New Techniques to Reduce Observer Inconsistency in Psychoacoustic Experiments

Judi A. Lapsley Miller*

Psychophysics Laboratory,
Victoria University of Wellington, New Zealand

Abstract

One of the biggest problems still facing experimental psychophysics is observer inconsistency. Observer inconsistency degrades performance, hampering comparisons with other observers and theoretical predictions. By repeating an experiment multiple times using the same stimuli, observer inconsistency can be reduced and asymptotic errorless performance estimated by using group operating characteristic (GOC) analysis and function of replications combined estimation (FORCE) analysis, respectively. These new techniques are described and illustrated with results from various psychoacoustic experiments, showing that humans are sometimes capable of performing as well as an ideal observer once observer inconsistency is reduced.

Acknowledgements

* Now at the Naval Submarine Medical Research Lab, Groton.
Thanks to Linton Miller and Vit Drga for their valuable advice and help with data analysis, and to NSMRL for their support and the use of their resources.

Why Is Observer Inconsistency a Problem?

- Observer inconsistency occurs when an observer responds differently to repeated presentations of the same stimulus
- Inconsistency comes from noise in the observer's environment or from within the observer, e.g.:
  - Continuous background noise masker
  - Heartbeat, muscle tension, neural noise
  - Memory, inattention, and coordination when using response manipulanda
  - Transient external noise: cars, aircraft, voices
- Inconsistency degrades performance, making it hard to compare an observer's performance to theoretical models and to other observers
  - It is difficult, if not impossible, to control for all noise sources, or account for each of their effects specifically (i.e., by modeling the noise)
  - Another approach is to reduce or remove the effects of observer inconsistency by repeating the experiment multiple times and averaging out the error

An example of observer inconsistency

- 4 observers each repeated a frequency discrimination experiment 16 times
- Shown are the 64 empirical ROC curves and the theoretical ROC curve
- The empirical ROC curves are below the theory, and are a different shape
- The empirical ROC curves are highly variable, despite using identical stimuli on each replication
  - SIFC task; 64 point rating scale
  - Events were High tones and Low tones, with an overlap (even an ideal observer could not perform perfectly)
  - Continuous background masker to make task more difficult
  - The theoretical ROC curve is based on discrete uniform distributions

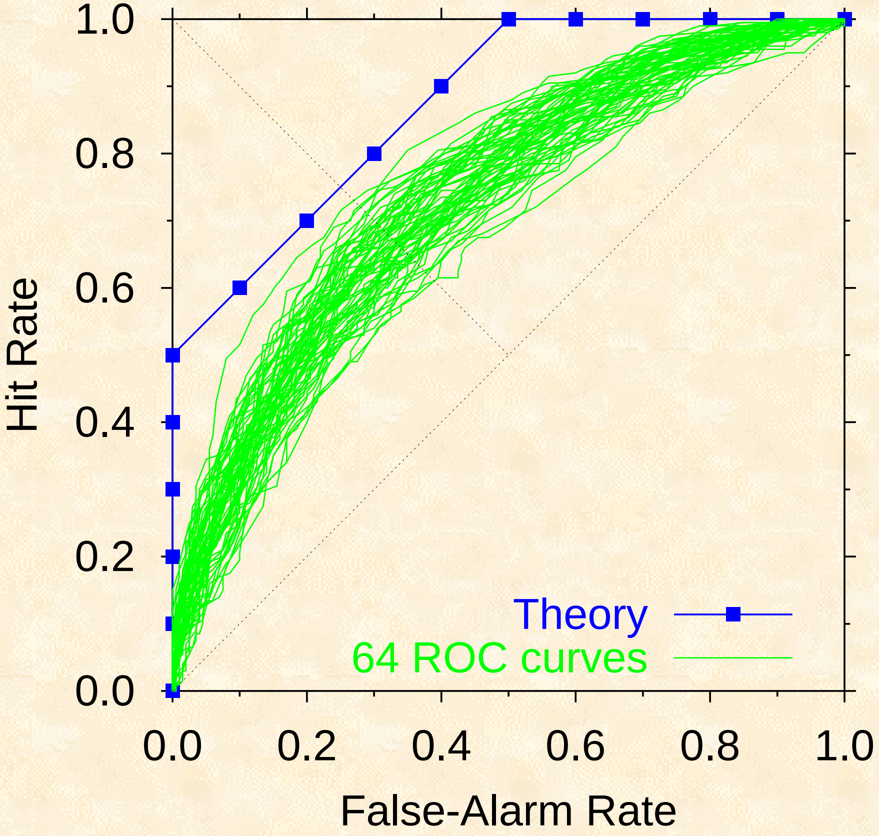

Fig 1: An example of observer inconsistency when repeating the same experiment 64 times

Reprinted with permission from Lapsley Miller, Scurfield, Drga, Galvin, & Whitmore (1999)

What can be done to reduce inconsistency?

- One technique for reducing the effects of observer inconsistency is Group Operating Characteristic (GOC) Analysis
  - Introduced by Watson (1963/64) and developed by Taylor, Boven, & Whitmore (1991) and Drga (1999).
  - GOC analysis has successfully been used in a variety of experiments, including amplitude discrimination of tones and noise, frequency discrimination, and Type II decision making.
  - GOC analysis can be used for subject groups as well as individuals
- The result of GOC analysis is a GOC curve in the ROC space, from which the usual measures of detectability may be taken.
- GOC analysis tends to improve performance in the ROC space, compared to the mean ROC curve
- Drga (1999) has developed a theory of GOC analysis; here the focus is on the practical application of this technique.
- The key is to average out the observer inconsistency in the decisions (ratings) for each stimulus. To do this, the experiment is repeated multiple times *using identical stimuli*
  - Both the signal and the masker need to be identical on each presentation.
  - Making identical acoustic stimuli is possible - even for noise maskers - with digital signal generation techniques.
- When ratings for each stimulus are averaged across replications, noise that is not common across all replications is averaged out.
  - this is *not* the same as calculating a mean ROC curve, where the averaging occurs across hit and false-alarm rates, rather than across ratings

How is GOC analysis done? (Part 1)

- The fictitious example in Table 1 shows the results of 4 replications of an experiment
  - There were three events: one noise-alone (with 10 stimuli) and two signal-plus-noise (with 5 stimuli each)
  - Decisions were indicated on a 9 point rating scale (see Table 1, columns 3-6)
  - Notice that the observer often made a different decision in response to the same stimulus
- For each stimulus, the observer's decisions (ratings) are averaged across replications (see Table 1, column 7)
  - Any type of averaging may be used, including summing (which is computationally more efficient)

Stimulus	Event	Rating				Sum-of-rating
		Rep 1	Rep 2	Rep 3	Rep 4	
1	N	3	5	8	1	17
2	N	1	2	2	4	9
3	N	4	3	5	2	14
4	N	2	4	5	2	13
5	N	6	1	2	6	15
6	N	1	2	1	1	5
7	N	5	4	6	3	18
8	N	1	1	4	1	7
9	N	9	3	4	1	17
10	N	1	1	4	2	8
1	SN_1	4	1	1	3	9
2	SN_1	4	5	5	6	20
3	SN_1	8	6	5	1	20
4	SN_1	9	4	2	1	16
5	SN_1	4	5	4	2	15
1	SN_2	6	1	3	4	14
2	SN_2	9	8	5	8	30
3	SN_2	7	7	9	9	32
4	SN_2	9	7	7	7	30
5	SN_2	8	6	4	8	26

Table 1: Ratings from 4 replications of a fictitious GOC experiment.

How is GOC analysis done? (Part 2)

- From here, analysis proceeds like a rating-scale ROC analysis
- The number of times each sum-of-ratings occurred is tallied
  - This tally is done separately for each event (see Table 2, columns 2-4)
  - The table is sorted so the sum-of-ratings are in order
- The tallies are cumulated, using each sum-of-ratings as a criterion, from highest sum-of-ratings to lowest (see Table 2, columns 5-7)
- Hit and False-Alarm Rates for the GOC curve are calculated by dividing each cumulative tally by the number of stimuli for that event (see Table 2, columns 8-10)

Sum-of-rating	Tally			Cumulative Tally			Rate		
	N	SN1	SN2	N	SN1	SN2	FAR	HR1	HR2
5	1	0	0	10	5	5	1.0	1.0	1.0
7	1	0	0	9	5	5	0.9	1.0	1.0
8	1	0	0	8	5	5	0.8	1.0	1.0
9	1	1	0	7	5	5	0.7	1.0	1.0
13	1	0	0	6	4	5	0.6	0.8	1.0
14	1	0	1	5	4	5	0.5	0.8	1.0
15	1	1	0	4	4	4	0.4	0.8	0.8
16	0	1	0	3	3	4	0.3	0.6	0.8
17	2	0	0	3	2	4	0.3	0.4	0.8
18	1	0	0	1	2	4	0.1	0.4	0.8
20	0	2	0	0	2	4	0.0	0.4	0.8
26	0	0	1	0	0	4	0.0	0.0	0.8
30	0	0	2	0	0	3	0.0	0.0	0.6
32	0	0	1	0	0	1	0.0	0.0	0.2
Total	10	5	5						

Table 2: Calculation of Hit and False-Alarm Rates from the sum-of-ratings from a fictitious GOC experiment.

An example of a GOC curve

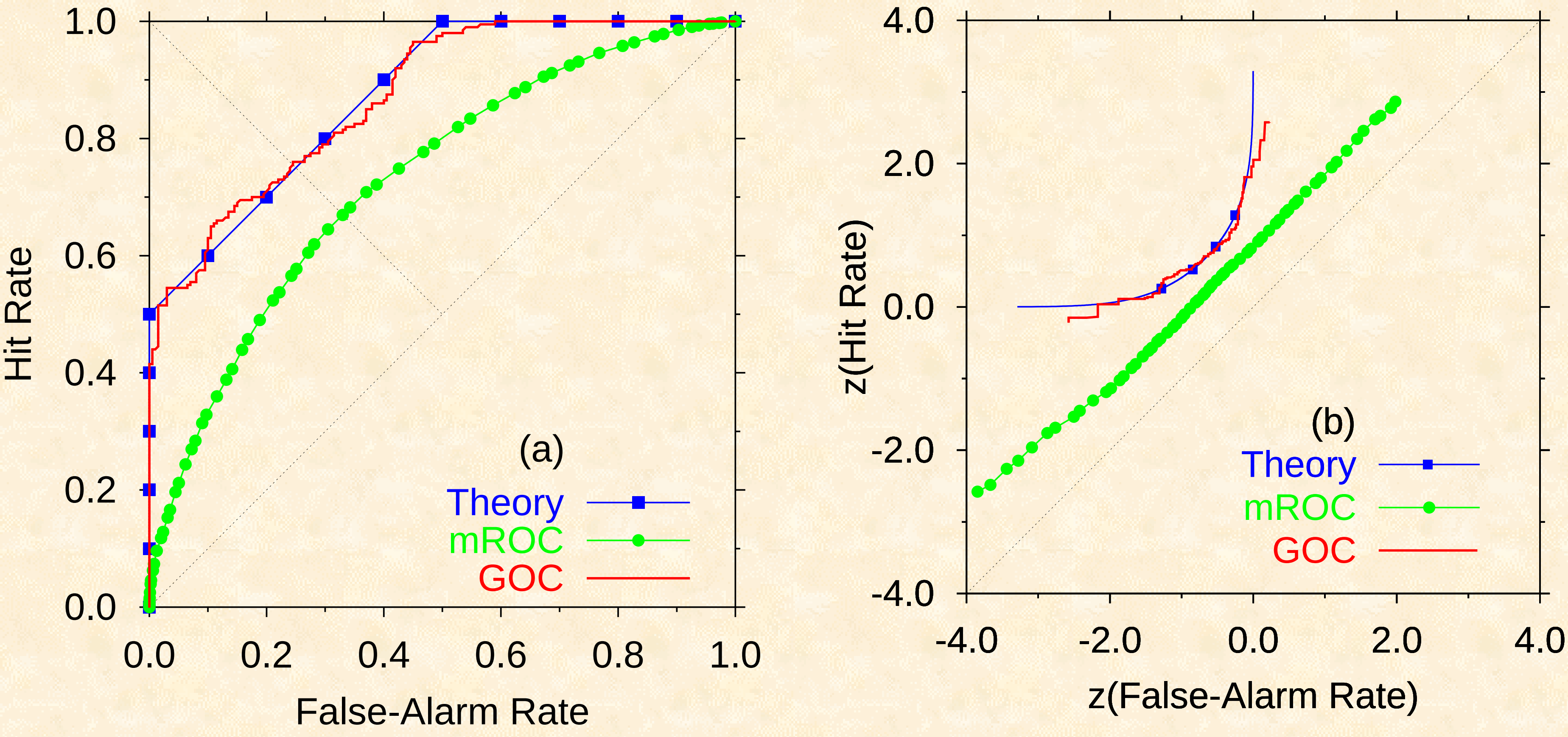

Fig 2. A GOC curve based on 64 replications of the experiment described in Fig 1. Also shown is the theoretical ROC curve and the mean ROC curve. (a) linear coordinates (b) normal coordinates. Reprinted with permission from Lapsley Miller et al. (1999)

What can GOC analysis do?

- Fig 2(a) shows the GOC curve based on the 64 replications shown in Fig 1.
  - Also shown is the mean ROC curve - calculated by averaging the hit and false-alarm rates at each criterion
- The GOC curve approximates the theoretical ROC curve well. In comparison, the mean ROC curve shows the average, inconsistency-degraded, performance
- In this case, the GOC analysis indicates that the observers' are able to perform as well as an ideal observer, once error due to their inconsistency is removed
- Fig 2(b) shows the same ROC curves in Normal coordinates. This coordinate system highlights the difference between the mROC and GOC curves
  - If only single replication, or mean ROC curves, were obtained, then it may be incorrectly concluded that the underlying distributions in this task were Normals with equal variance. In this case, because the task was contrived, we know that the underlying distributions are discrete Uniforms with equal variance, which is reflected accurately by the GOC curve
  - This tendency for inconsistency degraded performance to look Normal has meant that many researchers now assume that the Normal model is appropriate for virtually all detection tasks. GOC analysis can help in establishing which theory is more appropriate, if the competing theories predict different ROC curves.

How many replications are needed?

- How many replications are needed to reach inconsistency-free performance?
  - GOC experiments run at the Psychophysics Lab have ranged from 3 to 100 replications. As a rule-of-thumb, 16 replications will remove most of the inconsistency. Drga (1999) showed, however, that there was still some residual observer inconsistency even after 75 replications.
  - The actual number of replications required depends on the observer and the difficulty of the task, but more replications are always better!
- Is there a way to run fewer replications, but still get reliable estimates of inconsistency-free performance?
  - Yes! Vit Drga, Alan Taylor, and John Whitmore (Drga, 1999) developed FORCE analysis for just this purpose
  - FORCE (function of replications combined estimation) analysis enables estimation of inconsistency-free performance with fewer replications than GOC analysis alone
  - Good estimates may be made with around 10 replications, and reasonable estimates with as few as six replications
  - FORCE analysis estimates inconsistency-free measures of detectability, rather than a GOC curve, however, this still allows estimation of psychometric functions

An Overview and Example of FORCE Analysis

- FORCE analysis extrapolates GOC performance to an infinite number of replications, where, theoretically, all inconsistency is averaged out
  - To do this, the GOC performance is plotted as replications are added, for the number of replications that were run (green points, ±1SD error bars)
  - Because there is nothing inherently special about the order the replications were run, all possible combinations of one replication are calculated and averaged, then all combinations of two replications, etc.
  - A function is fitted to the data (blue)
  - The *asymptote* of the fitted function is an estimate of asymptotic inconsistency-free performance (red)

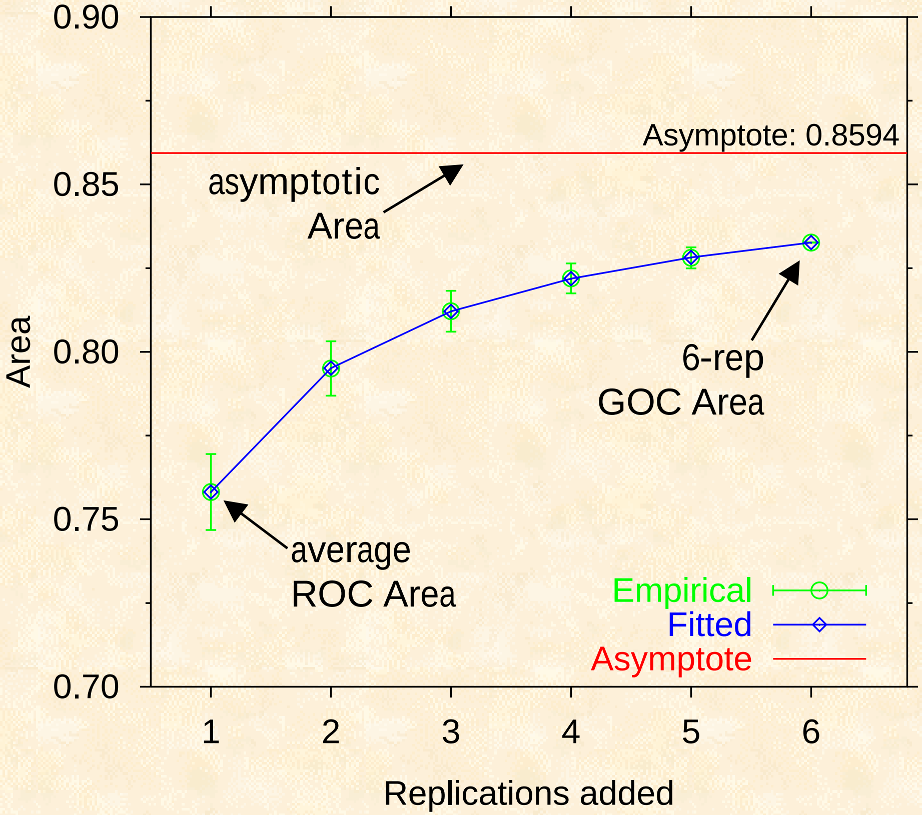

Fig 3a: An example of the results from a six replication GOC experiment

(noise-in-noise SIFC task, WT=1, W=40Hz, T=25ms, SNR = 8dB (from Lapsley Miller, 1999b))

Doing FORCE Analysis

- All-Combinations Analysis (ACA)
  - Plot the *average* measure of detectability (here we'll use the Area under the GOC curve), as a function of replications added, for *all combinations* of *r* replications, sampled from the total set of *n* replications
  - The number of replication combinations is given by the binomial coefficient ${}^{n}C_{r}$ e.g., for the 6 replications of the experiment in Fig 3:
    - the first point is the average Area of all six ROC curves (the GOC curve for one replication is just the ROC curve)
    - the second point is the average Area of all possible two-replication GOC curves, which can be selected from the six replications - fifteen GOC curves in total
    - the third point is the average Area from the twenty three-replication GOC curves
    - the fourth point is from the fifteen four-replication GOC curves
    - the fifth point is from the six five-replication GOC curves
    - the sixth point is from the one six-replication GOC curve
- This plot shows the average improvement in the Area as replications are added. It gives a "Function of Replications Added" or "FORA" curve

Doing FORCE Analysis, cont...

- Drga (1999) found that empirical FORA are invariably linear when plotted as the log of the *increments* in the Area, as a function of log replications-added
  - The linearity of virtually all examples he examined was on the order of 0.99 to 1.00, measured by Pearson's correlation coefficient
  - This linearity indicates that the series expansion of the Riemann-Zeta function may be used as the fitted FORA in *linear* coordinates
    - If the log-increment plot is not linear, then the estimate of the aysmptote is biased
- Asymptotic Detectability
  - Fit a Riemann-Zeta series to the empirical FORA (in linear coordinates) and calculate its asymptote
    - Drga (1999) shows how this may be done using a non-linear least squares gradient-descent method
  - Fig. 3b shows an example of the empirical FORA from Fig 3a and the fitted FORA, in log-increment coordinates

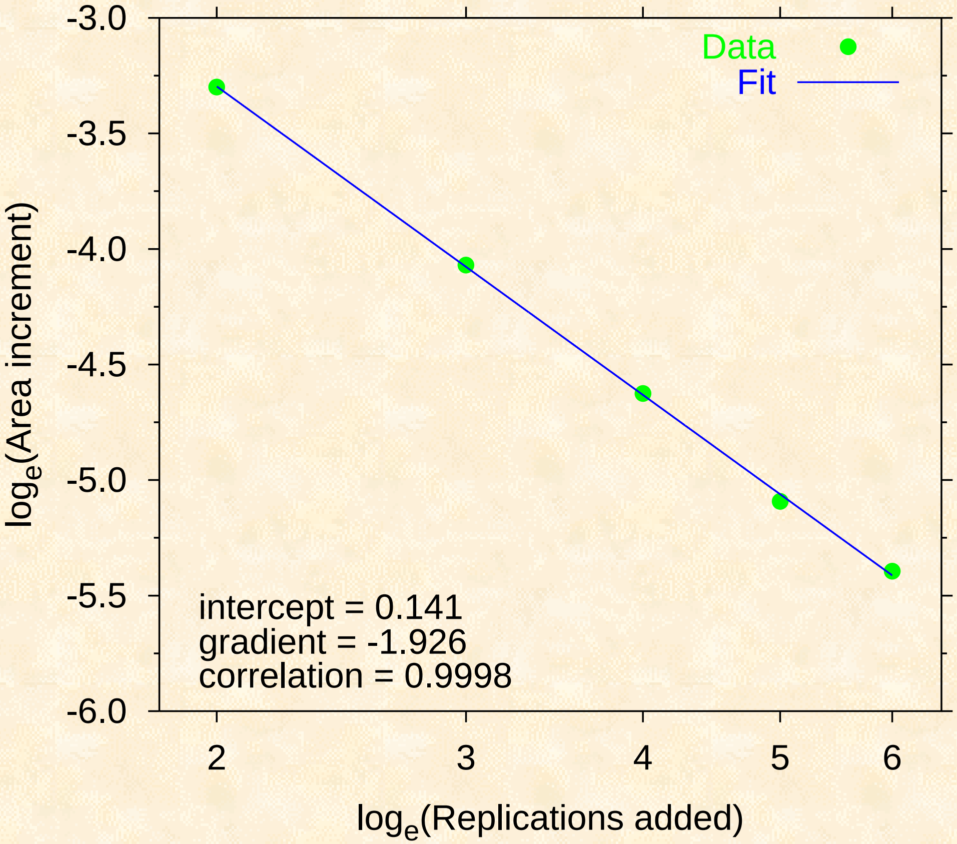

Fig 3b: An example of a log-increment plot from the 6 replication GOC experiment in Fig 3a

(noise-in-noise SIFC task, WT=1, W=40Hz, T=25ms, SNR = 8dB (from Lapsley Miller, 1999b))

Another example of FORCE analysis

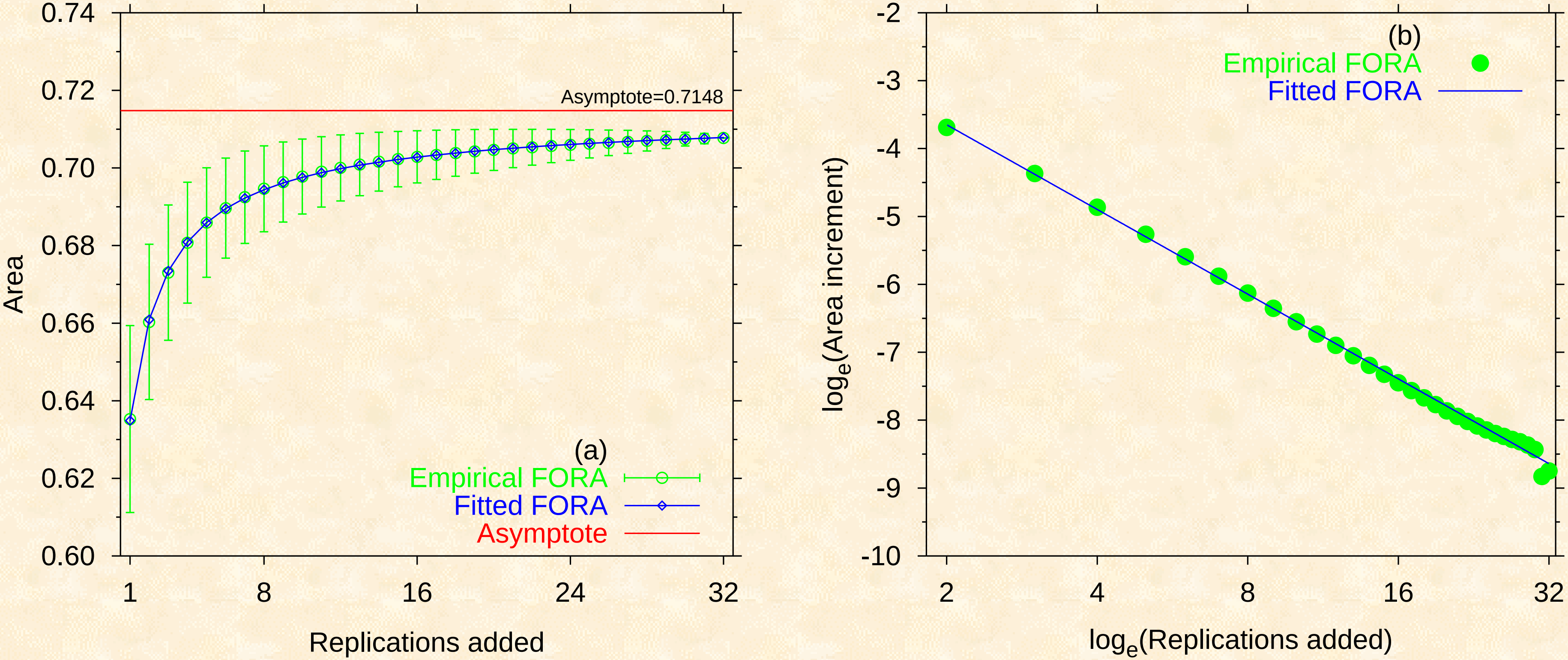

Fig 4. An example of FORCE analysis, showing the (a) empirical FORA, fitted FORA, and the asymptotic Area estimate, and (b) the empirical and fitted FORA in log-increment coordinates.

Based on data from Lapsley Miller et al. (1999): noise-in-noise SIFC experiment, WT=1, SNR = 7.5 dB.

An example of FORCE analysis, cont...

- Fig. 4 is based on data from a noise-in-noise SIFC experiment that was repeated 32 times by one observer.
  - Performance is still noticeably improving, even after 32 replications.
  - Even after 32 replications, the asymptotic Area is still above the GOC Area
    - The improvement in the Area from one replication to thirty-two replications is equivalent to 2.8 dB, assuming an energy or envelope detector model
    - The improvement from the 32-replication GOC to asymptotic performance is equivalent to 0.3 dB
    - The total improvement from single replication performance to asymptotic performance is 3.1 dB.
  - The equivalent improvement for Fig 3 is 2.7 dB from mean ROC to 6-replication GOC, 1.1 dB from GOC to asymptotic performance, for a total improvement of 3.8 dB

Improvement in the Psychometric Function

- The improvement in the psychometric function can be considerable, even from only 6 replications (examples from Lapsley Miller, 1999a,b)
- Fig 5a shows the data from Fig 4, plus data from four other signal levels
  - The theoretical psychometric function is the full-linear model, which for WT=1 is very similar to an energy detector

- Fig 5b shows a similar example where WT=2 (W=40 Hz, T=50ms)
  - The fitted model is the full-linear model, which for WT=2 performs *better* than the energy detector
  - **Once inconsistency has been removed, sometimes human observers can perform essentially as well as an ideal observer**

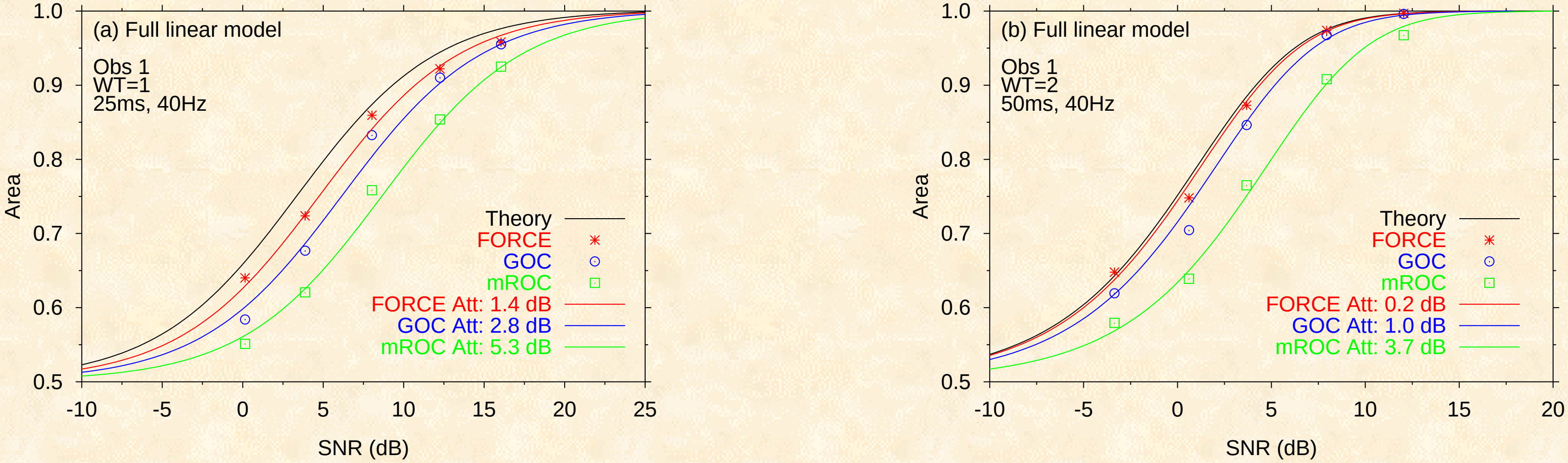

Fig 5: Plotted are the empirical psychometric function points from the mean ROC, GOC, and FORCE analysis, and the theoretical and linearly attenuated psychometric functions. (a) WT=1, (b) WT=2

Summary

- Group operating characteristic (GOC) analysis can remove error in psychophysical tasks that is due to observer inconsistency
- FORCE analysis extrapolates GOC performance to infinity, as a function of replications added, giving an estimate of asymptotic inconsistency-free performance
- These error-reducing techniques allow a clearer view of the processes underlying detectability, and can show, in some cases, that humans are able to perform as well as an ideal observer.

References

- Drga, V. (1999) *Theory of group operating characteristic analysis in discrimination tasks.* Unpublished doctoral dissertation, Victoria University of Wellington, New Zealand. Available on request (vi**t.drga@psychophysics.org**)
  - Contains an exhaustive list of pertinent GOC and FORCE references
- Lapsley Miller, J. A. (1999a) *The detectability of essentially band-limited and time-limited Gaussian noise by humans.* Manuscript submitted for publication.
- Lapsley Miller, J. A. (1999b) *The role of the bandwidth-duration product WT in the detectability of diotic signals.* Unpublished doctoral dissertation, Victoria University of Wellington, New Zealand. Available on request (j**udi@psychophysics.org** or on the web at http://www.**psychophysics.org)**
- Lapsley Miller, J. A., Scurfield, B. K., Drga, V., Galvin, S. J., & Whitmore, J. (1999) *Non-parametric relationships between single-interval and two-interval forced-choice tasks in the theory of signal detectability.* Manuscript submitted for publication.
- Taylor, A., Boven, R., & Whitmore, J. (1991) Reduction of unique noise in the psycho-physics of hearing by group operating characteristic analysis. *Psychological Bulletin*, *109*(1), 133-146.
- Watson, C. S. (1964) *Signal detection and certain physical characteristic of the stimulus during the observation interval.* (Doctoral dissertation, Indiana University, 1963). Dissertation Abstracts International, 24, 2995.

Software

- Software for running FORCE analysis is available on request from Linton Miller (**linton@psychophysics.org**)