To appear in A. Oliveira (Ed.) *Fechner Day 2004*. Coimbra, Portugal: International
Society for Psychophysics. (in press)

CONSTRAINED SCALING: ACHIEVING QUANTITATIVE CONVERGENCE ACROSS LABORATORIES

Michael Baumann*, Graeme Moffat+, Larry E. Roberts+ & Lawrence M. Ward*
*University of British Columbia, +McMaster University
michael@baumann.ca, moffatgd@mcmaster.ca
roberts@mcmaster.ca, lward@psych.ubc.ca

Abstract

It is well known that, although psychophysical scaling produces good qualitative agreement between experiments using ostensively the same methods but run in different laboratories, quantitative agreement is much more difficult to achieve. Constrained scaling, in which observers first learn a standard scale and then make magnitude judgments of other stimuli using the learned scale, has achieved excellent quantitative agreement between individual observers' psychophysical functions and could theoretically do the same for across-laboratory comparisons. We report two experiments that were replicated, using constrained scaling, in two different laboratories as examples of the level of agreement achievable with this technique. In general, we found across-laboratory agreement superior to that typically obtained with magnitude estimation.

Several authors have described the difficulties involved in obtaining quantitative convergence in psychophysical scaling. Such convergence is important if psychophysics is to become as useful as other sciences. Imagine if the charge of the electron had a range of values that appeared depending on who was doing the measurement experiment, or if the gas constant or the speed of light were not "constant" but depended on which investigator was writing about them. Unfortunately, this is the case for exponents of psychophysical power functions. Although S.S. Stevens (e.g., 1) argued that canonical exponent values should be adopted for all of the sensory continua, it has not been the case that the values he suggested could be achieved by every investigator, despite attempts to use the same methods and stimuli. A striking demonstration is the paper by Marks (2) in which across-observer-average exponents ranging from 0.37 to 0.80 from various laboratories are reported for ratio scaling of loudness of pure tones around 1000 Hz. Thus, if a theory predicted that the exponent for loudness of a 1000 Hz tone should be 0.60 based on physiological and physical considerations, that theory would be disconfirmed by the a sizable fraction of scaling experiments reported to date, although the average exponent is indeed around 0.60. Poulton (3) attempted to classify and model all of the various kinds of bias that affect such judgments and presumably give rise to the unacceptable level of variability of exponents (and other properties). Others, e.g., Laming (4), have suggested that such variation is the source of major interpretational problems with direct scaling results. Yet others, e.g., Lockhead (5), have suggested that the attempt to achieve canonical psychophysical scales is fundamentally misguided. Still, such scales do have vast usefulness, both in designing buildings, assessing environmental impact, and other applied contexts, and also in informing fundamental theories (e.g., 6). Therefore we have taken the approach that achieving canonical scales is a worthwhile goal.

The first step toward that goal was realized by West, Ward & Khosla (7), who showed that variation of scaling exponents across individuals could be substantially reduced by training observers to use a standard scale and then, while keeping them calibrated on that scale, having them judge other stimuli than those used in training (*constrained scaling*). West *et al.* (7) trained observers to the Stevens's sone scale for loudness of 1000 Hz pure tones, in which sones are a power function of sound pressure with an exponent of 0.6 (West et al. used $R = 16.6\ P^{\ 0.60}$). They then had observers judge tones of other frequencies on this same scale, reproducing the usual finding that exponents are larger for lower frequencies but doing so for every observer and with extremely little between-observer variability in the relation between exponents. West et al also had observers judge the brightness of lights on the same scale, in this case reproducing Stevens's standard finding of an exponent near 0.3 for brightness, ½ of the exponent for loudness. West *et al*. speculated that constrained scaling could also be used to achieve quantitative reproduction of results across laboratories. In this paper we report a first attempt at achieving this goal.

It is necessary to find a way to characterize the existing level of quantitative reproducibility of scaling results. After considering many alternatives, we decided to use the two indicators used by West *et al*. (7): the standard deviation of a set of exponents divided by the mean of that set (*SD/M*, the coefficient of variation), and the ratio of the highest to the lowest exponent in the set (*High/Low*). West *et al*. were able to reduce these indicators, calculated across individuals, from values ranging from 0.19 to 0.45 for *SD/M* and 1.6 to 6.0 for *High/Low* (from the literature summarized in their Table 1) to 0.045 to 0.100 for *SD/M* and 1.2 to 1.4 for *High/Low* (for experiments where observers were trained to the sone scale). To obtain such values for average exponents from groups of subjects, we selected 25 average exponents reported by various authors and summarized in Table 1 of Marks (2) for magnitude estimations of 1000 Hz tones, and 13 additional average exponents from magnitude estimation experiments reported in various papers by one of the authors (Ward, including two from West *et al*.; references supplied on request). For these 38 average exponents, the value of *SD/M* was 0.12/0.54 = 0.22 and that of *High/Low* was 0.80/0.37 = 2.2. We aimed to better these values.

In order to investigate reproducibility across laboratories, we collaborated in replicating two experiments in our two different (University of British Columbia and McMaster University) laboratories using the same constrained scaling protocol and graphical user interface, but different sound generation, presentation, and calibration apparatus. A standard SoundBlaster sound card, a custom artificial ear, and Kenwood KPM-510 headphones were used at UBC, and an Aardvark soundcard, a Bruel & Kjaer Artificial Ear, and Sennheiser HDA-200 headphones were used at McMaster. In the protocol observers first learned the sone scale using 1000 Hz tones (according to $R = 16.6\ P^{\ 0.60}$) and then produced judgments of pure tones of several different frequencies, including 500 Hz, and 5000 Hz in both experiments, and 65 Hz in addition in the second experiment. We also manipulated the number of training, judgment and calibration stimuli in an effort to discover the limits of the technique. In the first experiment [designated UBC52 (n=10) and McM52 (n=17)], observers made 52 training judgments of 1000 Hz tones with feedback, then 52 "calibration" judgments of the same stimuli with feedback, for a total of 104 judgments with feedback. They then made 52 judgments of 500 Hz tones without feedback interleaved with 52 judgments of 1000 Hz tones with feedback, followed by 52 judgments of 5000 Hz tones without feedback interleaved with 52 judgments of 1000 Hz tones with feedback. In each case, the levels of the 52 stimuli ranged from 40 dB to 90 dB in 1-dB steps and 1 trial with no stimulus was also included; stimuli were presented one time each in a shuffled order in each run. In the second experiment [UBC17 (n=15) and McM17 (n=15)], only 17 judgments were made in each set, instead of 52, and in addition 17 judgments of 65 Hz tones without feedback were made interleaved with 17 judgments of 1000 Hz tones with feedback. The 17 stimuli consisted of levels from 40 dB to 88 dB in 3-dB steps, and presented one time each in a shuffled order. Responses on the no-stimulus trial (52-trial studies), and any responses of "0" were not included in the curve fitting. Power functions were fitted to the individual judgments using linear regression on the logarithms of sound pressures and responses. Thus, exponents, m, in $R = aP^m$, where R is the response and P is the sound pressure of the stimulus, were estimated from $\log R = \hat{m} \log P + \log \hat{a} + error$. West et al. (7) discussed the problem of fitting power functions to such data, including how to estimate the effect of statistical regression on the exponent (the latter guided by comments of Stanley Rule). This effect can be measured by $m' = m\ r_{RP}$ where r_{RP} is the correlation coefficient between an observer's responses and the stimuli to which they were made. Thus, exponents estimated from a set of judgments are always smaller than the "true" exponent unless r_{RP} = 1, which is never the case. Here, we judged that the regression effect would be too large for accurate estimation of the exponent if r_{RP} was less than about 0.82 ($r_{RP}^2 < 0.67$). This never occurred for any of the 52-trial 1000-Hz runs but did occur for one or more 1000-Hz run for about half of the observers in the 17-trial experiments (overall 10/75 and 12/75 runs for UBC17 and McM17 respectively), and at least one of 500 Hz, 5000 Hz, or 65 Hz judgment sets in 1 of 17 observers in McM52, 2 of 10 in UBC52, 5 of 15 in UBC17 and 7 of 15 in McM17. Table 1 is based on all 1000-Hz runs, regardless of r_{RP}^2, but we do not include in Table 2 the data of these latter observers.

Figure 1 displays representative psychophysical functions from the "best" and "worst" observers (in terms of r_{RP}^2) from the four experiments for the 1000 Hz runs (selected from those made post-training). The functions from UBC52 and McM52 are very comparable to those reported by West *et al.* (7). The worst observers in the UBC17 and McM17 experiments, however, as mentioned earlier have r_{RP}^2 values lower than our nominal criterion of 0.67. Clearly for these observers 17 trials of training is not enough. It should be stressed that *individual responses* are plotted in these functions, in contrast to usual psychophysical functions that, even when plotted for individual observers, consists of points based on several to many judgments per stimulus. Table 1 summarizes the data from the 1000 Hz runs from the four experiments. All of these data are very comparable to those of West *et al.* (7), both the 1000 Hz runs with feedback and also their 1000 Hz runs without feedback, which had identical exponents and comparable *SD/M* and *High/Low* statistics across individuals. Most important, it is easy to see that the exponent values themselves are quantitatively very similar across laboratories, as are the *SD/M* values. There is, as might be expected, more variability across individuals in the 17-

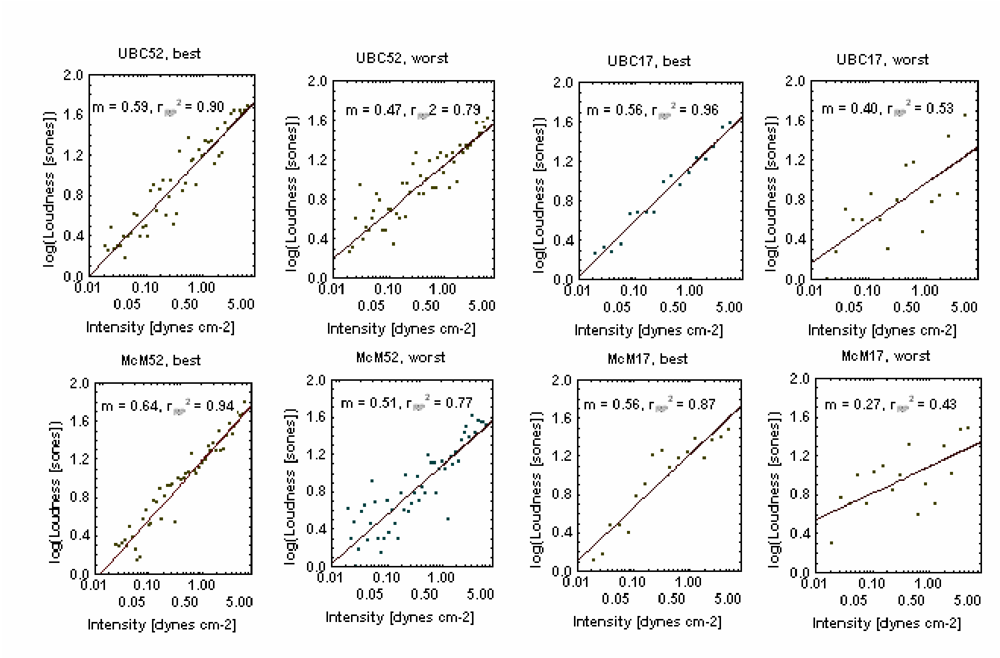

Figure 1. Representative psychophysical functions of "best" and "worst" observers for judgments of loudness of 1000-Hz pure tones from the four experiments in the two different laboratories.

Table 1. Average across observers of exponents (*Mean m*) and *SD/M* values for judgments of the loudness of 1000 Hz pure tones during several runs with feedback. WWK=West *et al.*(7)

Study	Training	Recal 1	Recal 2	Recal 3	Recal 4
UBC52					
Mean m	0.55	0.53	0.52	0.49	NA
SD/M	0.06	0.08	0.06	0.12	NA
McM52					
Mean m	0.58	0.56	0.55	0.51	NA
SD/M	0.14	0.08	0.12	0.10	NA
UBC17					
Mean m	0.57	0.55	0.52	0.51	0.47
SD/M	0.21	0.18	0.21	0.17	0.21
McM17					
Mean m	0.56	0.54	0.48	0.48	0.47
SD/M	0.21	0.19	0.22	0.16	0.19
WWK					
Mean m	0.59	0.54	0.54	0.55	0.56
SD/M	0.03	0.04	0.07	0.07	0.07

stimulus experiments: the *SD/M* values are about the same as the best of the standard magnitude estimation experiments surveyed by West *et al.* Clearly there is a cost of estimating power function exponents from psychophysical functions with so few trials.

West et al. had observers perform 1000 judgments in total in their first experiment, of which 800 were at 1000 Hz (600 with feedback), and 200 at 65 Hz. In contrast, in UBC17 and MCM17, observers performed only a total of 85 judgments of 1000 Hz tones with feedback. Thus, their total experience was not even as great as the first training run in West *et al.*'s Experiment 1. We can summarize by comparing the *SD/M* and *High/Low* statistics for these replications using constrained scaling with those for conventional magnitude estimation mentioned earlier. Over the 22 average constrained scaling exponents displayed in Table 1, *SD/M* = 0.035/0.53 = 0.07, and *High/Low* = 0.59/0.47 = 1.6, clearly an improvement over the conventional technique (0.22 and 2.2 respectively).

A more stringent test, perhaps, is the reproducibility of the judgments of off-training stimuli across laboratories. Table 2 presents the data for judgments of 500 Hz, 5000 Hz and 65 Hz pure tones. Here we see a similar but slightly more complicated picture. First, we find excellent agreement across laboratories in the exponents for 500 Hz and 5000 Hz, although the *SD/M* across observers is somewhat larger than for the 1000 Hz tones. Over the 4 replications of each set of judgments, the *SD/M* and *High/Low* statistics are respectively 0.017/0.52 = 0.03 and 0.54/0.50 = 1.1 for 500 Hz, and 0.051/0.52 = 0.10 and 0.57/0.46 = 1.2 for 5000 Hz. Although probably somewhat underestimated because of the small set of values available, these numbers represent excellent reproducibility across laboratories, comparable to that achieved across individuals by West *et al.* (7). The results for the 65 Hz judgments, however, are less encouraging. Although not reliably different from each other (*t* = 0.66), the average exponent values displayed in Table 2 are significantly larger than those reported by West et al. in their two main experiments, viz. 0.70 and 0.67. Nonetheless, all of the experiments replicate the usual finding that exponents for 65 Hz are significantly larger than those for 1000 Hz. The reason for the quantitative disagreement is not clear; it may arise from the fact that observers in UBC17 and McM17 had trouble judging these stimuli at all. The *SD/M* values of 0.25 and 0.30 in those experiments are the highest ever seen using constrained scaling, and this is after excluding nearly half of the observers from the

Table 2. Average across observers of exponents (*Mean m*) and *SD/M* values for judgments of the loudness of 500 Hz, 5000 Hz, and 65 Hz pure tones without feedback.

Study	**500 Hz**	**5000 Hz**	**65 Hz**
UBC52			
Mean m	0.51	0.56	NA
SD/M	0.12	0.12	NA
McM52			
Mean m	0.54	0.57	NA
SD/M	0.12	0.20	NA
UBC17			
Mean m	0.52	0.51	1.02
SD/M	0.13	0.27	0.25
McM17			
Mean m	0.50	0.46	0.93
SD/M	0.16	0.16	0.30

average because at least one r_{RP}^2 value was less than 0.67. Perhaps the small number of judgment trials, and the concomitant lack of practice judging these hard to hear stimuli (thresholds typically 40 dB or higher; several observers couldn't hear many of the lower-level stimuli), reduced the efficacy of the technique. Indeed two of the UBC17 observers had non-monotonic psychophysical functions, indicating that they were simply guessing the appropriate response. These observers may have had an undiagnosed low-frequency hearing loss. The fact that the calibration judgments of 1000 Hz stimuli interleaved with the 65 Hz judgments for these observers remained normal indicates that the problem was only with the 65 Hz stimuli. As mentioned earlier, West et al.'s observers performed 200 judgments of the 65 Hz stimuli; perhaps more practice with these stimuli in our experiments would have led to a closer replication. Nevertheless, this failure to reproduce the 65 Hz exponents across laboratories needs to be investigated further, and indicates that the minimal implementation of constrained scaling might not be good enough for scientific purposes in some cases.

Overall these constrained scaling experiments and those of West et al. (7), despite differences in lab equipment, observers, and numbers of training and test trials, demonstrate across-laboratory reproduction of quantitative results substantially superior to that obtainable with conventional techniques. Other modifications of conventional direct scaling techniques also could yield superior reproducibility, e.g., the CR-100 scale of Borg & Borg (8) and the master scaling technique of Berglund (9). One or more of these techniques should be adopted by convention in order to create reproducible canonical scales of sensory and other stimuli. Moreover, the UBC17 and MCM17 experiments demonstrate that accuracy comparable with the best of conventional techniques, often requiring hundreds of judgments per condition, can be obtained in constrained scaling with only 17 training judgments and 17 + 17 judgments (17 calibration and 17 test) per test condition. This means that this technique will be useful in the clinic, where we are already using it to measure tinnitus magnitude.

References

1. Stevens, S.S. (1975). *Psychophysics*. New York, Wiley.
2. Marks, L.E. (1974). On scales of sensation: Prolegomena to any future psychophysics that will be able to come forth as science. *Perception & Psychophysics*, *16*, 358-376.
3. Poulton, E.C. (1989). Bias in Quantifying Judgments. Hillsdale, NJ: Erlbaum.
4. Laming, D. (1997). *The Measurement of Sensation*. Oxford: Oxford University Press.
5. Lockhead, G.R. (1992). Psychophysical scaling: Judgments of attributes or objects? *Behavioral and Brain Sciences, 15*, 543-601.
6. Norwich, K.H. (1993). *Information, Sensation and Perception.* New York, Academic.
7. West, R.L., Ward, L.M. & Khosla, R. (2000). Constrained scaling: The effect of learned psychophysical scales on idiosyncratic response bias. *Perception & Psychophysics*, *62*, 137-151.
8. Borg, G. & Borg, E. (2001). A new generation of scaling methods: Level anchored ratio scaling. *Psychologica*, *28*, 15-45.
9. Berglund, M. B. (1991). Quality assurance in environmental psychophysics. In S.J. Bolanowski & G.A. Gescheider (Eds.), *Ratio Scaling of Psychological Magnitudes* (pp. 140-162). Hillsdale, NJ: Erlbaum.