

Stimulus-Specific Learning: Disrupting the Bow Effect in Absolute Identification

Pennie Dodds¹, Christopher Donkin², Scott D. Brown¹, Andrew
Heathcote¹, A. A. J. Marley³

¹School of Psychology, University of Newcastle, Australia

²Department of Psychology & Brain Sciences, Indiana University, Indiana

³Department of Psychology, University of Victoria, Canada

Counts

Abstract: 135

Body: 6049

Captions: 249

Figures: 5

Address correspondence to:

Pennie Dodds
School of Psychology
University of Newcastle
Callaghan NSW 2308
Australia
Ph: (+61)249216959
Email: Pennie.Dodds@newcastle.edu.au

Abstract

The “bow effect” is ubiquitous in standard absolute identification experiments - stimuli at the centre of the stimulus-set range elicit slower and less accurate responses than others. This effect has motivated various theoretical accounts of performance, often involving the idea that end-of-range stimuli have privileged roles. Two other phenomena (practice effects, and improved performance for frequently-presented stimuli) have an important but less explored consequence for the bow effect: standard within-subjects manipulations of set size could disrupt the bow effect. We found this disruption for stimulus types that support practice effects (line length and tone frequency), suggesting that the bow effect is more fragile than thought. Our results also have implications for theoretical accounts of absolute identification, which currently do not include mechanisms for practice effects, and provide results consistent with the literature on stimulus-specific learning.

The absolute identification paradigm explores a fundamental limit – that the number of separate categories that can reliably be identified along a single physical dimension is very small (about 7 ± 2 , according to Miller, 1956). In a typical absolute identification experiment, a participant is presented with a set of stimuli that vary along only one dimension (e.g., lines varying in length, or tones varying in intensity). These stimuli are labelled with the numerals #1 to # N in order of increasing magnitude. The participant is then shown one stimulus at a time in a random order and asked to respond with its label. Despite the task's apparent simplicity, absolute identification data reliably exhibit a great many phenomena, some of which are quite complex (for reviews see Petrov & Anderson, 2005, and Stewart, Brown & Chater, 2005).

In this paper we focus on one of the most fundamental of these phenomena: that performance is better for stimuli at the outer edges of the stimulus range, and worse for those in the centre. This phenomenon is called the *bow effect* because a U-shaped curve is observed when accuracy is plotted against stimulus magnitude and an inverted U-shaped curve when plotting response time (RT). These bow effects are robust phenomena that are consistent across manipulations of stimulus magnitude (Lacouture, 1997), the number of stimuli ("set size" - Stewart et al., 2005), and sensory modalities (Dodds, Donkin, Brown & Heathcote, 2011).

However, recent evidence that practice improves absolute identification performance (Dodds et al., 2011; Rouder, Morey, Cowan & Pfaltz, 2004) implies that the bow effect could be disrupted by practice when the effects of practice are stimulus-specific. Previously, it was widely believed that even extended practice did not lead to much improvement in absolute identification (see, e.g., Miller, 1956; Shiffrin &

Nosofsky, 1994) but recent research has shown that improvements can be made for some types of stimuli. For example, Dodds et al. demonstrated that practice improved performance a great deal when the stimuli were lines varying in length or tones varying in frequency, but very little when the stimuli were tones varying in loudness. Using tones varying in frequency, Cuddy (1968, 1970) found that presenting some stimuli more often than others resulted in an overall improvement – for all stimuli. This effect was limited however, to trained musicians, and to a task more akin to categorization than standard absolute identification. Using a more standard paradigm, and untrained participants, Cuddy, Pinn and Simons (1973) demonstrated improved performance across the entire stimulus range when one stimulus was presented more often than the others (but see Chase, Bugnacki, Braida & Durlach, 1982, for conflicting results).

We generalize these earlier findings in several ways. We examine more than one kind of stimulus dimension (not just tones varying in frequency), and we also use a more standard paradigm in which – during each block of trials – all stimuli were presented equally often. This latter constraint is important because presenting some stimuli more frequently than others encourages participants to bias their responses. Instead of using unequal presentation frequency within blocks, we employ a different experimental manipulation that encourages stimulus-specific learning; changing the stimulus set size on a within-subject basis between blocks of trials. In our design, a participant would first be asked to identify two stimuli (“set size two”, denoted “ $N=2$ ”), and, in a later phase, be asked to identify these two stimuli along with six others, in an $N=8$ condition. The stimulus set for the smaller set size is created from the middle stimuli of the larger set size (e.g., the two stimuli for $N=2$ are the same as the middle two stimuli from $N=8$).

In many other paradigms practice has stimulus-specific effects, that is, extensive practice with some stimuli does not confer a benefit upon other similar stimuli. For example, there is an extensive literature on perceptual learning that has almost uniformly shown poor generalization (for a review see Petrov, Doshier & Lu, 2005). This background makes Cuddy's (1968, 1970) results (general improvement in absolute identification after practice with just one stimulus) quite surprising. However, this contrast is complicated because of Cuddy's non-standard identification paradigm. If we find, using a standard absolute identification paradigm, that practice improves performance in a stimulus-specific (rather than task-wide) manner, and if the improvements persist across changes in set size, prior exposure to the $N=2$ condition might improve performance on the middle two stimuli for the $N=8$ condition. This performance boost would disrupt the bow effect for the larger set size.

Re-examination of existing AI data lends some preliminary support to this hypothesis. As a baseline comparison, first consider Stewart et al.'s (2005) Experiment 1, in which set size was manipulated between-subjects - some participants performed an absolute identification task with six tones of varying frequency, others with eight tones, and still others with ten tones (i.e., $N=6, 8$ or 10). The data from this experiment (Figure 1a) exhibit the standard bow effect in each set size, with poorest performance for the middle stimuli.

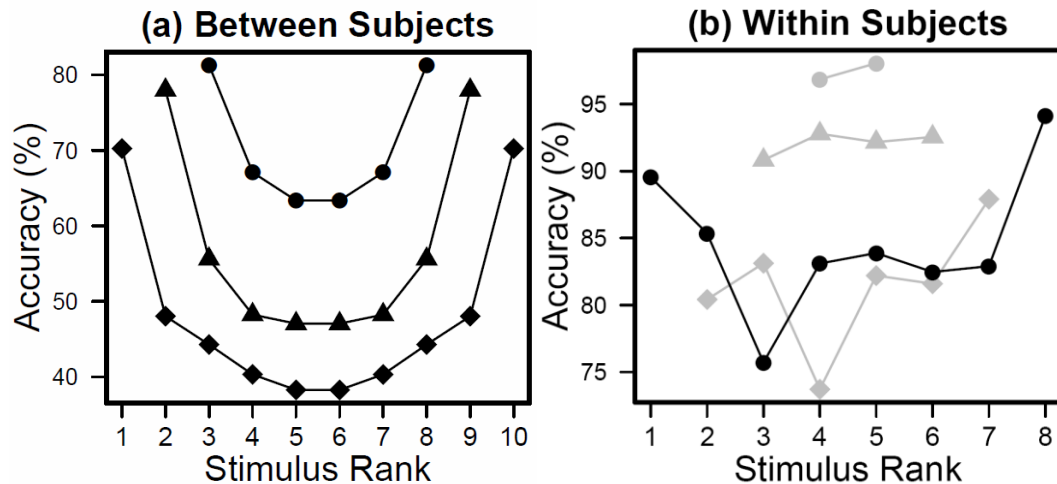


Figure 1. (a) Between-subjects data from Stewart et al.'s (2005) Experiment 1 and (b) within-subjects data from Kent and Lamberts' (2005) Experiment 2. Each line on each graph represents a different set size. All set sizes in Figure 1 (b) are in grey except set size $n=8$, for ease of comparison. Both plots show accuracy, measured by the percentage of correct responses, averaged over participants, for different set sizes. Stewart et al. "symmetrized" their data by averaging responses to small and large stimuli in corresponding pairs; we additionally averaged their data over stimulus spacing (wide vs. narrow).

In contrast, Kent and Lamberts (2005) had three participants perform absolute identification with dots varying in separation, using four different set sizes ($N=2, 4, 6$, and 8), manipulated within-subjects. Each participant experienced all set sizes, one after the other¹, and each participant showed a disruption (flattening) of the bow effect. Figure 1b illustrates a disrupted bow effect, with much shallower bows than in Stewart et al.'s data. Most pertinent for our study is the result for set size $N=8$ (bold in Figure 1b). Kent and Lamberts' data from this condition display a much shallower bow effect than before. Identification of the middle stimuli is enhanced to the point where response accuracy for

¹ Set size was partially counterbalanced between subjects. Participant 1 saw $N=10, 4, 2, 8$ then 6 , participant 2 saw $N=10, 6, 8, 2$ then 4 and participant 3 saw $N=4, 8, 2$ then 6 . We did not include set size $N=10$ in the figure because not all participants experienced this condition.

the central stimuli (#4 and #5) is just as good as response accuracy for the next-to-extreme stimuli (#2 and #7). In comparison, the standard effect (as in Figure 1a) exhibits a deep bow, so that there is a large difference between these pairs of stimuli. Our statistical analyses are motivated by this pattern, and test for a standard bow effect by assessing performance differences between the central stimuli and the next-to-edge stimuli².

Although the data in Figure 1 are suggestive, they must be interpreted with caution. There were many differences between Stewart et al.'s (2005) and Kent and Lamberts' (2005) experiments beside the between- vs. within-subject manipulation of set size: for example, different stimulus modalities, different amounts of training per participant, and different set sizes. Further, the W-shape in Figure 1b is quite clear in Kent and Lamberts' data when averaged over their three participants, but further examination reveals large differences between participants; differences that do not uniformly support the hypothesis that pre-training on smaller set sizes will disrupt the bow effect for larger set sizes. Our Experiments 1 and 2 were an attempt to clarify the evidence that stimulus-specific practice can disrupt the bow effect, and to test our hypothesised explanation of this disruption that it is due to stimulus-specific practice effects caused by unequal stimulus-presentation frequencies.

² Comparing the central stimuli (#4/#5) against their neighbours (#3/#6) provides little power to detect a standard bow effect, because the bow curvature is smallest in the centre (as in Figure 1a). On the other hand, comparing the central stimuli (#4/#5) against the edge stimuli (#1/#8) will classify all but the most severe disruptions of the bow effect as “standard bows” - e.g. the disrupted bow effect in Figure 1b still has better performance on edge stimuli than central stimuli.

Experiment 1

Dodds et al. (2011) found that practice can improve identification of line length but not of tone loudness. Hence, our hypothesis makes a clear prediction: if stimulus-specific practice disrupts the bow effect, a within-subjects manipulation of set size should disrupt the bow effect when the stimuli are lines varying in length but not when they are tones varying in loudness.

Method

Twenty-three participants were randomly allocated to either an absolute identification task using tone loudness (12 participants) or line length (11 participants). The stimuli for the line length task were eight pairs of small white squares, varying in horizontal separation. Each square had sides of length 3.3mm, and was shown at high contrast. The stimuli are referred to as lines varying in length because the participant is essentially making a judgment of length. The eight horizontal separations were 23.5, 26.0, 29.1, 32.2, 35.6, 39.3, 43.4, and 47.4 mm. The viewing distance was not physically constrained, but was approximately 700mm (so the stimuli subtended visual angles ranging from 3.3° to 6.7°). For the tone loudness condition, the stimuli were eight 1000Hz pure sine tones with loudness varying from 79db to 100db, in increments of 3db. Tones were generated using Matlab 2009a, with stepped onsets and offsets (although, by definition, the sine waves started and finished at zero amplitude because their duration was an integer-multiple of their frequency).

Before beginning the experiment, participants were presented with each of the eight stimuli, one at a time, along with the corresponding label. On every trial, participants were first shown a fixation cross for 300ms, which was removed when the

stimulus was presented. In the line-length condition, the stimulus remained on screen for 1 second, after which a mask appeared. The mask consisted of approximately 50 white squares of the same size used for the stimuli, randomly scattered across the screen. In the tone loudness condition, the tone played for 1 second followed by silence. In both conditions, participants were able to respond at any point after the stimulus presentation onset. Responses were made by pressing the appropriate numeral key (from 1-8) on the top line of the keyboard. Participants were given one opportunity to respond, after which feedback was provided.

Each participant took part in two one-hour sessions, on separate days, for a total of 20 blocks of 80 trials each. The first five blocks in the first session used only the middle two stimuli ($N=2$) and all subsequent blocks used all stimuli ($N=8$). When the participants were presented with only the middle two stimuli in the first 5 blocks in the first session, they responded to these with the numerals 4 and 5. In total, every participant received 200 presentations each of stimuli #4 and #5 when $N=2$ and 150 presentations of each of the eight stimuli when $N=8$. This meant that, across the whole experiment, each participant received 350 presentations each of stimuli #4 and #5 and 150 presentations of each of the other stimuli.

Results

Responses in the $N=2$ condition were quite accurate and rapid in both the length and loudness conditions: mean accuracy was 78% for length and 86% for loudness, and mean RT was 1.03 sec for length and 0.84 sec for loudness. Figure 2 shows mean response accuracy and the mean RT for correct responses, both conditional on stimulus rank, for the $N=8$ condition, separately for line length and tone loudness. Across all

stimuli, average accuracy was very similar for loudness and line length (46.8% and 47.1%, respectively), but the pattern of performance was quite different. A typical, deep, bow effect was observed for tone loudness: the mean accuracy was significantly higher for stimuli #2 and #7 ($M_{\#2/7}=46\%$) than for stimuli #4 and #5 ($M_{\#4/5}=33\%$) and the mean RT was significantly faster ($M_{\#2/7}=1.25$ sec., $M_{\#4/5}=1.37$ sec.). These differences were statistically reliable according to linear contrasts comparing the two group means (for accuracy, $F(1,77)=31.4$, $p<.001$ and for RT, $F(1,77)=10.9$, $p=.001$).

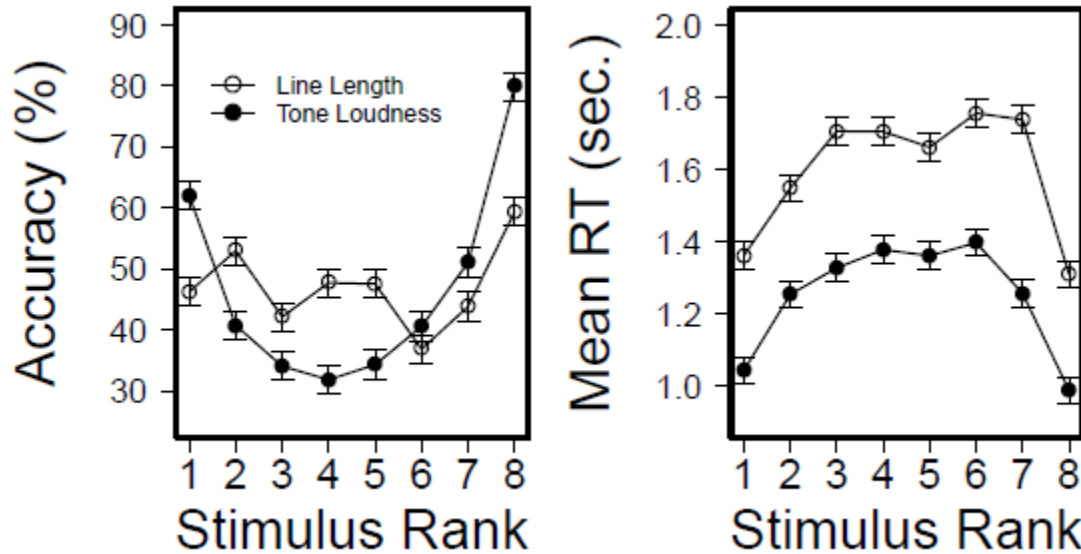


Figure 2. Accuracy and mean RT as functions of ordinal stimulus magnitude for Experiment 1. The lines represent different stimulus sets – line length or tone loudness. Error bars are 95% confidence intervals calculated in the repeated-measures manner described by Loftus and Masson (1994), separately for the two between-subjects conditions.

For the line length condition, however, the bow effect was clearly disrupted, with stimuli in the middle of the range (#4 and #5) eliciting faster and more accurate responses than for some other stimuli. In particular, linear contrasts showed that neither the mean accuracy nor the mean RT for stimuli #2 and #7 ($M_{acc}=48\%$; $M_{RT}=1.64$ sec.) was

significantly different than the mean accuracy or mean RT for stimuli #4 and #5

($M_{acc}=48\%$, $M_{RT}=1.68$ sec.; both $F_s < 1$).

The above analyses show a standard bow effect for tone loudness but not for line length. Nevertheless, these results do not directly support the conclusion that the bow effect was shallower in the line length condition than the tone loudness condition (since “the difference between 'significant' and 'not significant' is not itself statistically significant”; Gelman & Stern, 2006). To directly test this hypothesis, we calculated an interaction contrast comparing the depth of the bow effect between the two conditions, by taking the difference of the two contrasts reported above and testing it against the appropriate error variance term from the mixed ANOVA. These contrasts confirmed that there was a deeper bow effect for tone loudness than for dot separation response accuracy data ($F(1,147)=10.7$, $p < .001$). For RT data, the comparison was not significant ($F(1,147)=1.63$, $p = .10$).

Discussion

Participants in Experiment 1 first practiced the identification of two central stimuli (in an $N=2$ condition) and then the identification from the full set of eight stimuli. Data from those participants who identified tone loudness were quite standard, with the poorest performance for middle stimuli, and deep bow effects. However, for those participants who identified line lengths, performance for the central stimuli improved, to the point where it was not significantly poorer than performance on the next-to-edge stimuli. A potential weakness of this result is that null findings may be due to limited statistical power. To foreshadow, Experiment 2 addresses this concern, and obtains

similar differences to Experiment 1, using a different design with the same line length stimulus set.

The results of Experiment 1 suggest that the effects of pre-exposure to the central stimuli can, for certain stimulus dimensions, persist for long time intervals on the order of hours, rather than minutes as other authors have observed for the effects of stimulus presentation frequency (e.g., Petrov & Anderson, 2005). The performance bonus that we found in the line length condition persisted into the second experimental session, which was, on average, a full day after the extra presentations of the two central stimuli (in the $N=2$) condition. This was true even when we limited analyses to data from session two, during which the $N=2$ condition was not experienced.

Performance curves similar to those obtained here have also been observed by Kent and Lamberts (2005; Experiment 2) and Lacouture, Li and Marley (1998; Experiment 1), each of whom observed relatively flattened bow effects even when stimuli were presented equally often in each condition. The main difference between those experiments and those yielding a typical bow effect (e.g., Stewart et al., 2005) appears to be the within-subjects manipulation of set size. This manipulation results in more frequent presentations of centre stimuli than others, across the entire experiment, which could explain the corresponding performance benefit. Experiment 2 tests a potential confound to this presentation-frequency hypothesis present in Experiment 1.

Experiment 2

In Experiment 1, the two central stimuli were both presented *before* the others and presented *more often* than the others. Either, or both, of these factors could be the cause of the disrupted bow effect observed for line length stimuli in Experiment 1. In

Experiment 2, using only line lengths, we balanced the number of presentations per stimulus over the entire experiment. Thus, if the bow effect is disrupted in Experiment 2, then the results of Experiment 1 may be explained by some stimuli being presented before others. Alternatively, if the typical bow effect re-appears in Experiment 2, then the results may be explained by the differences in presentation frequency.

Method

We used the same procedure and stimuli as in the line length condition of Experiment 1, with 21 new undergraduate participants from the University of Newcastle. The experiment was divided into three sections. Participants, however, were only told of the first two sections. In section 1, participants completed two blocks of 100 trials each with just the central two stimuli ($N=2$). In section 2, participants completed 1000 trials in ten blocks, with all stimuli ($N=8$). However, the central two stimuli in this section were presented only 50 times each, while the other stimuli appeared 150 times each. Thus, at the end of the first two sections, each stimulus had appeared exactly 150 times. Participants were not explicitly told that the presentation of certain stimuli would be reduced in the second section. The third section reverted to five blocks of 80 trials each with all eight stimuli appearing equally often. Altogether, the three sections took participants approximately two hours, which they completed in a single testing session. One-minute breaks were provided regularly throughout the experiment. Participants were also given a single, extended five-minute break at the halfway point.

Results

The data from two participants were removed from analysis due to low accuracy (< 25% correct across the entire experiment, which was much lower than other participants in the experiment, $M=48\%$). Mean accuracy and mean RT for the $N=2$ condition were 80% and .99 sec., respectively. Figure 3 shows mean accuracy and mean RT for the $N=8$ condition. The final section of the experiment, during which each stimulus was presented equally often, is shown in black, and the unequal-frequency (middle) section is shown in grey. Accuracy was poorer, and mean RT longer, for the central two stimuli than all others in the critical third section of the experiment (when all stimuli were presented equally often).

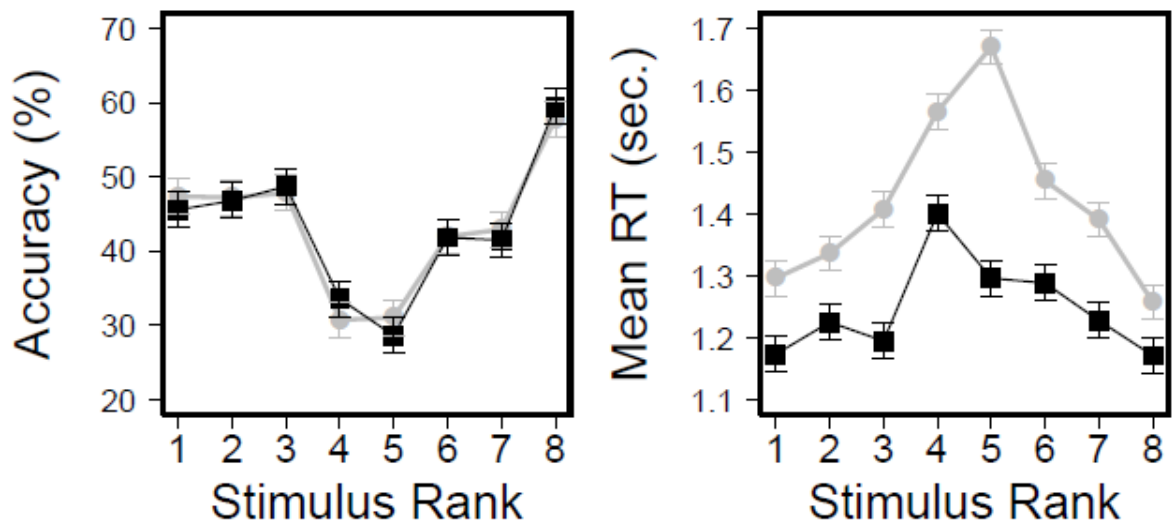


Figure 3. Accuracy and mean RT as functions of ordinal stimulus magnitude for Experiment 2. Each line represents a different section of the experiment. Grey lines represent the Section 2 (unequal presentation), black lines represent Section 3 (equal presentation). Error bars are as in Figure 2.

Repeated measures ANOVAs on accuracy and mean RT from the final phase of the experiment confirmed main effects of stimulus magnitude (accuracy: $F(7,126)=9.37$, $p<.001$; RT: $F(7,126)=4.56$, $p<.001$). As would be expected in a standard bow effect,

linear contrasts showed that the mean accuracy for stimuli #2 and #7 ($M=44\%$) was significantly greater than for the mean accuracy of stimuli #4 and #5 ($M=31\%$; $F(1,126)=17.9, p<.001$), and mean RT was significantly faster ($M_{\#2/7}=1.23$ sec., $M_{\#4/5}=1.35$ sec.; $F(1,126)=11.2, p=.001$). Furthermore, this effect was even evident when comparing middle stimuli with their immediate neighbours (stimuli #3 & #6; $M_{acc}=45\%$, $F(1,126)=21.2, p<.001$; $M_{rt}=1.24$, $F(1,126)=8.59, p=.004$)

Figure 3 shows very little difference in accuracy between section 2 and section 3 in Experiment 2 ($F<1$). Our hypothesis about over-presentation of certain stimuli would suggest the occurrence of a W-shape in accuracy in early trials of section 2, due to the over-presentation of the two central stimuli in section 1; however, we do not see such a pattern, likely due to a lack of power. There was, however, a clear reduction in RT in section 3 relative to section 2 (paired samples t-test: $t(7) = 5.54, p<.001$), which is likely due to a general improvement by practice, with participants trading the possibility of improved accuracy for improvements in speed.

Discussion

When the number of presentations per stimulus was manipulated so that participants were eventually exposed to an equal number of presentations of all stimuli, performance on the central stimuli was poorer than on all other stimuli, as in a standard bow effect, for both accuracy and for RT. This suggests that it is the over-presentation of certain stimuli, not presentation order, which leads to improvement in performance for those stimuli.

Experiment 2 also provided an estimate of the difference that might be expected between the central pair of stimuli (#4 and #5) and the next-to-edge stimuli (#2 and #7),

when a standard bow effect is observed. In particular, the central stimuli were correctly identified 13% less often than the next-to-edge pair. A power analysis shows that, if such a difference had been present in the line length condition of Experiment 1, the corresponding linear contrast would have detected a significant difference with almost perfect power ($>99\%$). Indeed, even if the true difference was only half as large (6.5%) the power would still have been close to perfect ($>99\%$). This suggests that the combined null results for both accuracy and RT from the line length condition of Experiment 1 were very unlikely to have been caused by a lack of statistical power.

Experiment 3

Experiment 1 confirmed a prediction arising from Dodds et al.'s (2011) investigation of learning effects in absolute identification: if the bow effect is disrupted in within-subjects designs because of differential stimulus-specific practice, this effect should be modulated by the susceptibility of the stimulus type to learning. As predicted, Experiment 1 showed a standard bow effect when tone loudness (which does not show strong learning) was judged, but not when line length (which does show a strong learning effect) was judged. A necessary weakness of such an experiment is the comparison of performance on different stimulus types, as these will often have different pairwise discriminability and other characteristics. Experiment 3 remedies this weakness by comparing results for two conditions that both use the same stimulus type; tones varying in frequency. The conditions differ only in the order in which different set sizes are practiced, testing the prediction that one order enhances the advantage for the central pair and other reduces it (i.e., it disrupts the bow effect).

This design enables a confirmation of the findings of Experiment 2 and supports a direct comparison of the bow effects from the two conditions, avoiding the problem of confirming a null hypothesis. It also tests a subtler version of a prediction from Dodds et al.'s (2011) work. Dodds et al. showed that learning effects for tones varying in frequency were smaller than those for line length, but larger than those for tones varying in loudness. Thus, if the bow effect is disrupted by practice, this disruption should also appear for tones varying in frequency, but the disruption should be less marked than for line lengths.

Method

We used the same procedure as Experiment 1, with 25 participants randomly assigned to one of two conditions that differ only in presentation order. Participants were given stimulus sets of either set size $N=2$ then $N=8$, or the reverse, which we will refer to as the 2-then-8 and 8-then-2 conditions (with 13 and 12 participants, respectively). Each participant took part in 20 blocks of practice, over two hours. Regular one-minute breaks were provided between blocks with a compulsory five-minute break after approximately one hour. Each block consisted of 80 trials. The $N=2$ condition was practiced for five blocks, and the $N=8$ condition for 15 blocks. In the 2-then-8 condition, the $N=2$ condition was practiced first, followed by 15 blocks of the $N=8$ condition. This was reversed in the 8-then-2 condition. The stimuli were eight one-second, 67db tones with frequencies taken from Stewart et al. (2005; wide spaced condition): 672, 752.64, 842.96, 944.11, 1057.11, 1184.29, 1326.41, and 1485.58hz. Tones were generated as pure sine waves using Matlab 2009a and were presented through Sony headphones (model MDR-NC6), with the noise cancelling function turned off.

Results

As before, the data from the $N=2$ condition showed high accuracy and fast mean RT, both for the 2-then-8 condition (98% and .65 sec.) and the 8-then-2 condition (98% and .58 sec.). Figure 4 shows mean accuracy and RT as functions of stimulus rank for the $N=8$ conditions. We extended the linear interaction contrasts employed in Experiment 1 to directly test the difference between conditions by comparing the linear contrast from the two groups (against the appropriate pooled error term from a mixed ANOVA with factors stimulus rank and experimental condition). This contrast showed that the difference between the mean accuracy for stimuli #2 and #7 and stimuli #4 and #5 was significantly larger in the 8-then-2 condition compared to the 2-then-8 condition ($F(1,161)=3.86, p=.03$). The corresponding test for the RT data was not significant ($p=.07$). Separate linear contrasts for each condition (as in the analysis of Experiment 1) further confirmed these trends for the accuracy data: in the 2-then-8 condition, the mean accuracy for stimuli #2 and #7 was not significantly different than for stimuli #4 and #5 ($p=.20$), but this comparison was significantly different in the 8-then-2 condition ($F(1,77)=10.3, p<.001$). The corresponding tests for the mean RT data did not reach significance.

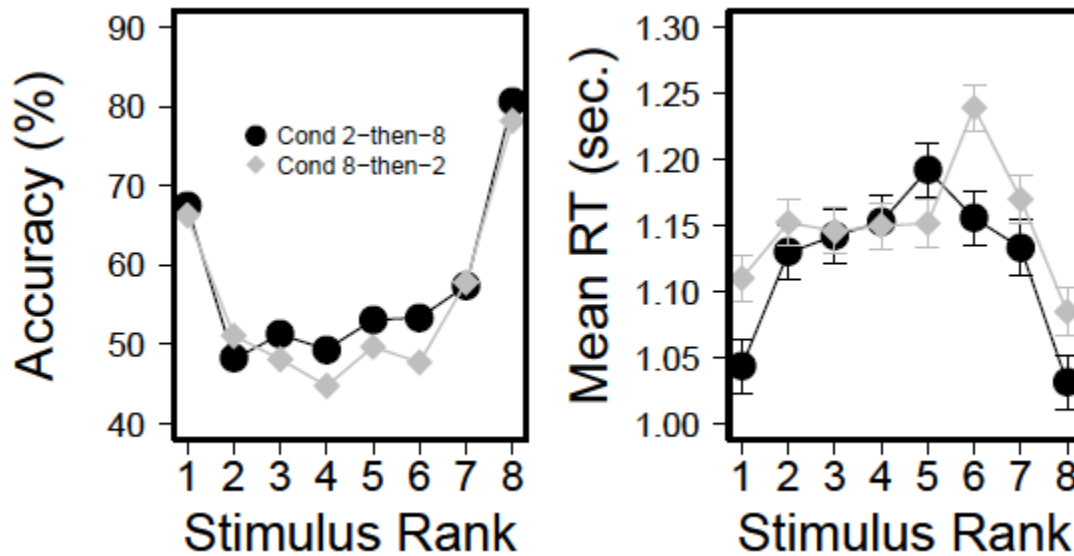


Figure 4. Accuracy and RT as a function of stimulus magnitude for Experiment 3. Note that there are two conditions: some participants experienced the $N=2$ set size first, then $N=8$ (condition 2-then-8) and others experienced the reverse order (condition 8-then-2). Error bars are as in Figure 2.

Discussion

Experiment 3 demonstrates a significant difference in the bow effects observed in the 8-then-2 vs. the 2-then-8 condition. The bow effect for response accuracy (but not RT) was deeper in the 8-then-2 condition than in the 2-then-8 condition, which is consistent with the hypothesis that the bow effect was disrupted (flattened) by pre-exposure to the central stimuli in the 2-then-8 condition. The null effects for the RT data are surprising, given the large RT differences in the corresponding test in Experiment 1. This null effect might be due to low power, because the variability we observed in RT data was much larger than for accuracy data (e.g., see the very slow mean RT peaks for some stimuli in Figure 4). Alternatively, the null effect might have been caused by a speed-accuracy tradeoff. If performance is improved for frequently-presented stimuli, participants can choose to exhibit that improvement either as improved decision accuracy,

or as improved response time (or both). Such tradeoffs are complex, and often accompany improved performance in simple decision tasks (see, e.g., Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009).

Response Biases

In most categorization paradigms, increasing the presentation frequency of some stimuli relative to others alters participants' *a priori* response biases in such a way as to improve performance for the frequently-presented stimuli (for a general review, see Healy & Kubovy, 1981; or see Petrov & Anderson, 2005, for an example in absolute identification). It is possible that the tendency for our participants to demonstrate improved accuracy for stimuli in the centre of the range is due to such response biases. Here, we attempt to rule this explanation out, leaving open the possibility that practice improved discrimination performance itself.

Results from Experiment 1 provide an initial insight into this issue. Experiment 1 provided a direct comparison between the identification of tone loudness and line length. Performance was improved for over-presented lines, but not for over-presented tones and one would expect that if results were caused solely by response biases towards over-presented stimuli, performance should have increased for both stimulus modalities. However, there are difficulties comparing such results across different stimulus modalities, in our case because the dimensions studied do not have identical Weber fractions (i.e., the minimum separation required between adjacent stimuli so that each are equally perceptually discriminable).

To address this issue, we further analyse the results from Experiment 3, which are ideal for examining issues of response bias. In Experiment 3, both conditions use the same stimuli, so the pairwise discriminability of the stimuli is identical by design. Additionally, Experiment 3 did not include any blocks with unequal stimulus presentation frequencies, so participants were not given any *a priori* reason to employ unequal response bias. Figure 5 shows the marginal response probability – that is, the probability that each stimulus label is used as a response. Response probability typically demonstrates similar phenomena (including the bow effect) to accuracy and RT (Petrov & Anderson, 2005), but allows examination of bias. Given that the stimuli were presented equally often, increased marginal response probability for a stimulus indicates a response bias towards that stimulus on the part of the observer. For consistency with earlier analyses, and because of the similarity between patterns in response probability and patterns in response accuracy and RT, we have used the same inferential analyses for these data as used above. Linear contrasts comparing the central stimuli for both conditions with the next-to-edge stimuli (#2 and #7), using the appropriate mixed ANOVA error term, showed that there were no significant differences between conditions in the amount of bow observed in response probability ($p=.22$). A power analysis showed that a relatively small difference between conditions (e.g., a difference in the depth of the bow effect for the two conditions of just 2% in marginal probability) would have been detected with probability 76%, using a Type I error rate of .05. These results suggest that differences in response biases are unlikely to be a contributing factor to the significant improvements in performance found for participants in the 2-then-8 condition.

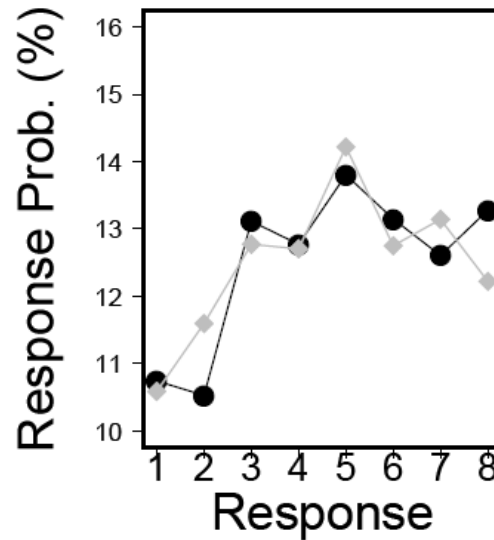


Figure 5. Response probability for Experiment 3

General Discussion

Experiments 1 and 2 indicated that the bow effect can be disrupted by design factors, such as within-subject manipulations and stimulus presentation probabilities, at least when the stimuli are line lengths. In Experiment 1, we found that, following the presentation of the $N=2$ condition, the stimuli in the middle of the range were identified more accurately compared to the surrounding stimuli for line lengths but not for tone loudnesses. Experiment 2 indicated that the disrupted bow effect for line lengths in Experiment 1 was due to unequal stimulus presentation frequency across the experiment, induced by a standard within-subject manipulation of set size. Experiment 3 suggests that the modulation of the bow effect depends on stimulus modality, as would be expected from Dodds et al.'s (2011) results.

Previous Results on Unequal Stimulus Presentation Frequency

Petrov and Anderson (2005) manipulated the presentation frequency of stimuli (dots varying in separation) in an absolute identification experiment and found that

correct responses were more likely for stimuli presented more frequently. However, Petrov and Anderson's presentation frequency manipulation was counterbalanced over short time scales, such that long-term learning effects (as we observed) would not be expected in longer time-scale averages. Cuddy (1968, 1970) found that training participants on just one particular stimulus, out of a set of nine tones varying in frequency, resulted in improvement for the entire set. However, this result was limited only to highly trained musicians – regular participants showed little improvement. In the most similar work to our own, Cuddy et al. (1973) found that regular participants were able to greatly improve their performance when trained by presenting three tones out of a set of nine more frequently than others. However, Chase et al. (1983) replicated Cuddy et al.'s experiment and found very small improvements (8%, as opposed to Cuddy et al.'s 50% improvement).

Our experiments extend these earlier findings in several ways. Firstly, we examined performance in conditions where all stimuli are presented equally often (after having manipulated presentation frequency earlier). This more faithfully represents performance in standard absolute identification tasks. Cuddy (1968, 1970) used a similar procedure, but found changes in performance only for musically trained participants. Secondly, we demonstrated effects of the presentation frequency for stimuli across the entire experiment. That is, stimuli that were presented more often over the entire experiment were identified more accurately, even when every block of trials contained equal presentation frequencies for all stimuli in the block (Experiments 1 and 3). This manipulation mirrors standard within-subject manipulations of stimulus set size, and avoids creating a situation that rewards response biases in favour of more frequent

stimuli. Thirdly, our experiments systematically examine different stimulus types, which have predictable and large effects on the results.

Absolute identification can be thought of as a variant of categorization, in which each stimulus defines its own category. In standard categorization tasks, where many different stimuli are mapped to the same response (a single category), there have been many investigations of the effect of unequal stimulus presentation frequency, with results that are consistent with ours. For example, Nosofsky (1988) found that frequently-presented category exemplars were classified more accurately and rated as more typical of the category than less-frequently-presented exemplars. This effect generalized to unseen exemplars that were very similar to the more-frequently-presented exemplars, but not to less similar ones; analogous to our stimulus-specific findings.

Theoretical Implications

Our results are indicative of long-term learning. This adds weight to recent findings that practice can improve performance in absolute identification (e.g., Rouder et al., 2004) and that these effects are larger for line length and tone frequency than for tone loudness (Dodds et al., 2011). An additional theoretical implication from our results is that learning is stimulus-specific. For example, suppose learning effects were instead driven by time-on-task (or the total number of absolute identification decisions). Under that assumption, additional presentations of some stimuli would not lead to improved performance for those particular stimuli above others, contrary to our results. Further, improved performance for frequently presented stimuli was observed to last for hours or days, long after uniform presentation frequencies were re-established. This suggests that theoretical accounts of improved performance for frequently presented stimuli based on

short-term biasing mechanisms (e.g., Petrov & Anderson's, 2005, ANCHOR model) are not sufficient.

Although current theories for absolute identification do not include mechanisms by which practice can improve performance, there are several obvious candidate mechanisms. Some of these candidates seem better suited to meeting the challenges described above than others. For example, exemplar-based models (e.g., Kent & Lamberts, 2005) naturally predict that increased exposure to some stimuli enriches the representation of those stimuli above others. Kent (2005, Chapter 9) suggests a precise mechanism that would have this effect - a particular relationship between the number of exemplars and the associated psychological distances.

The selective attention component of the SAMBA (Brown, Marley, Donkin & Heathcote, 2008) and ANCHOR (Petrov & Anderson, 2005) models, both explain the phenomenon known as “contrast” (the tendency for a response on the current trial to be biased away from those presented more than one trial previously) by assuming that recently-presented stimuli have privileged representations in memory – psychological space effectively expands around these representations, increasing their distances from other stimulus representations. Such mechanisms might naturally accommodate improved performance due to extra stimulus presentations, because extra presentations of a stimulus usually lead to a higher probability of that stimulus having been presented in the recent past. However, both SAMBA and ANCHOR assume that these changes are very short-lived (lasting only a few trials, or perhaps on even shorter time scales – see Matthews & Stewart, 2009). This assumption would have to be altered to allow the contrast mechanisms to explain our results.

One further theoretical constraint – the observed differences between stimulus types – has interesting implications for these possible accounts based on contrast mechanisms. Standard contrast effects occur for all stimulus types (e.g. Ward & Lockhead, 1971), so it is not immediately clear why a contrast mechanism (in SAMBA or ANCHOR) should allow for disrupted bow effects using line length and tone frequency, but not for tone loudness. An intriguing possibility was raised by Dodds et al.'s (2011) finding that, when extended practice improves performance, the standard contrast effect disappears. It is possible that extra practice with frequently presented stimuli alters the contrast mechanism, to the extent that learning occurs, by fixing in place the expanded psychological representation.

It is a matter for future research to identify why this might occur for some stimulus sets (such as line lengths and tone frequencies) but not others (such as tone loudness). This account might be tested in future work by examining contrast effects in paradigms that, as in ours, involve differential stimulus presentation frequencies. Existing experiments, including ours, are not suitable for such analyses because the frequently presented stimuli have always been the central stimuli, and contrast effects are not observed for those stimuli (Ward & Lockhead, 1971).

Acknowledgments

This research was supported by Australian Research Council Discovery Project 0881244 to Brown & Heathcote and by Natural Science and Engineering Research Council Discovery Grant 8124-98 to the University of Victoria for Marley. We thank Chris Kent for providing us with the data from Kent and Lamberts' (2005) Experiment 2 and Kent's (2005) Experiment 5.

References

- Brown, S. D., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological review*, 115(2), 396-425.
- Chase, S., Bugnacki, P., Braida, L. D. & Durlach, N. I. (1982). Intensity perception: XII. Effect of presentation probability on absolute identification. *Journal of the Acoustical Society of America*, 73(1), 279-284.
- Cuddy, L. L. (1968). Practice effects in the absolute judgment of pitch. *The Journal of the Acoustical Society of America*, 43(5), 1069-1076
- Cuddy, L. L. (1970). Training the absolute identification of pitch. *Perception & Psychophysics*, 8(5A), 265-269
- Cuddy, L. L., Pinn, J. & Simons, E. (1973). Anchor effects with biased probability of occurrence in absolute judgment of pitch. *Journal of Experimental Psychology*, 100(1), 218-220.
- Dodds, P., Donkin, C., Brown, S. D., Heathcote, A. (2011). Practice effects in absolute identification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 37(2), 477-492.
- Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review*, 16(6), 1026-1036.
- Gelman, A. & Stern, H. (2006). The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant. *The American Statistician*, 60(4) 328-331.

- Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory*, 7(5), 344-354.
- Kent, C. (2005) . *An exemplar account of absolute identification*. Ph. D. Thesis, Department of Psychology, The University of Warwick.
- Kent, C., & Lamberts, K. (2005). An exemplar account of the bow effect and set-size effects in absolute identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2), 289-305.
- Lacouture, Y. (1997). Bow, range, and sequential effects in absolute identification: A response-time analysis. *Psychological Review*, 60, 121-133.
- Lacouture, Y., Li, S., & Marley, A. A. J. (1998). The roles of stimulus and response set size in the identification and categorisation of unidimensional stimuli. *Australian Journal of Psychology*, 50(3), 165-174.
- Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subjects designs. *Psychonomic Bulletin & Review*, 1, 476-490.
- Matthews, W. J. & Stewart, N. (2009). The effect of interstimulus interval on sequential effects in absolute identification. *The Quarterly Journal of Experimental Psychology*, 62(10), 2014-2029.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits in our capacity for processing information. *Psychological Review*, 63(2), 81-97.
- Nosofsky, R.M. (1988). Similarity, frequency and category representations. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14, 54-65.

- Petrov, A. A., & Anderson, J. R. (2005). The dynamics of scaling: A memory-based anchor model of category rating and absolute identification. *Psychological Review*, 112(2), 383-416.
- Petrov, A. A., Doshier, B., & Lu, Z. (2005) The dynamics of perceptual learning: An incremental channel reweighting. *Psychological Review*, 112(4), 715-743.
- Rouder, J. N., Morey, R. D., Cowan, N., & Pfaltz, M. (2004). Learning in a unidimensional absolute identification task. *Psychonomic Bulletin & Review*, 11(5), 938-944
- Shiffrin, R. M. & Nosofsky, R. M. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review*, 101(2), 357-361.
- Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, 112(4), 881-911.
- Ward, L. M. & Lockhead, G. R. (1971). Response system processes in absolute judgment. *Perception & Psychophysics*, 9(1B), 73-78.