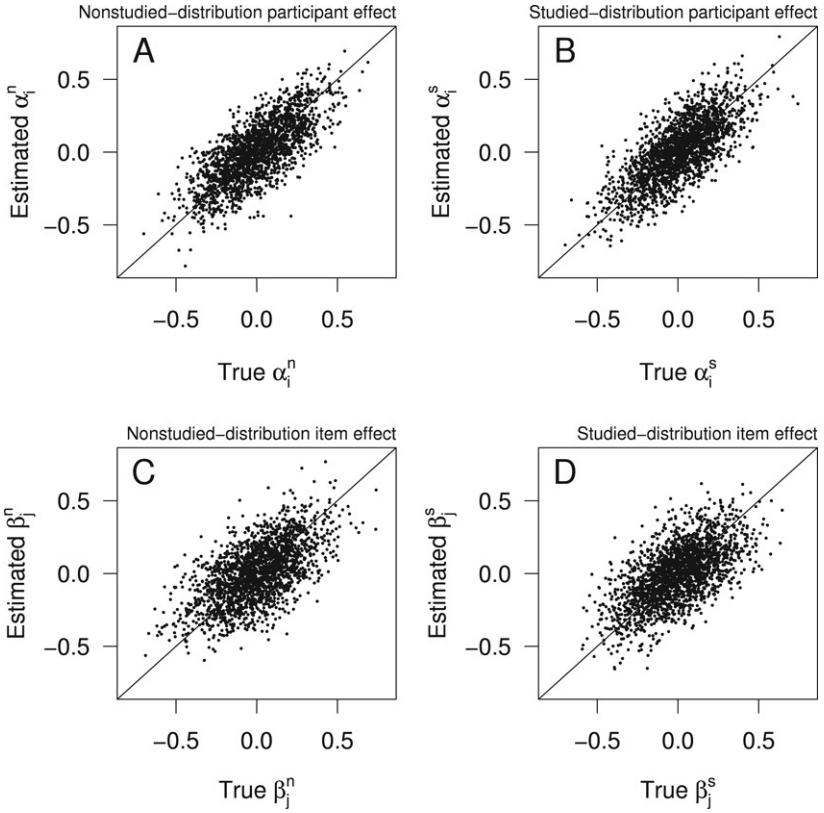

Fig. 6. Estimates of participant and item effects as a function of true value in Simulation 1A. Scatterplots of participant and item effects were thinned by factors of 5 and 10, respectively.

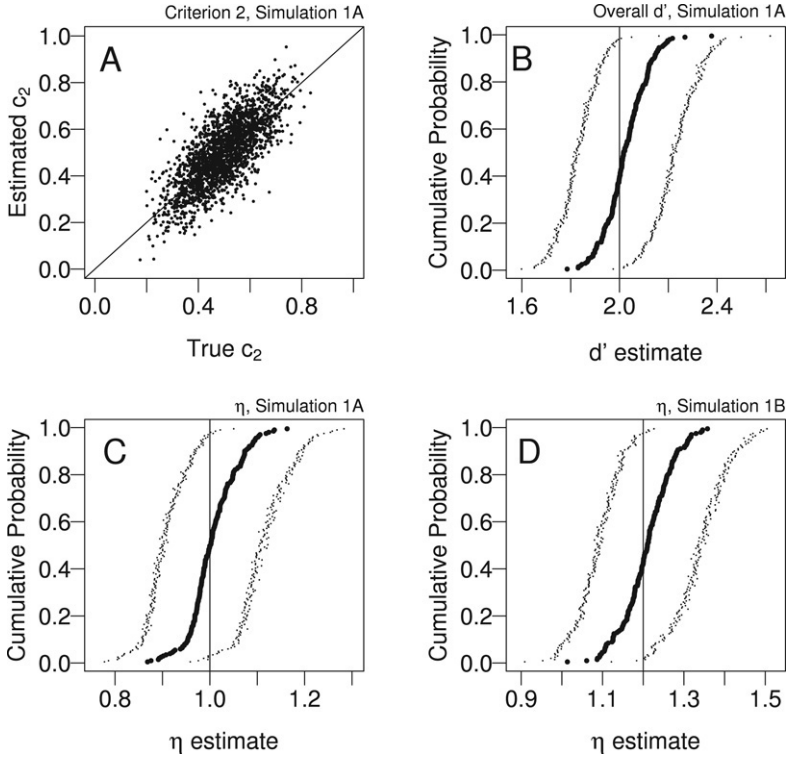

Fig. 7. Results of simulations. A: Estimates of middle criterion c_2 as a function of true value for Simulation 1A. The scatterplot was thinned by a factor of 5. B: CDF-coverage plot of overall sensitivity d'. Thick points are posterior means of d' arranged as a cumulative distribution function across the 200 replicates. The points to the sides of the posterior means denote the endpoints of the 95% credible interval sorted by posterior mean. The solid vertical line is the true value. C and D: CDF-coverage plots of η in Simulation 1A (true value of $\eta = 1$) and Simulation 1B (true value of $\eta = 1.2$), respectively.

Please cite this article in press as: Morey, R. D., et al. Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. Journal of Mathematical Psychology (2008), doi:10.1016/j.jmp.2008.02.001

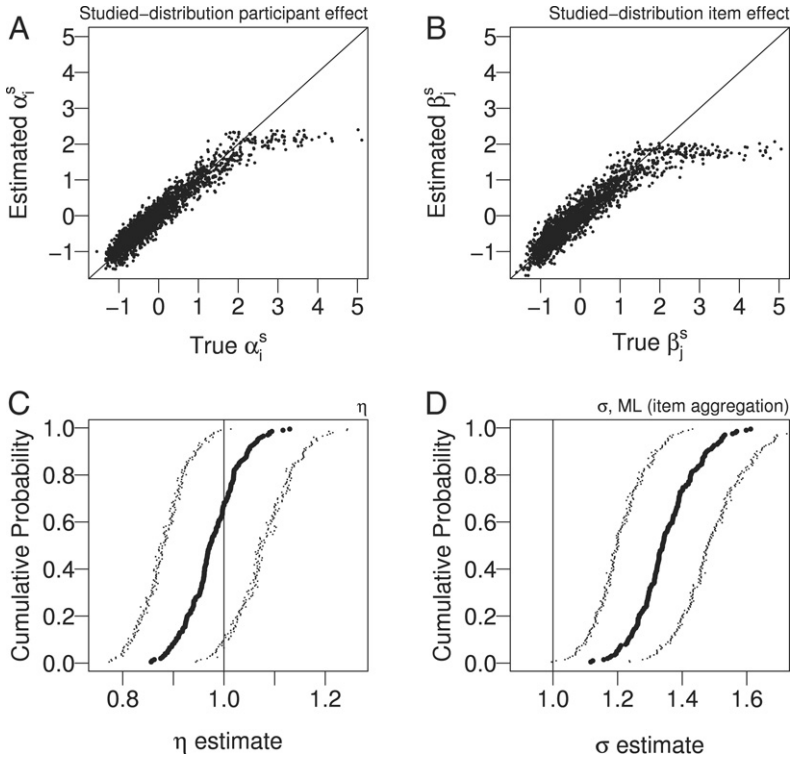

Fig. 8. Results of Simulation 3. A: Estimates of $\alpha_i^{(s)}$ as a function of true value. B: Estimates of $\beta_j^{(s)}$ as a function of true value. C: CDF-coverage plot of η. The solid vertical line is the true value. D: CDF-coverage plots of item-aggregated ML estimates of σ with 95% confidence intervals. Scatterplots of participant and item effects were thinned by factors of 5 and 10, respectively.

increase in strength from study can never be negatively valued, and is similar to an informal model suggested by Wixted (2007).

Fig. 8 shows the results of Simulation 2. Panels A and B show estimates of $\alpha_i^{(s)}$ and $\beta_j^{(s)}$ as a function of the zero-centered true value. Most random effects are recovered accurately with poor recovery only for extremely large values. This inaccuracy is the direct result of misspecification—the exponential distribution has more skew than the normal assumed in the model. Importantly, this misspecification affects the estimates of η only slightly. Panel C shows the CDF-coverage plot of η estimates. The mean estimate of η shows a small under-estimation of about 2.4%. This underestimation is non-asymptotic; estimates of all parameters will converge to the true value in the large-sample limit. To demonstrate that this convergence is sufficiently rapid, we ran an additional simulation with 400 items per participant and found the bias in estimating η was reduced to. 2%. An experiment with 400 items is large, but not unrealistic. Glanzer et al. (1999), for example, report an experiment with 480 items per participant.

We also estimated standard signal detection model parameters using maximum likelihood applied to item-aggregated data. Fig. 8D shows the CDF-coverage plots for the ML estimate of σ. The intervals between the small points denote the 95% confidence intervals computed from the Fisher information (Rice, 1998). The mean estimate of σ is 1.35, far above the true value of 1. In fact, only a single confidence interval in 200 (.5%) includes the true value. Unlike the small bias in the hierarchical Bayesian estimates, this bias is not only extreme; it is asymptotic. The reason for the poor performance of ML is aggregation. In this simulation, the item effects are more variable for studied than nonstudied items. The artifactual increase in $\hat{\sigma}$ in the ML analysis is due to an inability to account for this increased variability. The aggregated analysis includes item variability in the estimate of σ, severely inflating the estimate.

5.6. General discussion

In this paper, we have provided an extension of Rouder and Lu's (2005) Bayesian hierarchical signal detection model to account for participant, item, and process variability in confidence-rating paradigms. We recommend researchers use this approach as an alternative to aggregation. There is one caveat needed, however, for application. There is a subtle though important assumption in the fixed-criteria parametrization. This assumption is most easily discussed by contrast with a different and more general fixed-criterion model. Consider a parametrization in which the interior criterion is fixed to 0 and the standard deviation of the non-studied familiarity distribution is fixed to 1.0. This model has the same likelihood as the one we advocate; the natural hierarchical priors, however, are more general than the model we advocate. This generality may be seen by comparing the number of parameters, including those in the prior. In the model we advocate, there are $2IJ + (K-3)I + 2$ parameters. In the hierarchical extension of this alternative fixed-criteria model, there are $2IJ + I(K-2) + 1$ parameters. The difference is $I - 1$ criterion parameters. This reduction in parameters is achieved by assuming common scaling of criteria across

Please cite this article in press as: Morey, R. D., et al. Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. Journal of Mathematical Psychology (2008), doi:10.1016/j.jmp.2008.02.001

people. With the more general priors, there is a noticeable bias toward higher values of σ^2. The advocated model does not have this bias. We do not know if the common scaling assumption is warranted and urge practitioners to check residuals for evidence of misspecification (see, for example, Morey, Rouder, and Speckman (2008).

The presented critique of aggregation should not be viewed as an idiosyncrasy of signal detection. The critique that aggregation distorts inference has a long history in mathematical psychology. Examples include the critique that aggregation affects the form of learning curves (Estes, 1956; Heathcote, Brown, & Mewhort, 2000), the fit of similarity-choice models (Ashby, Maddox, & Lee, 1994), and the assessment of selective influence in process dissociation memory models (Curran & Hintzman, 1995; Rouder et al., in press). These critiques should be viewed as reflecting the same underlying problem. Models of psychological process are typically nonlinear. The assumption tacit in aggregation is that aggregated data may be used to estimate averaged parameter values. In nonlinear models, however, model outputs of averaged parameter values are never the average of outputs from individual parameters. Whereas all of these critiques reflect the same underlying problem, process-oriented hierarchical models such as the one presented here are broadly applicable in many domains.

Acknowledgments

This research is supported by NSF grant SES-0351523 and NIMH grant R01-MH071418.

Appendix

A.1. Joint posterior density

Let $\mathbf{Y}$ be the collection of all data and $\boldsymbol{\theta}$ be the collection of all model parameters. The joint posterior density of all parameters given the data $\mathbf{Y}$ is

$$
\begin{aligned}
[\boldsymbol{\theta} \mid \mathbf{Y}] \propto & \prod_{i=1}^{I}\prod_{j=1}^{J}\left[\sum_{k=1}^{K} I_{(y_{ij}=k)} I_{(c_{i(k-1)}<w_{ij}<c_{ik})}\right] \\
& \times (\sigma_n^2)^{-\frac{N_n}{2}} \exp\left\{-\frac{1}{2\sigma_n^2}(\mathbf{w}_n - \mathbf{X}_n\boldsymbol{\lambda}_n)^T(\mathbf{w}_n - \mathbf{X}_n\boldsymbol{\lambda}_n)\right\} \\
& \times (\sigma_s^2)^{-\frac{N_s}{2}} \exp\left\{-\frac{1}{2\sigma_s^2}(\mathbf{w}_s - \mathbf{X}_s\boldsymbol{\lambda}_s)^T(\mathbf{w}_s - \mathbf{X}_s\boldsymbol{\lambda}_s)\right\} \\
& \times |\boldsymbol{\Sigma}_\lambda^{(n)}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\lambda}_n^T(\boldsymbol{\Sigma}_\lambda^{(n)})^{-1}\boldsymbol{\lambda}_n\right\} \\
& \times |\boldsymbol{\Sigma}_\lambda^{(s)}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\lambda}_s^T(\boldsymbol{\Sigma}_\lambda^{(s)})^{-1}\boldsymbol{\lambda}_s\right\} \\
& \times (\sigma_n^2)^{-(a+1)} \exp\left\{-\frac{b}{\sigma_n^2}\right\} (\sigma_s^2)^{-(a+1)} \exp\left\{-\frac{b}{\sigma_s^2}\right\} \\
& \times (\sigma^2_{(\alpha^{(n)})})^{-(e+1)} \exp\left\{-\frac{f}{\sigma^2_{\alpha^{(n)}}}\right\} \\
& \times (\sigma^2_{(\alpha^{(s)})})^{-(e+1)} \exp\left\{-\frac{f}{\sigma^2_{\alpha^{(s)}}}\right\} \\
& \times (\sigma^2_{(\beta^{(n)})})^{-(e+1)} \exp\left\{-\frac{f}{\sigma^2_{\beta^{(n)}}}\right\} \\
& \times (\sigma^2_{(\beta^{(s)})})^{-(e+1)} \exp\left\{-\frac{f}{\sigma^2_{\beta^{(s)}}}\right\}. \qquad (15)
\end{aligned}
$$

A.2. Full conditionals

Proof of Fact 1. Inspection of (15) reveals that

$$
\begin{aligned}
[w_{ij} \mid \cdot] & \left[\sum_{k=1}^{K} I_{(y_{ij}=k)} I_{(c_{i(k-1)}<w_{ij}<c_{ik})}\right] \\
& \times \exp\left\{-\frac{1-s_{ij}}{2\sigma_n^2}\left(w_{ij} - \mu^{(n)} - \alpha_i^{(n)} - \beta_j^{(n)}\right)^2\right\} \\
& \times \exp\left\{-\frac{s_{ij}}{2\sigma_s^2}\left(w_{ij} - \mu^{(s)} - \alpha_i^{(s)} - \beta_j^{(s)}\right)^2\right\}.
\end{aligned}
$$

Using the definitions of $d_{ij}^{(n)}$ and $d_{ij}^{(s)}$,

$$
\begin{aligned}
\left[w_{ij} \mid s_{ij}=0, \cdot\right] & \propto \exp\left\{-\frac{1}{2\sigma_n^2}\left(w_{ij} - d_{ij}^{(n)}\right)^2\right\} \\
& \quad \times \sum_k \left[I_{(y_{ij}=k)} I_{(c_{i(k-1)}<w_{ij}<c_{ik})}\right], \\
\left[w_{ij} \mid s_{ij}=1, \cdot\right] & \propto \exp\left\{-\frac{1}{2\sigma_s^2}\left(w_{ij} - d_{ij}^{(s)}\right)^2\right\} \\
& \quad \times \sum_k \left[I_{(y_{ij}=k)} I_{(c_{i(k-1)}<w_{ij}<c_{ik})}\right].
\end{aligned}
$$

Because the sum can only be nonzero if $y_{ij}=k$, the right hand side can be further simplified as

$$
\begin{aligned}
\left[w_{ij} \mid s_{ij}=0, \cdot\right] & \propto \exp\left\{-\frac{1}{2\sigma_n^2}\left(w_{ij} - d_{ij}^{(n)}\right)^2\right\} \\
& \quad \times I_{(c_{i(y_{ij}-1)}<w_{ij}<c_{i(y_{ij})})}, \\
\left[w_{ij} \mid s_{ij}=1, \cdot\right] & \propto \exp\left\{-\frac{1}{2\sigma_s^2}\left(w_{ij} - d_{ij}^{(s)}\right)^2\right\} \\
& \quad \times I_{(c_{i(y_{ij}-1)}<w_{ij}<c_{i(y_{ij})})}.
\end{aligned}
$$

The right-hand side expressions are proportional to the density functions for the corresponding truncated normal distributions in Eq. (12). □

Proof of Fact 2 and 3. Proof of Fact 2 and 3 are standard and may be found in Gelman, Carlin, Stern, and Rubin (2004). □

Proof of Fact 4. Proof of Fact 4 is standard. For example, see Rouder, Morey, Speckman, and Pratte (2007). □

Proof of Fact 5. The proof may be found in Albert and Chib (1993). □

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*, 144–151.

Chen, M. H., & Shao, Q. M. (1998). Monte carlo methods for Bayesian analysis of constrained parameter problems. *Biometrika*, *85*, 73–87.

Christensen, R. (1996). *Plane answers to complex questions: The theory of linear models* (2nd ed.). Berlin: Springer-Verlag.

Curran, T. C., & Hintzman, D. L. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 531–547.

Dorfman, D., & Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals-rating-method data. *Journal of Mathematical Psychology*, *6*, 487–496.

Estes, W. K. (1956). The problem of inference from curves based on grouped data. *Psychological Bulletin*, *53*, 134–140.

Gelfand, A., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman and Hall.

Glanzer, M., & Adams, J. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *16*, 5–16.

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley, Reprinted by Krieger, Huntington, NY, 1974.

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

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review*, *7*, 185–207.

Heathcote, A., Raymond, F., & Dunn, J. (2006). Recollection and familiarity in recognition memory: Evidence from ROC curves. *Journal of Memory and Language*, *55*, 495–514.

Hintzman, D. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Holmes, C. C., & Held, L. (2006). Bayesian auxiliary variable models for binary and polychotomous regression. *Bayesian Analysis*, *1*, 145–168.

Kintsch, W. (1967). Memory and decision aspects of recognition learning. *Psychological Review*, *74*, 496–504.

Malmberg, K. J. (2002). On the form of rocs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 380–387.

Malmberg, K. J., & Xu, J. (2006). The influence of averaging and noisy decision strategies on the recognition memory ROC. *Psychonomic Bulletin & Review*, *13*, 99–105.

Morey, R. D., Rouder, J. N., & Speckman, P. L. (2008). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology*, *52*, 21–36.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*, 609–626.

Murdock, B. B. (1993). Todam2: A model for the storage and retrieval of item, association, and serial order information. *Psychological Review*, *100*, 183–203.

Peters, H. (1936). The relationship between familiarity words and their memory value. *American Journal of Psychology*, *48*, 572–584.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *20*, 763–785.

Ratcliff, R., Sheu, C. F., & Grondlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*, 518–535.

Rice, J. (1998). *Mathematical statistics and data analysis*. Monterey, CA: Brooks/Cole.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin and Review*, *12*, 573–604.

Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. A hierarchical process dissociation model. *Journal of Experimental Psychology: General* (in press).

Rouder, J. N., Lu, J., Sun, D., Speckman, P. L., Morey, R. D., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Pratte, M. S. (2007). Detecting chance: A solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin and Review*, *14*, 597–605.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 1–17.

Schwartz, F., & Rouse, R. (1961). The activation and recovery of associations. *Psychological Issues*, *3*, 1–141.

Singer, M., Gagnon, N., & Richard, E. (2002). Strategies of text retrieval: A criterion shift account. *Canadian Journal of Experimental Psychology*, *56*, 41–57.

Shepard, R. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*, 156–163.

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

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

Yonelinas, A. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *25*, 1415–1434.

Yonelinas, A., & Parks, C. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*, 800–832.

Please cite this article in press as: Morey, R. D., et al. Problematic effects of aggregation in *z*ROC analysis and a hierarchical modeling solution. Journal of Mathematical Psychology (2008), doi:10.1016/j.jmp.2008.02.001