Revisiting Miller's Limit:

Studies in Absolute Identification

Pennie Dodds – BPsych (Hons)

Thesis Submitted for Doctorate of Philosophy, September 2011

Statement of Originality

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying subject to the provisions of the Copyright Act 1968.

Acknowledgment of Collaboration

I hereby certify that the work embodied in this thesis has been done in collaboration with other researchers. I have included as part of the thesis a statement clearly outlining the extent of collaboration, with whom and under what auspices.

Thesis by Publication

I hereby certify that this thesis is in the form of a series of published papers of which I am a joint author. I have included as part of the thesis a written statement from each co-author, endorsed by the Faculty Assistant Dean (Research Training), attesting to my contribution to the joint publications.

Pennie Dodds

Publications Included in Thesis

In order of reference:

Dodds, P., Donkin, C., Brown, S. D. & Heathcote, A. (2011) Increasing Capacity: Practice Effects in Absolute Identification *Journal of Experimental Psychology: Learning, Memory & Cognition, 37(2), 477-492.*

Dodds, P., Donkin, C., Brown, S. D., Heathcote, A., & Marley, A. A. J. (2011) Stimulus-Specific Learning: Disrupting the Bow Effect in Absolute Identification. *Attention, Perception & Psychophysics*

Brown, S.D., Marley, A.A.J., **Dodds, P**., & Heathcote, A. (2009) Purely relative models cannot provide a general account of absolute identification. *Psychonomic Bulletin & Review, 16, p.583-593*

Dodds, P., Brown, S. D., Zotov, V., Shaki, S., Marley, A. A. J. & Heathcote, A. (2011). *Absolute production and absolute identification.* Manuscript submitted for publication

Dodds, P., Donkin, D., Brown, S.D., Heathcote, A. (2010) Multidimensional scaling methods for absolute identification data *In S. Ohlsson & R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society. Portland, OR: Cognitive Science Society*

Dodds, P., Rae, B. & Brown, S. D. (2011). *Perhaps Unidimensional is not Unidimensional.* Manuscript submitted for publication

THE UNIVERSITY OF NEWCASTLE AUSTRALIA

19th September 2011

To whom it may concern,

This letter outlines Pennie Dodds' contribution to the series of papers that are submitted as a part of her PhD. All papers that are contributing to her thesis are listed below, with a statement of her contribution for each.

Regards,

Associate Professor
Scott Brown

Professor
Andrew Heathcote

Emeritus Professor
A. A. J. Marley

Babette Rae
PhD Candidate

Doctor
Chris Donkin

Doctor
Samuel Shaki

Doctor
Vladimir Zotov

Endorsed By

Dodds, P., Brown, S. D., Zotov, V., Shaki, S., Marley, A. A. J. & Heathcote, A. Reconciling absolute identification and absolute production: a method of examining internal magnitude representation. *Submitted*

> *This project was led by Pennie, 50% contribution. Pennie coordinated and supervised data collection, completed all data analyses, and took the lead role in manuscript preparation. Other authors contributed as follows: S. Brown (10%), V. Zotov (10%), S. Shaki (10%), A. A. J. Marley (10%), A. Heathcote (10%).*

Dodds, P., Rae, B. & Brown, S. D. When unidimensional is not unidimensional. *Submitted*

> *This project was led by Pennie, 40% contribution. Pennie completed all data analyses, took the lead role in manuscript preparation, and collected most of the data that were analysed. The other PhD student on the project (Babette Rae) contributed 40%, Associate Professor Brown contributed 20%.*

Dodds, P., Donkin, C., Brown, S. D., Heathcote, A., Marley, A. A. J. (2011) Stimulus-Specific Learning: Disrupting the Bow Effect in Absolute Identification. *Attention, Perception & Psychophysics, 73*(6), 1977-1986

> *This project was led by Pennie. She conducted all data collection and all analyses, and was primarily responsible for manuscript preparation. Numerically, the contributions from the authors were: Pennie Dodds, 50%; Chris Donkin, 20%; Scott Brown, Andrew Heathcote & A.A.J. Marley, 10% each.*

Dodds, P., Donkin, C., Brown, S. D. & Heathcote, A. (2011) Increasing Capacity: Practice Effects in Absolute Identification. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *37*(2), 477-492

> *This project was very large (many experiments, over several years) and was jointly led by Pennie Dodds and Scott Brown. Pennie was responsible for*

almost all of the extensive data collection, all of the data analyses and most of the manuscript preparation. Numerically: Pennie Dodds 50%, Chris Donkin 30%, Scott Brown 10%, Andrew Heathcote 10%.

Dodds, P., Donkin, D., Brown, S.D., Heathcote, A. (2010) *Multidimensional scaling methods for absolute identification data* In S. Ohlsson & R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society. Portland, OR: Cognitive Science Society.

Pennie was responsible for all data collection, analyses and manuscript preparation, with little except advice and feedback provided by the other authors. This equated to approximately a 90% contribution.

Brown, S.D., Marley, A.A.J., **Dodds, P**., & Heathcote, A.J. (2009) Purely relative models cannot provide a general account of absolute identification. *Psychonomic Bulletin & Review, 16*, p.583-593

Pennie was responsible for experiment design, data collection and manuscript review. Approximately a 30% contribution, with Brown and Marley contributing 25% each and Heathcote 20%.

Acknowledgements

For responding to an infinite number of emails and always being willing to indulge my questions and concerns, I would like to thank Associate Professor Scott Brown. Without such an enthusiastic, infectiously happy and supportive supervisor, I would not be in a situation to submit this thesis. I would also like to thank Professor Andrew Heathcote for his support and encouragement throughout my years in the Newcastle Cognition Lab.

To the people that have gone through the Cognition Lab in the last few years – thanks for the chats and coffee trips that kept me awake and motivated. Without your support and friendship the lab would not be the comfortable workplace that it is. I can say without doubt I will miss spending each day with every one of you.

To David Elliott for putting up with my constant questions, for emergency IT support and for being a constant source of great stories and wisdom – thank you for your time, patience and friendship. A special mention also goes to Bebe Gibbins, a good listener, supportive friend and constant bundle of energy and laughs.

And most of all, thank you to my ever-supportive and incessantly even-tempered husband, Ray – no words can thank you enough for putting up with my complaints and anxiety over the past few years. I am continually astounded with your patience and grace and I am forever grateful for your love and support over the past few years.

Table of Contents

Abstract

Absolute Identification is a seemingly simple cognitive task that provides researchers with a number of interesting and complex phenomena. The task provides evidence towards an information processing capacity, which Miller (1956) popularised with his magical number 7±2 – a number of which he suggested is reminiscent of the number of unidimensional items (or chunks in short term memory) that an individual should be able to learn to perfectly identify. This limit has long since been accepted as a truism of absolute identification research, with much further model development accepting this as a known intrinsic "quirk" of absolute identification performance. This thesis begins with results that are in stark contrast to Miller's findings – we find that given moderate practice, participants are able to improve their performance significantly. The following chapters describe further investigation into this contrary result, and provide an overview of current models of AI performance. Through a series of published and submitted papers, we investigate the possibility that rather than disproving Miller's theory of an information processing capacity, we might have further refined the absolute identification paradigm. Close examination of common stimuli used in absolute identification tasks reveals that while common stimuli such as line lengths and dot separation are physically unidimensional, the psychological representation of these stimuli may be multidimensional. Interestingly, the sole stimulus modality that did *not* exhibit learning effects – tone loudness - did *not* appear to be represented on multiple dimensions. Without the assumption of uni-dimensionality, we cannot suggest the results are due to some difference in information processing capacity, but are rather more likely an artefact of stimulus perception. These results have significant consequences for the future of absolute identification research: it would appear that

absolute identification researchers should restrict their use of stimulus modalities to only tones varying in loudness.

Miller's (1956) seminal review of short term memory and absolute identification popularised the phenomenon of a limitation to our memory. Despite the seemingly infinite memory people have for every-day objects such as faces, names and letters, Miller highlights a severe and robust limit to the number of items we are able to store in short term memory, whether they be *chunks* of information or specific stimuli from a then-common cognitive task called absolute identification. This limit of 7±2 stimuli was referred to as a *bottleneck* in information processing capacity – using information theory, Miller likened this limit to a narrow pathway in a communication channel, creating a bottleneck and resulting in a limit to our processing capacity, or creating a *capacity limit*. In order to examine this limitation to our memory, we can use a task called absolute identification, a deceptively simple cognitive memory experiment.

A typical absolute identification (AI) task uses stimuli that vary on a single dimension. For example, tones varying in intensity or frequency, or lines varying in length or angle. The participant is first presented with a set of these stimuli, each labelled with a unique referent – usually a number from 1 through to *n*. The participant is then presented with randomly selected stimuli from the set, and asked to try and remember the label that was previously associated with it. If a participant is incorrect, the correct answer is displayed. This seemingly simple task can be used to examine something infinitely more complex: human information processing capacity. By calculating the amount of information transferred from stimuli to responses, where an increase in information transfer is analogous to an increase in memory for stimulus items, researchers can infer processing capacity and examine limitations to memory.

Benchmark Phenomena

Despite the seeming simplicity of the task, absolute identification provides a

plethora of interesting phenomena. A majority of this thesis is devoted to an examination of existing benchmark phenomena in absolute identification performance, with a further section describing the theoretical implications of these findings. A brief description of each of the phenomena encountered is provided below, together with a description of how it relates to this thesis.

Limitation in Learning

One of the most robust phenomena associated with performance in absolute identification tasks is perhaps the fundamental limit to learning: Miller's (1956) review of the absolute identification literature linked this limit to chunking in short term memory, suggesting that people are able to only learn to perfectly identify approximately 5 to 9 stimuli (his magical number 7±2). This notion was supported by volumes of literature and was assumed to be resistant to practice (e.g. Pollack, 1952; Garner, 1953; Weber, Green & Luce, 1977; Lacouture, Li & Marley, 1998; Shiffrin & Nosofsky, 1994). The accepted doctrine in the field is that people are able to improve their performance slightly, but reach a low level asymptote and do not improve any more.

Pollack (1952) and Hartman (1954) both used absolute identification of tones varying in pitch and found that even up to eight weeks of testing failed to result in perfect performance. This phenomenon has even been replicated in participants with perfect (or absolute) pitch: the ability to identify any given musical note (e.g. Hsieh & Saberi, 2007). Those fortunate enough to claim this ability find it difficult to identify pure tones varying in pitch because musical tones (to which they are accustomed) contain overtones and harmonics that increase the number of dimensions to the stimulus.

Garner (1953) also provides evidence for the lack of learning in the absolute

identification of tones varying in loudness. Garner found that even when participants completed 12,000 trials, maximum information transmission was still low (1.62 bits, equivalent to perfect identification of only 3.1 stimuli). Weber et al. (1977) perhaps provide the most evidence against substantial learning in absolute identification: they also used 12,000 trials, and compared the performance on the initial and last 2,000 trials to calculate change in accuracy. Even with monetary incentives and significant practice however, identification accuracy was only shown to improve by an average of 0.06. Interestingly, Weber et al. (1977) only used six stimuli, which one would expect could facilitate perfect performance.

The notion of imperfect performance was further popularised by Shiffrin and Nosofsky (1994), who provided a humorous anecdote as part of their prominent review on the conundrum of absolute identification: as a young graduate student, Nosofsky believed that he could surely improve his identification of a set of tones varying in loudness. After locking himself up in a sound proof booth for several weeks, he was forced to conclude that the only thing that had improved was his need for psychotherapy.

The consequence of literature that supports this notion of imperfect performance is that this became a truism of absolute identification performance: people cannot learn to perfectly identify beyond nine stimuli. As a further consequence, no current model of AI performance takes into account learning effects. At the most, models only account for limited trial-to-trial memory, used to create sequential effects (e.g. SAMBA; Brown, Donkin, Heathcote & Marley, 2008) or short-term learning of the stimulus-to-response mapping (Petrov & Anderson, 2005). More recent research however, questions this truism and highlights a need for model revision (Rouder, Morey, Cowan & Pfaltz, 2004; Dodds, Donkin, Brown & Heathcote, 2011).

Contrary to the above literature, Rouder et al. (2004) and Dodds et al. (2011a; Section One, Chapter One of this thesis) describe situations in which participants are able to improve their performance significantly. Rouder et al. (2004) describe an experiment where participants are asked to identify a number of different sets of lines varying in length. One participant was able to learn to perfectly identify up to 20 line lengths – a number significantly beyond Miller's (1956) limit of 7±2 stimuli. Dodds et al. (2011a) replicated this learning effect across several different stimulus dimensions, including line length, tone frequency, line angle and dot separation. The one exception to this new phenomenon appeared to be tone loudness – performance did not improve very much for this stimulus modality, and in general appeared to resemble the same low-level performance pattern shown in earlier research. Section One, Chapter One (Dodds et al., 2011a) discusses this concept in more detail, and begins the investigation into why this phenomenon might be occurring for only some stimulus modalities.

Bow & Set Size Effect

Another common benchmark phenomenon in absolute identification is the bow effect. The bow effect is so-named because of the shape the curve makes when stimulus magnitude is plotted against accuracy or reaction time (RT). These curves are called *serial position curves*. Performance is better and reaction times are faster for stimuli at the edges of the stimulus range compared to the centre of the stimulus range, creating a bow shape in the curve (e.g. Kent & Lamberts, 2005; Lacouture & Marley, 2004; Lacouture, 1997). It has been shown that the bow effect is resistant to numerous experimental manipulations, including stimulus spacing (Lacouture, 1997), set size (or number of stimuli in the set; Stewart, Brown & Chater, 2005), and is even consistent across different stimulus modalities (Dodds et al., 2011a). Figure 1 provides an example of the bow effect taken from Dodds et al. (2011a).

The set size effect is a related phenomenon, and refers to the increase in the prominence of the bow in a serial position curve as set size increases (e.g. Pollack, 1953; Weber, Green & Luce, 1977). Stimuli in the set are identified more poorly, and the bow in the serial position curve *increases,* as more stimuli are added to the set. See Figure 2 for an example of the set size effect. This increase in the bow effect as set size increases is interesting, as it is still apparent even when *the same stimuli are used* for each set size in question. For example, an individual can often perfectly identify a set of two stimuli ($N = 2$), but when N is increased to 4, 8 or more, even if those same two stimuli are used as a part of these new stimulus sets, performance for these unchanged stimuli decreases markedly, both in accuracy and reaction time.

Section One, Chapter Two of this thesis examines the bow effect, and discusses how apparently simple experimental manipulations influence what was a previously assumed to be a robust phenomenon. Dodds, Donkin, Brown, Heathcote and Marley (2011b) discuss how a simple within-subjects manipulation of set size can create abnormalities in the bow plot. This appears to be due to over-presentation of certain stimuli – for example, in a within-subjects manipulation of set size, certain stimuli are naturally presented more often than others. Dodds et al. (2011b) demonstrate that we are able to learn to identify certain stimuli relatively well, and so for these same stimuli on which we are able to learn, we also see this abnormality in the bow effect (Dodds et al., 2011b), suggesting the ability to learn these stimuli is increasing identification accuracy and distorting the serial position curve. Chapter Two (Dodds et al., 2011b) discusses this concept further, examining the implications of this for absolute identification experimental design and model development.

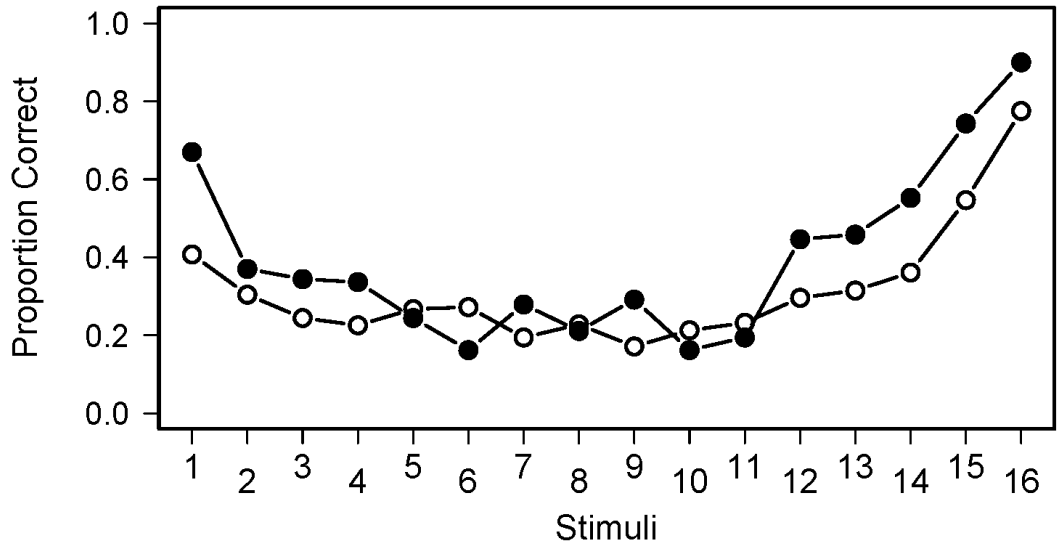


Figure 1 . The bow effect shown in a serial position curve from Dodds et al. (2011a). The plot shows improvement in performance for stimuli at the edges of the stimulus range, in comparison to those in the centre. Different lines represent the first and last sessions of the experiment – the experiment from which this was taken consisted of ten sessions of training in the task.

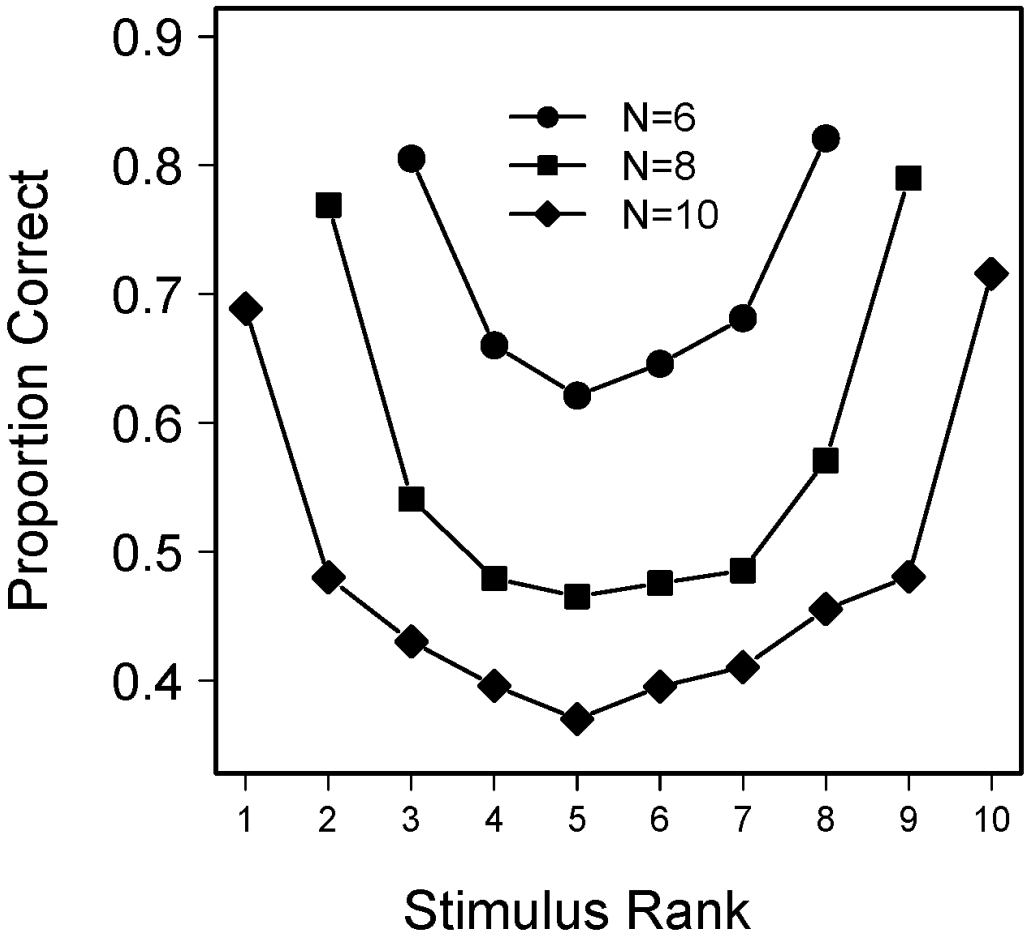


Figure 2. An example of the set size effect, from Stewart et al. (2005) Experiment 1

Sequential Effects

Sequential effects refer to the effect that the previous stimuli have on current responses (e.g. Garner, 1953; Lacouture, 1997; Petrov & Anderson, 2005; Brown et al., 1998). There are two main types of sequential effects: assimilation and contrast. Assimilation refers to the tendency for the current response to be biased *towards* the stimulus presented on the previous trial. Contrast refers to the tendency for the current response to be biased *away* from stimuli presented on trials further back than *n*-1. Figure 3 shows an idealised version of an impulse plot: a plot that graphs assimilation and contrast.

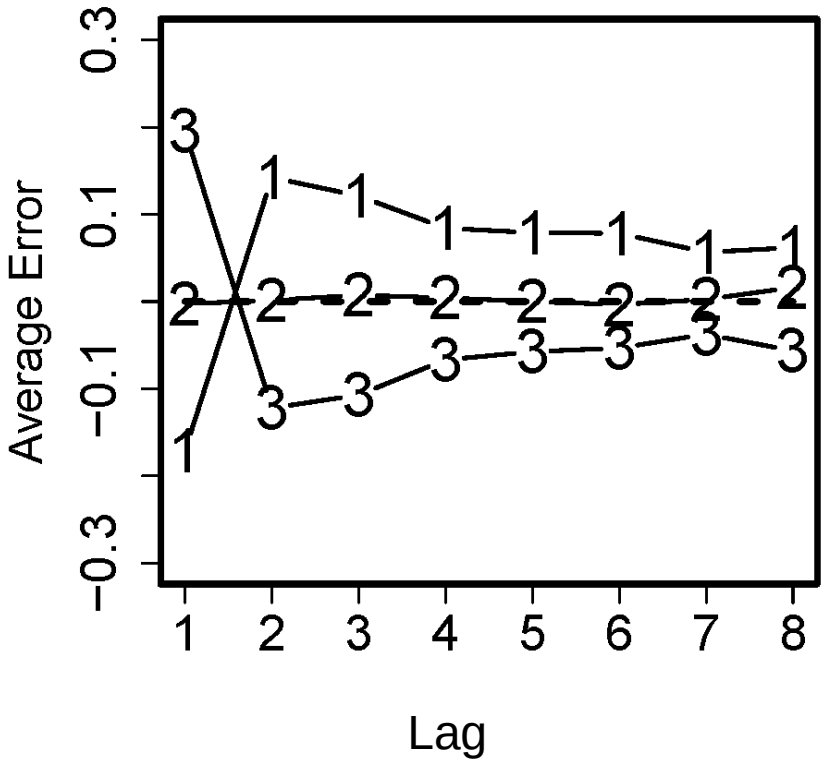


Figure 3. An idealised impulse plot taken from Dodds et al. (2011a). In Dodds et al. this was created from simulated data. Numbers on the lines refer to stimulus groups present on previous trials: 1 refers to low magnitude stimuli, 2 refers to middle range magnitude stimuli, and 3 refers to high magnitude stimuli. Assimilation is shown at lag = 1: on trial n-1, error rates are likely to be negative when previous stimuli were low in magnitude and high when previous stimuli are high in magnitude, because response are bias towards the n-1 stimulus. Contrast is shown at lag > 1, and shows the opposite pattern: error rates are likely to be high for low magnitude stimuli and low for high magnitude stimuli.

Sequential effects are a robust phenomenon of AI performance; they appear in every data set, albeit in different magnitudes. Some authors propose that these effects provide valuable insight into information processing (Stewart et al., 2005). Section One, Chapter One (Dodds et al., 2011a) discusses how sequential effects may also provide some insight into the mechanisms behind the learning effect – Dodds et al. show that as practice increases, contrast substantially reduces. This is adaptive, as contrast contributes towards errors in responses.

Stimulus Separation

Generally it is assumed that so long as stimuli are perfectly perceptually discriminable (that is, they are separated by a value much greater than the Weber fraction for that stimulus modality; see Laming, 1986; Teghtsoonian, 1971), increasing the spacing between adjacent stimuli in an absolute identification task leads to only small improvements in performance (e.g. Pollack, 1952, Lacouture, 1997). For example, Pollack (1952) examined the difference between identification of a set-number of tones varying in pitch over a small range (400Hz) and a large range (8000Hz) and found little difference in performance. This concept is analogous to the concept of channel capacity: there appears to be a fixed limit to the amount of information that can be transferred from stimulus to response, regardless of both practice (as described earlier) and other experimental manipulations such as stimulus spacing.

Absolute vs. Relative Models

The manipulation of stimulus separation also provides a unique opportunity to distinguish between two broad types of absolute identification models: absolute and relative models. The fundamental difference between the two types of models is the use of long term memory elements: *relative* models (e.g. Laming, 1984; Stewart et al.,

2005) deny the use of any long term memory elements, whereas *absolute* models (e.g. Marley and Cook, 1984; Petrov and Anderson, 2005; Lacouture and Marley, 2004) utilise the concept of long term representation.

Absolute models incorporate long term memory by assuming that individuals maintain long term magnitude information. For example, Petrov and Anderson's (2005) ANCHOR model posits that "anchors" in memory are compared to the current stimulus in order to generate a magnitude estimation. On the other hand, a purely relative model would base stimulus judgments only on the most recently-seen stimuli (the judgments of current stimulus magnitude are *relative* to the previously displayed stimuli). In practice however, even relative models require a form of long term memory. For example, Stewart et al.'s (2005) Relative Judgment Model (RJM) requires long term memory for the spacing between stimulus magnitudes (λ) and for the stimulus set size.

Lockhead and Hinson's (1986, Experiment 1) unequally-spaced stimuli experiment provides an opportunity to distinguish between absolute and purely relative models. This is discussed further in Section Two, Chapter Three, *Purely Relative Models Cannot Provide a General Account of Absolute Identification* (Brown, Marley, Dodds & Heathcote, 2009). In summary, we replicated Lockhead and Hinson's (1986) design, with three conditions –one condition used evenly spaced stimuli, and two used unevenly spaced stimuli. A purely relative model would assume a single value for stimulus separation e.g. 2dB, which, in the case of the evenly spaced stimuli, would not present a problem for either relative nor absolute models. In unevenly spaced stimulus sets however, a memory for a *single* stimulus separation value is insufficient, hence relative models should not predict appropriate patterns of performance. Therefore such unevenly spaced stimulus sets provides a unique opportunity to distinguish between these two general model categories.

Psychological Representation of Stimuli

Several papers within this thesis have been based on the main findings from the Section One, Chapter One *Increasing Capacity: Practice Effects in Absolute Identification* (Dodds et al. 2011a) – that learning is possible in absolute identification tasks for certain stimulus modalities. Dodds et al. demonstrate a learning effect for line length, line angle, dot separation and tone frequency, but not for tone loudness. These results could represent an important theoretical finding, something previous research has neglected to observe. Alternatively however, there might be a more simple, methodological explanation. For example, in order to examine capacity limitations, stimuli must vary on only a single *dimension*. As more dimensions are added to stimuli, they become more easily identifiable – e.g. we have an apparently unlimited memory for names, or faces. It is possible that the type of stimuli employed in Rouder et al. (2004) and Dodds et al. (2011a) were responsible for the learning effect, rather than some higher cognitive process.

Section Two, Chapter Five *Multidimensional Scaling Methods for Absolute Identification Data* (Dodds et al., 2010) and Chapter Six *Perhaps Unidimensional is not Unidimensional* (Dodds, Rae & Brown, submitted) examine the possibility that those stimulus types that support learning effects have more complex psychological representations compared to those that do not. In the same way that colour is perceived on multiple dimensions, and yet only varies on a single physical dimension, Rouder et al.'s (2004) and Dodds et al.'s (2011a) stimuli may give rise to more complex psychological representations than originally anticipated. If stimuli are indeed more complex, the learning effects found by Rouder et al. and Dodds et al. are unsurprising and do not challenge Miller's (1956) fundamental limit – because that it is accepted that we have good memories for complex, or multidimensional, stimuli.

There are several means of examining the issue of stimulus complexity, two of which are discussed in this thesis. The first technique that we examine is Multidimensional Scaling (Cox & Cox, 1994; 2001). The second is a structural forms algorithm (Kemp & Tenenbaum, 2008). The two will be briefly discussed here, but are discussed in more detail in Section Two, Chapters Five and Six.

Multidimensional Analysis (MDS)

Multidimensional Analysis (MDS; Cox & Cox, 1994; 2001) refers to a range of statistical techniques employed by a variety of fields, designed to examine underlying structure in distance or proximity data. Given a matrix of either true distance data (e.g. kilometres, centimetres), or more subjective proximity data (e.g. similarity ratings), MDS allows the observer to infer the relationship between the objects in question by placing them within a higher-dimensional vector space. It also is possible to infer dimensionality with MDS by comparing analyses based on vector spaces with different dimensionality.

The most common example of the application of MDS analyses, is its use for geographic data. Say you were given a matrix of distance data that provides the distances between three cities: using MDS analyses, a spatial representation of the location of each of the cities can be developed, together with a stress value (a goodness-of-fit measure). There are two main types of MDS analyses: metric and non-metric MDS. Metric MDS is used for the analysis of true-distance data e.g. the distances between cities. Non-metric MDS on the other hand, is used for more subjective data, such as similarity ratings. The distinction between the two is important: non-metric MDS relaxes the metric assumptions placed on the distance measures, and instead assumes only that the observed distance measurements are monotonically related to some true distance measurement. Non-metric MDS is therefore well suited to

psychological data such as similarity ratings, as we can be relatively confident that the *rank* of these ratings is informative, but not their absolute values.

For this thesis, the purpose of MDS was to examine the dimensionality of our stimuli – particularly those stimuli which have previously exhibited learning. In Chapter Five (Dodds et al., 2010), we took 16 line lengths used in Dodds et al. (2011a) and asked participants for similarity ratings for every possible pair within the set (twice). The aim of this was to examine the structure of these stimuli and determine whether there were additional, unexpected dimensions. The difficulty with MDS arises in the interpretation of results: in order to examine the structure, one must choose a number of dimensions to test. For example, for the line lengths used in this experiment, one, two and three dimensions were tested. MDS outputs a goodness-of-fit measure (the stress value) to indicate how well any particular model fits the data, but does not provide any means for comparing between models. The stress value is also highly influenced by model complexity – as we increase the complexity of the model, the stress value decreases – not necessarily because the model is most appropriate for the data, but rather because the lower-dimensional models are nested within the higher-dimensional models. In addition, MDS provides no framework for inference, which is necessary when attempting to distinguish between several dimensions. Further discussion and results can be found in Section Two, Chapter Five *Multidimensional Scaling Methods for Absolute Identification Data* (Dodds et al., 2010).

Structural Forms Algorithm

An alternative method for examining the complexity of our stimuli is the structural forms algorithm (Kemp & Tenenbaum, 2008). The structural forms algorithm is a Bayesian technique based on a universal graph grammar that not only provides a framework for inference but also examines the fit of a range of different forms – not just

arrangements within a vector space - and the arrangement of objects within this form. For example, the technique provides a means of examining the plausibility of a chain, ring, cylinder, cluster, tree and hierarchical structure, among others. In these ways it is more suitable than MDS for examining dimensionality in our data because it allows coherent inferential testing. Importantly, the technique is also flexible in the type of data that is used. Kemp and Tenenbaum provide examples using a wide range of data types – e.g. feature rating data, distances between cities and colour similarity. In Section Two, Chapter Six, *Perhaps Unidimensional is Not Unidimensional* (Dodds, Rae & Brown, submitted), we discuss the use of this technique further, and its application using confusion matrices taken from absolute identification experiments.

Summary and Overview of Main Body

Table 1 provides a list of the papers included in my thesis, with section and chapter references. Chapters are grouped under two broad sections: empirical results and theoretical implications.

Table 1. A list of all papers included in my thesis, with chapter references

Section	Chapter	Reference
1	1	Dodds, P., Donkin, C., Brown, S. D. & Heathcote, A. (2011) Increasing Capacity: Practice Effects in Absolute Identification *Journal of Experimental Psychology: Learning, Memory & Cognition, 37(2), 477-492.*
	2	Dodds, P., Donkin, C., Brown, S. D., Heathcote, A., & Marley, A. A. J. (2011) Stimulus-Specific Learning: Disrupting the Bow Effect in Absolute Identification. *Attention, Perception & Psychophysics*
2	3	Brown, S.D., Marley, A.A.J., Dodds, P., & Heathcote, A.J. (2009) Purely relative models cannot provide a general account of absolute identification. *Psychonomic Bulletin & Review, 16, p.583-593*
	4	Dodds, P., Brown, S. D., Zotov, V., Shaki, S., Marley, A. A. J. & Heathcote, A. (submitted). *Absolute production and absolute identification.*
	5	Dodds, P., Donkin, D., Brown, S.D., Heathcote, A. (2010) Multidimensional scaling methods for absolute identification data *In S. Ohlsson & R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society. Portland, OR: Cognitive Science Society.*
	6	Dodds, P., Rae, B. & Brown, S. D. (Submitted). *Perhaps Unidimensional is not Unidimensional*

Section One: Empirical Results

Section One of this thesis consists of two papers

1. Dodds, Donkin, Brown and Heathcote (2011) *Increasing Capacity: Practice Effects in Absolute Identification*
2. Dodds, Donkin, Brown, Heathcote & Marley (2009) *Stimulus-Specific Learning: Disrupting the Bow Effect in Absolute Identification*

These papers discuss the broad empirical findings from a series of experiments related to absolute identification. Many experiments described in this thesis concentrate on the first paper, *Increasing Capacity: Practice Effects in Absolute Identification* (Dodds et al. 2011a), where we discuss a series of experiments that demonstrate learning effects in absolute identification. Dodds et al. (2011a) conduct a series of experiments that examine learning effects for a range of stimulus types, concluding that of several types of stimuli tested, many were able to exhibit strong learning effects: tones varying in intensity were the only stimulus modality that presented performance patterns consistent with prior literature. Dodds et al. (2011a) examine this phenomenon and promote the further investigation into why this contrast might be found.

Chapter Two examines another popular and robust absolute identification phenomenon – the bow effect. Dodds et al. (2011b) manipulate stimulus presentation probability to demonstrate that the learning effect shown in Dodds et al. (2011a) is stimulus-specific. This has important consequences both for experimental purposes, and also model design: learning effects must be taken into account, but also introduced on a per-stimulus basis. In regards to experimental design – Dodds et al. (2011b) show that additional presentations of specific stimuli leads to improved performance for these stimuli – this means that any design in which additional presentations are naturally

introduced (e.g. a manipulation of set size on a within-subjects basis), should be avoided.

Section Two: Theoretical Implications

Section Two consists of four papers:

1. Brown, S.D., Marley, A.A.J., Dodds, P., & Heathcote, A.J. (2009). Purely relative models cannot provide a general account of absolute identification. *Psychonomic Bulletin & Review, 16, p.583-593*
2. Dodds, Brown, Marley & Heathcote (submitted). *Absolute Identification and Absolute Production*
3. Dodds, P., Donkin, D., Brown, S.D., Heathcote, A. (2010) Multidimensional scaling methods for absolute identification data In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Portland, OR: Cognitive Science Society
4. Dodds, Rae & Brown (submitted). *Perhaps Unidimensional is not Unidimensional*

This section introduces theoretical implications of the empirical results outlined in Section One. This broadly includes discussion on different types of AI models and how we might distinguish between them, and also an investigation into alternative explanations for the learning effects described in Chapter One.

Chapter Three provides a discussion of two general types of absolute identification models: absolute and relative models. We distinguish between the two broad categories of AI models by using unequally spaced absolute identification data. This paper provides insight into current model developments.

Chapter Four in this thesis attempts to find new ways of extending on current models of AI performance. In this chapter, we reconcile two similar tasks: absolute

identification and *absolute production* (AP). Absolute production is similar in concept to absolute identification, but requires the participant to *draw* the line in question, rather than recalling its label. The continuous nature of responses intrinsic to AP tasks has advantages over AI however, as it gives insight into the way models might replicate the internal representation of stimuli. We extend on previous work (Zotov, Shaki & Marley, 2010) and show that AP and AI share similar benchmark data characteristics. Further to this, we discuss some preliminary adaptations to a model of AI performance SAMBA (Brown et al., 2008), to replicate AP data patterns.

Chapters Five and Six begin an investigation into an alternative explanation for the effects discussed in Chapters One and Two (Dodds et al. 2011a & Dodds et al., 2011b). Several chapters in this thesis allude to the need to update current models of AI performance to accommodation learning effects – in the following two chapters however, we discuss the possibility that the learning effects were an artefact of the psychological perception of stimuli. In Chapter Five we use Multidimensional Scaling techniques to examine the underlying structure in our stimuli. While results were consistent with a single underlying dimension, a lack of framework for inference is of some concern. We extend upon this in Chapter Six (Dodds, Rae and Brown, submitted) *Perhaps Unidimensional is Not Unidimensional*. We discuss the issue of underlying structures again, but instead use a Bayesian algorithm developed by Kemp and Tenenbaum (2008), that examines underlying structures in similarity and distance data.

Absolute Identification provides a number of interesting and intricate phenomena – in this thesis we study two of these in detail, and describe how existing beliefs about AI performance may be misguided. We describe existing AI performance models and methods of distinguishing between them, before returning to examine the possibility of other explanations for the empirical findings of Chapters One and Two.

Taken together, the chapters in this thesis combine to make a coherent story of the development of our understanding of empirical AI phenomena and challenge our existing assumptions of AI performance.

References

Brown, S.D., Marley, A.A.J., Dodds, P., & Heathcote, A.J. (2009). Purely relative models cannot provide a general account of absolute identification. *Psychonomic Bulletin & Review*, 16, p.583-593

Brown, S.D., Marley, A.A.J., Donkin, C. & Heathcote, A.J. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review, 115*(2), 396-425

Cox, T. F. & Cox, M. A. A. (1994). *Multidimensional Scaling*. London: Chapman and Hall.

Cox, T. F. & Cox, M. A. A. (2001). *Multidimensional Scaling*. London: Chapman and Hall.

Dodds, P., Brown, S. D., Zotov, V., Shaki, S., Marley, A. A. J. & Heathcote, A. (2011). *Absolute production and absolute identification*. Manuscript submitted for publication.

Dodds, P., Rae, B., & Brown, S. D. (2011). *Perhaps unidimensional is not unidimensional*. Manuscript submitted for publication

Dodds, P., Donkin, C., Brown, S. D. & Heathcote, A. (2011a) Increasing Capacity: Practice Effects in Absolute Identification. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 37(2), 477-492.

Dodds, P., Donkin, C., Brown, S. D., Heathcote, A., Marley, A. A. J. (2011b) Stimulus-Specific Learning: Disrupting the Bow Effect in Absolute Identification. *Attention, Perception & Psychophysics,73*(6), 1977-1986

Dodds, P., Donkin, D., Brown, S.D., Heathcote, A. (2010) Multidimensional scaling methods for absolute identification data In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Portland, OR: Cognitive Science Society

Garner, W. R. (1953). An informational analysis of absolute judgments of loudness. *Journal of Experimental Psychology, 46*(5), 373-380.

Hartman, E. B. (1954). The influence of practice and pitch distance between tones on the absolute identification of pitch. *The American Journal of Psychology, 67*(1), 1-14.

Hsieh, I., and Saberi, K. (2007). Temporal integration in absolute identification of musical pitch. Hearing Research, 233, 108-116

Kemp, C. & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105,* 10,687-10,692.

Kent, C. & Lamberts, K. (2005). An exemplar account of the bow and set-size effects in absolute identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 289 –305.

Lacouture, Y. (1997). Bow, range and sequential effects in absolute identification: A response-time analysis. *Psychological Research, 60*(3), 121-133.

Lacouture, Y., & Marley, A. A. J. (2004). Choice and response time processes in the identification and categorization of unidimensional stimuli. *Perception and Psychophysics, 66*(7), 1206-1226.

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.

Laming, D. (1986). *Sensory Analysis*. London: Academic Press.

Lockhead, G. R., & Hinson, J. (1986). Range and sequence effects in judgment. *Perception and Psychophysics, 40*(1), 53-61.

Marley, A. A. J., & Cook, V. T. (1984). A fixed rehearsal capacity interpretation of limits on absolute identification performance. *British Journal of Mathematical and Statistical Psychology, 37,* 136-151.

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

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.

Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America, 24*(6), 745-749.

Pollack, I. (1953). The information of elementary auditory displays: II. *Journal of the Acoustical Society of America*, 25, 765–769.

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., & Nosofsky, R. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review, 101,* 357-361.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112*(4), 881-911.

Teghtsoonian, R. (1971). On the exponents in Stevens' Law and the constant in Ekman's Law. *Psychological Review, 78*(1), 71-80.

Weber, D. L., Green, D. M., & Luce, R. D. (1977). Effects of practice and distribution of auditory signals on absolute identification. *Perception and Psychophysics, 22*(3), 223-231

Zotov, V., Shaki, S., & Marley, A. A. (2010). Absolute production as a - possible - method to externalize the properties of context dependent internal representations. In A. V. Bastianelli (Ed.), *Fechner Day 2010. Proceedings of the 26th Annual Meeting of the International Society for Psychophysics* (pp. 203-209). Padua, Italy: The International Society for Psychophysics.

SECTION ONE: EMPIRICAL RESULTS

Included Papers

Dodds, P., Donkin, C., Brown, S. D. & Heathcote, A. (2011) Increasing Capacity: Practice Effects in Absolute Identification. *Journal of Experimental Psychology: Learning, Memory & Cognition,* 37(2), 477-492.

Dodds, P., Donkin, C., Brown, S. D., Heathcote, A., Marley, A. A. J. (2011) Stimulus-Specific Learning: Disrupting the Bow Effect in Absolute Identification. *Attention, Perception & Psychophysics*, 73(6), 1977-1986

Chapter One

Increasing Capacity: Practice Effects in Absolute Identification

Pennie Dodds[1], Christopher Donkin[2],
Scott D. Brown[1] & Andrew Heathcote[1]

[1]School of Psychology, University of Newcastle, Callaghan NSW 2308, Australia
[2] Department of Psychology & Brain Sciences, University of Indiana, Indiana

Counts
Abstract: 157
Body: 9827
Figures: 12
Tables: 4

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

In most of the long history of the study of absolute identification – since Miller's (1956) seminal paper – a severe limit has been observed on performance, and this limit has resisted improvement even by extensive practice. In a startling result, Rouder, Morey, Cowan and Pfaltz (2004) found substantially improved performance with practice in the absolute identification of line lengths, albeit only for three participants and in a somewhat atypical paradigm. We investigated the limits of this effect and found that it also occurs in more typical paradigms, is not limited to a few virtuoso participants nor due to relative judgement strategies, and that it generalizes to some (e.g., line inclination and tone frequency) but not other (e.g., tone loudness) dimensions. Apart from differences between dimensions, we also observed two unusual aspects of improvement with practice: a positive correlation between initial performance and the effect of practice; and a large reduction in a characteristic trial-to-trial decision bias with practice.

Human memory for complex items such as names, letters and faces is seemingly infinite. We are able to memorise a great number of these items across our lifespan, or even in a one-hour experimental task, with relative ease. For decades, however, a single, simple task has provided an exception to this rule: absolute identification. Absolute identification (also called "dead reckoning" by Miller, 1956) is the task of identifying which stimulus has been shown out of a set of stimuli that vary on only one physical dimension. For example, a participant might be given a set of *n* lines varying in length, or tones varying in intensity, labelled from 1 through to *n*. On each trial of an absolute identification task, the participant is then presented with one of these stimuli and asked to recall its label. Empirical research into absolute identification has a long history, with Miller's summary of early work identifying a surprisingly small capacity limitation – people are generally unable to accurately identify more than 7±2 stimuli in an absolute identification task. Miller noticed a similar limitation in short term memory performance, and the two limitations have often been treated as manifestations of a single phenomenon; that is, absolute identification performance is limited precisely because it relies on short term memory capacity, and so the study of absolute identification is interesting (in part) because of what it reveals about short term memory.

For decades, the received view has been that this capacity limitation is unaffected by manipulations that are otherwise very powerful. For example, Miller (1956) showed that the capacity limit was about the same for the absolute identification of many different kinds of stimuli including line length, taste, brightness, hue and loudness. There are many other stimulus manipulations which one might assume would improve performance, but these have all been demonstrated to have little or no effect on the capacity limitation (e.g., increasing the number, or separation of the stimuli: Pollack,

1952; Garner, 1953). Possibly the most intriguing finding is that the capacity limit is highly resistant to practice. For example, Garner's participants engaged in up to 12,000 judgements in a single condition, yet even at the end of the experiment they were still limited to identifying the equivalent of three or four stimuli correctly. Weber, Green and Luce (1977) had participants complete 12,000 trials identifying six white noise signals of varying loudness and found an improvement in response accuracy of just 6%. Final performance for these participants was well below ceiling, despite the large amount of practice, monetary incentives, and the apparently easy task of identifying just six separate levels of loudness. Hartman's (1954) participants also practiced over an eight-week period, and while they demonstrated substantial improvement, their best performance level was still well within Miller's limit: equivalent to the perfect identification of only five stimuli. Such results have established a truism about absolute identification – there is a severe limitation in human ability to identify unidimensional stimuli, and this limit is largely unaffected by practice.

In a departure from previous findings, Rouder, Morey, Cowan and Pfaltz (2004) demonstrated that substantial learning *is* possible in an absolute identification task. In particular, three participants showed large improvements in the identification of line length with practice. One participant, after 11,100 trials of practice, was able to correctly identify almost 20 different line lengths. The other two participants, with 18,740 and 5,040 trials of practice, were able to correctly identify about 13 lines. It is not clear what caused the difference between Rouder et al.'s result and earlier studies. For example, learning may have been improved because Rouder et al.'s participants were given chances to correct incorrect responses. Perhaps also the large improvement with practice is unique to the absolute identification of line lengths, and would not have been observed with, for example, the identification of tones of varying loudness

(consistent with Garner's 1953, results). This explanation seems especially attractive because, although line lengths have been used occasionally in the field (e.g., Thorndike, 1932; Lacouture, 1997; Rouder, 2001; Kent & Lamberts, 2005), previous demonstrations of the null effect of practice have mainly used tones varying in intensity. Another important difference between Rouder et al.'s methods and earlier work was the use of considerably larger stimulus sets (up to 30 different lines, rather than the more typical 8-12 stimuli).

These findings are particularly interesting because they might shed light on the deeper issue: although we seem to have practically unlimited memory for items such as faces and names, unidimensional stimuli have been highlighted as the exception to this rule. Through a series of experiments, we investigate whether unidimensional stimuli truly represent an exception to this short term memory limitation, and what characteristics of such stimuli affect overall learning. As well as identifying which kinds of stimulus sets support learning and which do not, we also investigate the mechanisms underlying improvement with practice. For example, participants may learn to increase the capacity of their short term memory, and so are better able to pair stimuli with their to-be-recalled labels. Alternatively, they may learn to avoid some of the well-documented decision biases that pervade absolute identification (the "sequential effects", see, e.g., Stewart, Brown & Chater, 2005). To foreshadow our conclusions, although our data strongly suggest improvements of the latter variety, model-based analyses implicate both kinds of learning.

Experiment 1

We begin our investigations by examining whether any of the atypical design features used in Rouder et al's (2004) study contributed to the large learning effect. The

most novel aspect of Rouder et al.'s design was their response technique, where participants were given two opportunities to respond instead of the standard single response. If the participant made an incorrect response they were allowed a second attempt. If they were incorrect on their second attempt, the correct answer was displayed. In Experiment 1a, we aimed to replicate Rouder et al.'s findings of significant learning with their response method – to ensure that Rouder et al. did not simply have an exceptional sample of participants. In Experiment 1b we investigated whether learning persists with a standard response method in a paradigm that is otherwise identical.

Method

Participants. Six participants took part in Experiment 1a, and a different six in Experiment 1b. Each was reimbursed $15 per session, and unless otherwise stated, this was the case for all following experiments, with six new participants recruited for each.

Stimuli. The stimuli were 30 lines of varying length, increasing in size according to a power function with an exponent of 3.5 (see Rouder et al., 2004, and see Table 1). Stimuli were presented in black on a white background, using a 21inch CRT monitor set at a resolution of 1152 x 864 pixels. Each pixel measured .39mm wide by .35 high. Images were positioned in the centre of the screen, with 22 x 22 pixel variation in position from trial to trial to discourage participants from using the edge of the monitor as a size cue.

Table 2. Line lengths in pixels for Experiment 1a, 1b and 5a.

Experiment 1a and 1b									
9	12	14	17	20	23	27	31	36	41
47	53	60	67	76	84	94	104	115	127
140	153	168	183	199	217	235	255	276	298
Experiment 5a									
15	18	22	27	33	41	50	61	74	90
110	134	164	200	244	298				

Procedure. In a brief study phase at the beginning of each session, participants were given each stimulus one at a time, labelled with a corresponding number, from 1 through to 30. In order to proceed through the study phase, the participant had to select the number on screen that corresponded to the numerical label. For example "This is line number 1. Press 1 to continue". During each trial in the main phase of the experiment, one stimulus was randomly selected and presented, and the participant was asked to respond with the numerical label that was attached to the stimulus in the study phase. Instructions given to the participants emphasised response accuracy over response time. This decision was made in light of our primary interest in how accurately participants could perform the task. Responses were made using the mouse to click buttons arranged onscreen in increasing numerical order. Three columns of 10 buttons were arranged on the left hand side of the screen and these remained onscreen throughout the experiment.

The only difference between Experiments 1a and 1b was the number of response opportunities per trial. In Experiment 1a, we replicated Rouder et al.'s two-response method. If participants were incorrect on the first response, they were given a second response opportunity. If they were incorrect again, the correct answer was displayed for 500ms. Whenever a correct response was recorded, the text "Correct" was displayed and the trial ended. In Experiment 1b, we used the traditional one-response absolute

identification feedback system, where participants were only given one opportunity to respond. If they were incorrect, the correct answer was displayed for 500ms. If they were correct, the text "Correct" was displayed. The stimulus always remained on screen until the final feedback was provided.

Participants took part in ten sessions, each of approximately one hour. Sessions were conducted on (mostly) consecutive days. The first three sessions consisted of 6 blocks of 90 trials, while the remaining seven sessions consisted of seven blocks. This resulted in 201 presentations per stimulus per participant. A minimum one minute break was enforced between blocks.

Results

Analyses were conducted on the first response only, to allow more valid comparison of Experiments 1a and 1b (cf. Rouder et al., 2004). Rather than focus only on response accuracy, we also calculated the amount of information transmitted from the stimulus to the response. Due to a historical focus on information-theoretic accounts of absolute identification performance (e.g., Hake & Garner, 1951), information transfer has become a standard descriptor for performance, and it is also particularly useful when comparing different stimulus set sizes (see Shannon, 1948; also Pollack, 1952; Garner, 1953). Information transfer attempts to measure how much uncertainty in the identity of the stimulus is removed by considering the observer's response. The amount of information transmitted from the stimulus to the response is measured in "bits", and 2^{bits} can be interpreted as the equivalent number of stimuli that could be perfectly identified (e.g., 3 bits of transmitted information corresponds to perfectly accurate identification of 2^3=8 stimuli).

We calculated the amount of transmitted information separately for each participant and each practice session. We quantified the amount of improvement

induced by practice using the minimum and maximum of these values. Note that these extrema did not always occur in the first or last sessions, but analyses based on the first and last sessions yield similar results. We employed the minimum and maximum values because of a trend for participants to lose some motivation in the final session of the experiment – across all experiments and all participants, the proportion of increases in performance from one session to the next was 72%, but this was significantly lower for the final session, at 41% (χ^2=4.2, p<.05).

Improvement with practice was apparent in both Experiment 1a and 1b, as illustrated in Figure 1. In Experiment 1a, where participants were given two response opportunities, the percentage of correct responses (i.e., accuracy) improved from 23% to 49%, compared to a chance performance level of just 3.3%. In terms of information transmission, this corresponds to an average improvement of 0.83 bits from 2.40 to 3.23. This meant that average maximum performance (across subjects) was equivalent to the perfect identification of approximately 9.4 stimuli. A one-way (session) repeated-measures ANOVA with a Greenhouse-Geisser sphericity correction confirmed that these effects were highly reliable for both accuracy (F(1.36,6.63) = 14.84, p=.005) and information transfer (F(1.37,6.72) = 26.17, p=.001).

In Experiment 1b, where participants were not offered a second response opportunity, we observed almost identical results. There was again highly significant improvement across the ten sessions as measured by accuracy (F(1.38,6.9)=24.58, p =.001) and information transfer (F(1.41,7.03)=31.41, p < .001). Accuracy improved from 22% to 46%, an average increase of 24%. Information transfer also increased by an average of 0.83 bits, from 2.28 to 3.11 bits, which is equivalent to the perfect identification of approximately 8.66 stimuli. Even though the subject-average peak performance was greater in Experiment 1a than 1b, this difference was not reliable

according to an independent-samples t-test ($p = 0.71$). Naturally, the statistical power of this test to identify between-experiment differences is very limited, due to the small sample size. Nevertheless, we note that several participants in Experiment 1b showed larger practice effects than some participants in Experiment 1a, making it seem unlikely that the two-response feedback procedure caused any large differences.

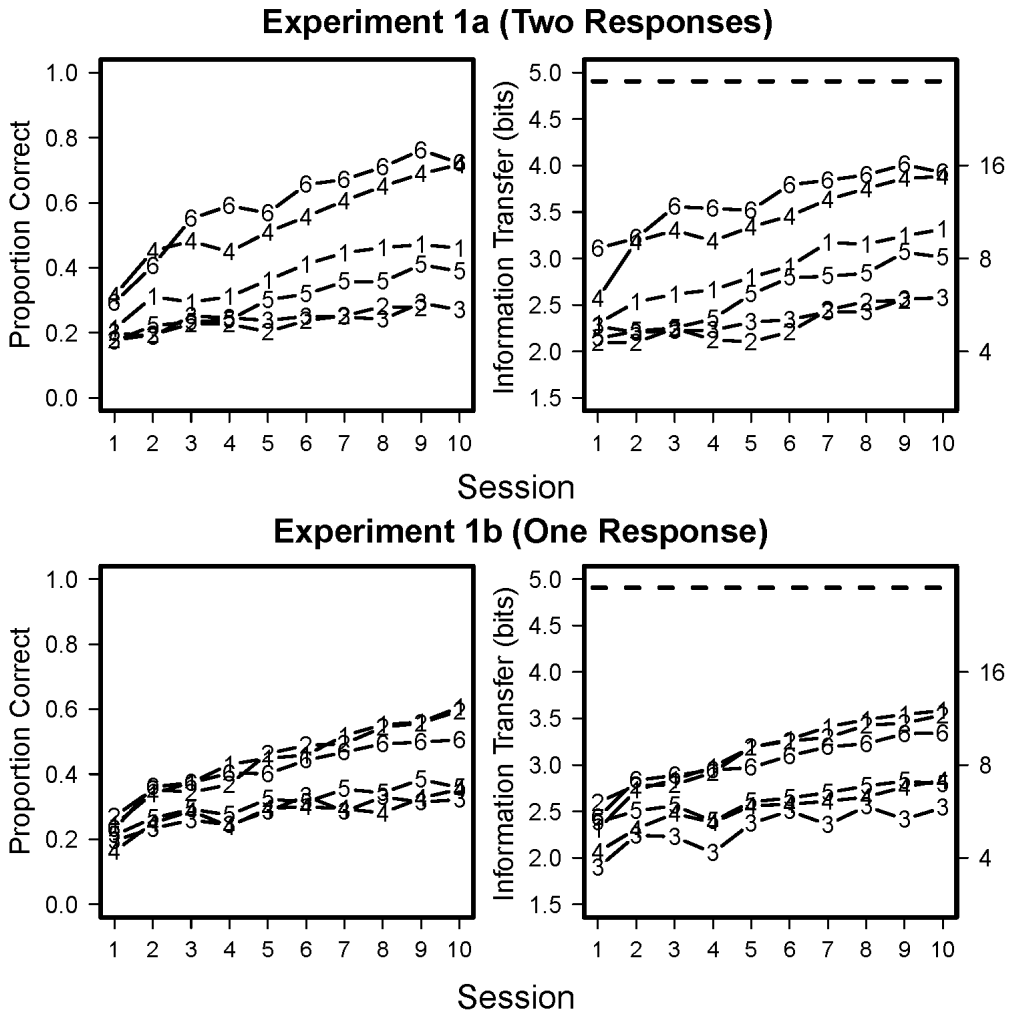


Figure 1. Proportion correct, and information transfer as functions of session for Experiments 1a and 1b (30 lines varying in length). The right hand axis on the information transfer graph shows the equivalent number of stimuli that were perfectly identified (2^{bits}). The dashed line indicates perfect performance: $\log_2$(*number of stimuli*)

Discussion

Participants in both Experiments 1a and 1b demonstrated significant improvements in performance, suggesting that the two response method was not responsible for the amount of learning observed. These experiments also confirm that Rouder et al.'s (2004) results were not due to unusual, virtuoso, participants. In both experiments, performance improved by more than 20% (~0.8 bits) after about 6,000 practice trials, and three of the six participants in each experiment exceeded Miller's (1956) bound of 7±2 stimuli.

A possible explanation for the improvement with practice at length judgment might invoke the development of a relative (or "referent") judgement strategy, rather than by improving absolute identification processes themselves. That is, participants judging line lengths might be able to compare the lines to external magnitude cues, such as the edges of the computer monitor or the response buttons that appeared on screen. In Experiment 1, and in Rouder et al.'s (2004) design, these strategies were discouraged by jittering the absolute location of the stimuli on screen from trial to trial. Nevertheless, some small amount of relatively imprecise information might still be gained by comparisons against visible reference points, and it might be that this information alone supports improvement with practice. In Experiment 2, we investigate this explanation, and also the idea that large effects of practice are only possible with large stimulus set sizes.

Experiment 2

Experiment 2 was conducted in a dark room. The edges of the monitor were obscured from view, and response buttons varied in size from trial to trial. Response buttons were never on screen at the same time as stimuli. We also included a second

condition, Experiment 2b, in which only half of the stimuli were presented, to determine whether the learning effect was due to the large amount of available information.

Method

Participants. Ten participants took part in this experiment, five in Experiment 2a and five in Experiment 2b. They were reimbursed in a similar fashion as the participants in the first experiment.

Stimuli. Each stimulus consisted of a pair of white dots on a black background, horizontally separated by intervals that were of the same lengths as the lines in Experiment 1.

Procedure. There were two conditions defined by the number of stimuli: Experiment 2a had 30 stimuli while Experiment 2b had only 15 stimuli. The stimuli in Experiment 2b were all of the odd-numbered stimuli from Experiment 2a, and so the pair-wise stimulus separation was twice as large in Experiment 2b as in Experiment 2a. We could have kept stimulus separation equal in the two experiments (e.g., by presenting only stimuli #1-#15 in Experiment 2b) but that would have instead confounded stimulus range with set size. We acknowledge that both solutions to this problem (either confounding stimulus range, or stimulus separation) are imperfect, but we chose the latter solution because performance is mostly unaffected by changes in stimulus for widely spaced stimulus sets (e.g., see Braida & Durlach, 1972, but also see Stewart et al. 2005 and Lacouture, 1997, for alternative findings).

The experiment was conducted in a dark room, where the only light was that emitted by the computer monitor (which was made as dark as possible). The edges of the computer monitor were obscured by black cardboard. To ensure that the response buttons could not be used as a cue for relative comparison with the size of the stimuli, two precautions were taken: the buttons were never present on screen at the same time

as the stimuli, and the size of the response buttons varied from trial to trial. That is, when a stimulus was presented the buttons were removed from the screen until the participant clicked a mouse button to indicate they were ready to respond, then the stimulus was removed and the response buttons were displayed. Participants took part in ten sessions, each about an hour in length. Each session consisted of 6 blocks of 90 trials, resulting in 180 presentations per stimulus in Experiment 2a, and 360 presentations per stimulus for Experiment 2b.

Results

Performance increased significantly across the ten sessions in both conditions. Participants in the 30 stimulus condition (Experiment 2a) increased their accuracy from 25% to 50%; an average improvement of 25% ($F(1.76,7.04)=29.74$, $p<.001$). Information transfer also increased by 0.93 bits across the ten sessions from 2.44 to 3.36 bits ($F(1.54,6.15)=60.88$, $p<.001$), so the average maximum performance was equivalent to perfect identification of approximately 10.3 stimuli.

Similarly, participants in the 15 stimulus condition also demonstrated highly significant improvements in both accuracy ($F(1.56,6.24)=22.25$, $p=.002$) and information transfer ($F(1.71,6.86)=19.43$, $p=.002$). Participants improved 33% from an average accuracy of 48% to 81%, and information transfer improved 1.08 bits from 2.07 to 3.15. Average maximum information transmitted was equivalent to identification of 8.85 stimuli. Figure 2 provides a comparison of the individual participant results in Experiments 2a and 2b.

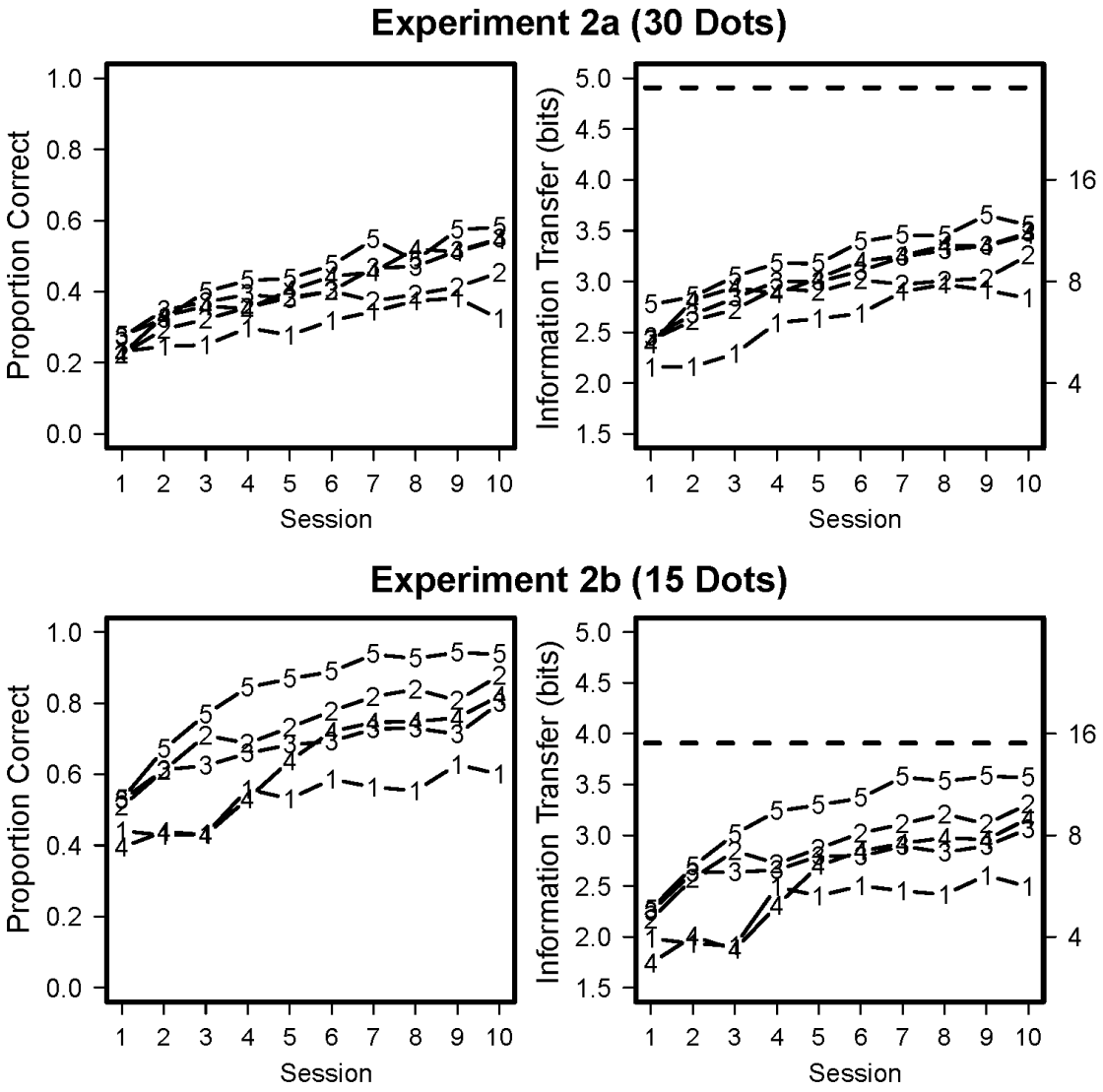


Figure 2. Proportion correct and information transfer as a function of session for Experiments 2a (30 stimuli) and Experiment 2b (15 stimuli), using dots varying in separation. The right hand axis on the information transfer graph is the equivalent number of stimuli that were perfectly identified (or 2^{bits}). The dashed line indicates the maximum amount of information transfer possible, $\log_2$ (*number of stimuli*)

The average maximum amount of information transfer reported for the 30 stimulus condition in the current experiment (M = 3.30) was not reliably different from that found in Experiment 1a (M = 3.25, p = 0.86) and Experiment 1b (M = 3.15, p = 0.50) [1]. This suggests that any external cues were not responsible for the learning effect in Experiment 1.

We also observed that, although one participant in the 15 stimuli condition reached almost perfect performance (94.3% accuracy), the average maximum information transmission for the 15 stimulus condition (M = 3.11) was not reliably different from the 30 stimulus conditions in Experiments 1a (p = 0.64) and 1b (p = 0.85) or Experiment 2a (p = 0.38). This suggests that maximum performance, in terms of information transmission, does not vary with set size.

Discussion

Participants in Experiment 2 demonstrated significant improvements in performance, and similar information transmission limits and amounts of learning to participants in Experiment 1. Once again, half of the participants demonstrated maximum information transfer rates that exceeded Miller's (1956) 7±2 bound. The similarity in results between Experiments 1 and 2 suggests that external cues were not responsible for the learning effect, and that the amount of available information does not determine the extent of learning, at least as long as performance is below ceiling.

Experiment 3

So far, substantial learning in absolute identification has only been demonstrated using line lengths, with null (or small) effects observed tones varying in intensity or

[1] Estimates of transmitted information are inflated by small sample sizes (see Norwich, Wong & Sagi, 1998). For this reason, for comparisons between experiments we always calculated information transfer using data divided into fairly long (540 trial) segments.

frequency. Unfortunately, this difference between stimulus modality has always been confounded with a procedural change: tones were only made available to participants for a short period of time (typically, one second), whereas lines were made available for as long as participants wish. There is some evidence to suggest that stimulus presentation time can influence performance. For example, Miller (1956) cites unpublished research by Pollack that found significantly smaller information transmission for lines varying in length when presented for short periods of time (2.6 bits), compared to longer presentation times (3.0 bits). Ward and Lockhead (1971) also found lower information transfer for a presentation time of 8ms (0.19 bits) compared to presentation for 200ms (1.07 bits), although they simultaneously manipulated luminance. In an attempt to examine whether unlimited presentation time may have encouraged the learning effect, in Experiment 3 line stimuli were masked after one second – in line with usual practice for auditory stimuli.

Method

Six participants took part in Experiment 3, using the same procedure as used for Experiment 2a, with the exception of presentation time. Stimuli were left on the computer monitor for only one second, after which they were covered by a mask consisting of white dots scattered randomly over a rectangle of dimensions 1024 x 684 pixels. The white dots were equal in size and luminance to the white dots used to construct the line stimuli. The mask remained on the screen until the participant had responded.

Results

The results are very similar to those in Experiment 1 and 2a, where participants were given stimuli with unlimited presentation time. Accuracy increased from 24% to 42% ($F(1.56,7.79)=18.2$, $p=.002$) and information transfer increased from 2.39 to 3.06

bits ($F(1.54,7.7)=22.27$, $p<.001$). The average maximum performance (3.06 bits) was equivalent to the perfect identification of 8.32 stimuli (see

Figure ***3***). Although average learning was slightly smaller, one participant still learned to identify more than Miller's (1956) upper limit of 9 stimuli. In addition, the maximum amount of information transmitted with masked stimuli was only about 5% smaller than the average amount for Experiments 1-2, and this difference was not statistically reliable ($p = .30$). The similar pattern in results for the current experiment suggests that long presentation times were not required for the learning effect.

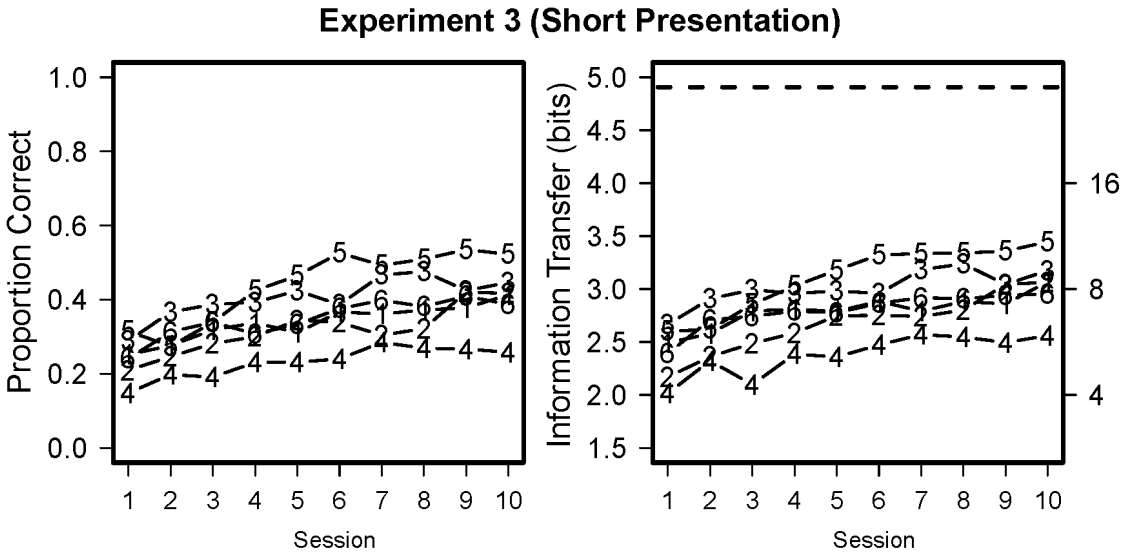


Figure 3. Proportion correct and information transfer as a function of session for Experiment 3 (dots varying in separation). The right hand axis on the information transfer graph is the equivalent number of stimuli that were perfectly identified (or 2^{bits}). The dashed line indicates the maximum amount of information transfer possible, $\log_2$ (*number of stimuli*)

Discussion

Even with limited stimulus presentation times there was significant improvement in performance with practice, and these results were comparable to earlier experiments with unlimited viewing time. The slightly lower performance reached with short presentation times was not significantly different from previous experiments. The direction of the effect, however, suggests that limited presentation time, or perhaps the addition of the mask, may have limited the amount that participants could improve via practice, even if our sample sizes provided insufficient statistical power to detect a reliable difference. Most importantly, however, participants did still manage to significantly improve their performance, and the amount of improvement was not much smaller than Experiments 1 and 2.

Experiment 4

Experiments 1-3 established that learning was not due to the more unusual aspects of Rouder et al.'s (2004) methods, nor to external cues, and that it was not much attenuated by a limitation on stimulus presentation time. We now test whether the strong practice effects we have observed are specific to visual lengths: lines varying in length or dots varying in separation. In Experiment 4 we investigated whether learning is possible with lines varying in angle of inclination.

Method

The methods were identical to Experiment 3 except that the stimuli were 30 lines whose angle of inclination varied from 1.5^{o} to 89.5^{o} in increments of 3^{o}. The lines were 12 x 210 pixel rectangles, and they were blurred by applying Gaussian kernel with a 7 pixel standard deviation (to prevent the use of pixel aliasing as a cue for angle). Stimuli were white on a black background, and were positioned within a square 300 x

300 pixels in size. To help prevent the use of both horizontal and vertical cues for angle judgments, lines were rotated around a central pivot point, and the screen position of that pivot point was varied randomly from trial to trial within a 22x22 pixel region. Each stimulus was presented for one second. If no response was made within one second, a mask was displayed and remained onscreen until the participant had made their response. Masks were 1024x1024 pixel squares containing a series of randomly positioned and randomly oriented lines of the same sort as the stimuli.

Results

Results were very similar to the prior experiments. Figure 4 shows that learning was highly significant across the ten sessions (accuracy: $F(2.05,10.3)=23.6$, $p<.001$; information transfer: $F(2.17,10.8)=26.37$, $p<.001$). Average accuracy improved by 22% from an initial value of 24% to 46%. Average information transfer also improved 0.81 bits from 2.37 to 3.18 bits, which made average maximum performance equivalent to the perfect identification of about 9.05 stimuli. Three of the six participants exceeded Miller's (1956) 7±2 limit after ten sessions practice. Neither initial performance, performance improvement, nor maximum performance were significantly different from Experiment 1a ($p = 0.93$, $p = 0.68$ and $p = 0.78$), nor Experiment 3, where presentation time was identical ($p = .87$, $p = .23$, $p = .37$).

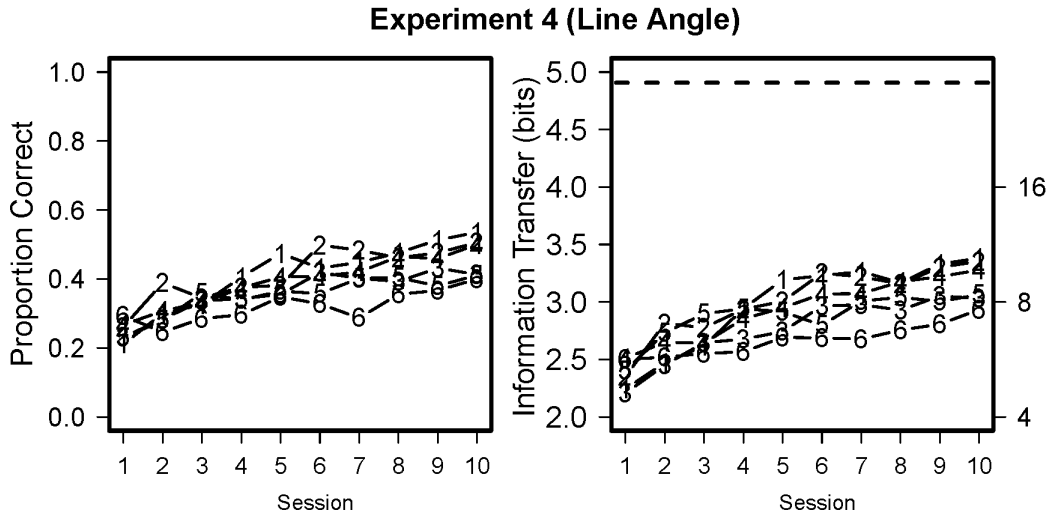


Figure 4. Proportion correct and information transfer as a function of session for Experiment 4 (angle of inclination). The right hand axis on the information transfer graph is the equivalent number of stimuli that were perfectly identified (or 2^{bits}). The dashed line indicates the maximum amount of information possible, $\log_2$ (*number of stimuli*)

Discussion

Participants in Experiment 4 demonstrated significant improvement in performance across the ten sessions, similar to that observed in the previous experiments. This result suggests that the learning effect may generalise to visual stimuli other than line length. We further explore whether learning occurs with other stimulus types in Experiment 5.

Experiment 5

Clearly people are able to substantially improve their performance in an absolute identification task when given significant practice, and we have shown that this learning is not specific to distance or length judgements. However, so far our investigation has

been limited to visual stimuli only. Miller (1956) noted visual modalities led to slightly greater information transmission (hence the "plus or minus" in his 7±2). More recent research has also suggested differences – Lacouture and Lacerte (1997) found better performance for lines varying in length than tones varying in intensity. This is particularly interesting here, because most previous studies showing no effect of practice used tones varying in intensity. In Experiment 5, we compared the effect of practice using tones varying in intensity and lines varying in length in order to determine whether it is modality which differentiates our (and Rouder et al.'s, 2004) findings from others.

Method

Methods were identical to Experiment 1a except that stimuli were either 16 lines varying in length (Experiment 5a), or 16 tones varying in intensity (Experiment 5b). Though we aimed to replicate our earlier experiments exactly, we found that we were limited to the use of just 16 (rather than 30) tones. This limit was identified through pilot testing, with naïve participants. Those tests showed that participants were able to make perfectly accurate discrimination judgments (lower/higher) between sequentially presented stimuli separated by a one second pause, when the stimulus difference was 3dB. This stimulus separation implied that range restrictions imposed by ethical considerations and the audio equipment itself limited us to 16 tones in total. We therefore also ran Experiment 5a using 16 lines (see Table 1 for line lengths in pixels) for ease of comparison of results with Experiment 5b. The 16 auditory stimuli were pure 1000Hz tones, ranging from 61dB to 106dB, in 3dB increments. Loudness was measured using a Brüel and Kjaer artificial ear (model 4152) and sound level meter (Brüel and Kjaer, model 2260), equipped with a condenser microphone (Brüel and Kjaer, model 4144). Tones were played for one second each, and were presented via

Sony headphones (model DR-220). For each of the ten sessions, participants in the lines condition completed 7 blocks of 80 trials, and those in the tones condition completed 7 blocks of 90 trials.

Results

Participants in the 16 line condition performed similarly to participants in previously reported line experiments: average accuracy significantly increased across the ten sessions, from 49% to 78% ($F(1.82,9.11)=23.43$, $p<.001$), and average information transmission increased significantly from 2.34 to 3.15 bits ($F(1.94,9.7)=20.92$, $p<.001$). Average maximum performance was equivalent to identification of 8.86 stimuli, and two of the six participants exceeded Miller's (1956) 7±2 limit. The maximum information limit reached in the 16 lines condition was not significantly different from those in Experiment 1a ($p = 0.58$), or from the results in Experiment 2b with a similar set size ($p=.95$).

Participants in the tone intensity condition, on the other hand, failed to exhibit the substantial learning found in all other experiments (see Figure 5). Participants given 16 tones of varying intensity had a lower average initial accuracy (31%) and only improved on average by 12%. Similarly, information transfer only increased on average by approximately 0.46 bits, from 1.49 to 1.95 bits, meaning that maximum performance was equivalent to the perfect identification of only 3.86 stimuli, and no participant exceeded Miller's (1956) 7±2 limit. In fact, all participants identified less than 5 stimuli perfectly correctly. However, the small effect of practice was statistically reliable (accuracy: $F(2.91,14.2)=4.8$, $p=.02$; information transfer: $F(2.08,10.2)=4.9$, $p=.03$).

Although the improvement for both modalities was reliable, loudness in Experiment 5b showed a significantly lower information transfer limit than lines in Experiment 5a ($M_{tones}=1.96$, $M_{lines}=3.09$; $t(9.79)=9.51$, $p<.001$). We also observed

reliably smaller maximum information transmission ($t(9.47)=3.17$, $p=.01$) for tones (M_{diff} = 0.42 bits) than lines (M_{diff} = .74 bits). The findings for loudness were consistent with previous findings of a low channel limit (e.g., Miller, 1956, Garner, 1953, Pollack, 1952) and little improvement with substantial practice (e.g., Weber, Green and Luce, 1977). It is also interesting to note that, in contrast to the slow increase in performance for tones varying in intensity, there was a much faster increase in performance for line length (Experiment 5a). This was particularly noticeable between Sessions 1 and 2, where participants in Experiment 5a improved their performance significantly more ($M=.16$) compared to the corresponding difference between Session 1 and 2 in Experiment 5b ($M=.02$; $t(6.8)=8.85$, $p<.001$). This suggests that participants in Experiment 5a (line stimuli) learned quickly to some upper limit, unlike participants in Experiment 5b.

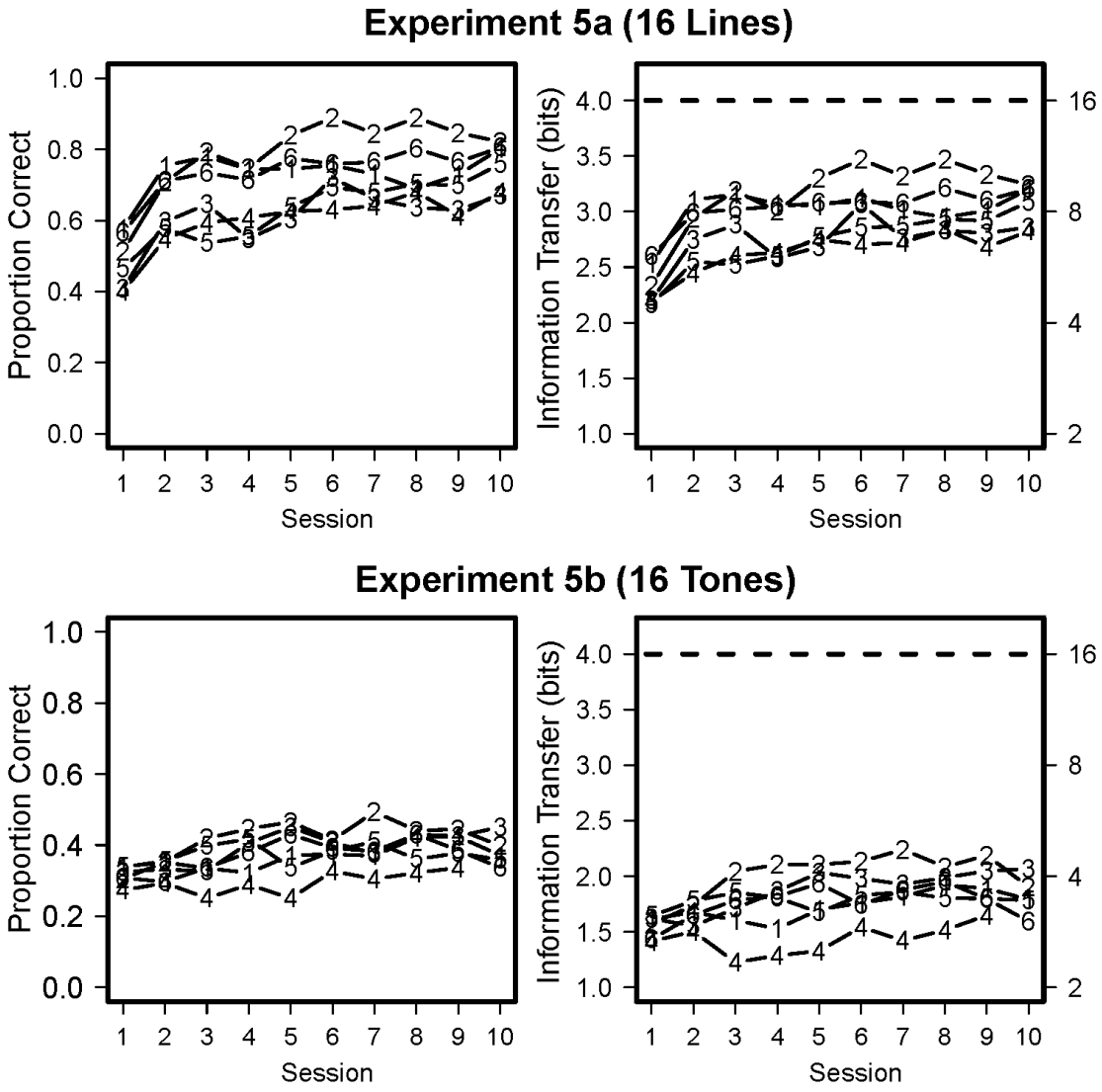


Figure 5. Proportion correct and information transfer as a function of session for Experiment 5a and 5b (lines varying in length and tones varying in loudness respectively). The right hand axis on the information transfer graph is the equivalent number of stimuli that were perfectly identified (or 2^{bits}). The dashed line indicates the maximum amount of information transmission, $\log_2$ (*number of stimuli*)

To better understand the difference between learning with lines and tone intensities, we further examined accuracy for each stimulus type. Figure 6 plots the proportion of correct identifications against ordinal stimulus magnitude, separately for the two stimulus continua, and separately for data from the beginning and end of practice. When practicing with line lengths (Experiment 5a), there was general improvement for all stimuli across the range, except where limited by ceiling effects for the smallest and largest lines. Although not shown here, corresponding plots for all other experiments show the same pattern as Experiment 5a. However, for tone intensities, there was no reliable improvement for tones in the middle of the range (#5-#11). This suggests that the limited amount of learning we observed for tone intensities was restricted to tones near the ends of the range.

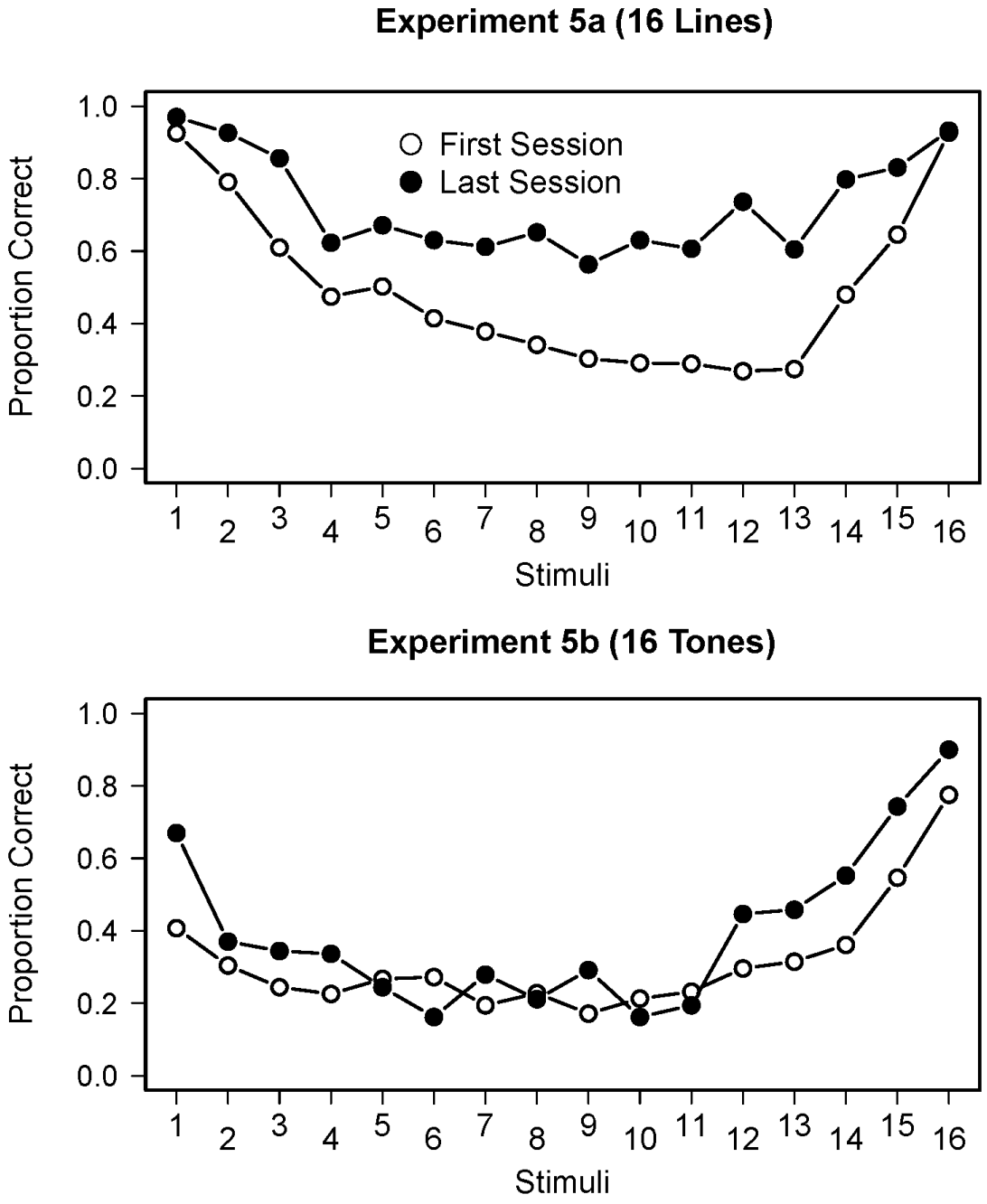


Figure 6. The proportion of correct identifications plotted against ordinal stimulus magnitude for Experiment 5a (Lines) and 5b (Tone Intensity), for both the first and the last 540 trials.

Discussion

Participants who practiced with either 16 lines varying in length or 16 tones varying in intensity both demonstrated significant improvements in performance. Even though the improvement for both experiments was statistically reliable, participants who practiced with tone intensities showed a much smaller learning effect and a significantly lower information transfer limit than those practicing with lines. Participants practicing with tone intensities, in contrast to participants who practiced with other continua, also failed to improve their performance consistently across the stimulus range (see Figure 6).

Experiments 1-5 together suggest an interesting possibility – that the amount of improvement through learning is closely related to the initial level of performance, prior to practice. For example, initial accuracy with tones of varying intensity (Experiment 5b) was poorer than for any other stimulus continuum, and so was the amount of improvement with practice. Conversely, accuracy with lines of varying length was initially very high, and so was the amount of improvement with practice. In Experiments 6 and 7, we explore the relationship between initial performance and learning, and further investigate the generality of the learning effect across different stimulus dimensions, using tone frequency.

Experiment 6

Experiment 6 uses tones of varying frequency. Other research (e.g., Pollack, 1952; Garner, 1953; Stewart, Brown, & Chater, 2005) has shown that pre-practice performance with tone frequency is similar to, but slightly better than, that for tone intensity. From this, we hypothesize that the amount of improvement from practice will be a little more than that observed for tones of varying intensity, but still less than that

observed for lines of different length.

Method

Stimuli. Stimuli were 36 tones varying in frequency. The range of frequencies (see Table 2) mimicked piano key frequencies, ranging from A3 to G#5 (220Hz to 1661Hz). Tones were pure sine waves, generated using Matlab R2008b, and were presented via headphones at a constant sound pressure, corresponding to 75dB at 1000Hz.

Table 3. Range of frequencies used in Experiment 6 and 7. Frequencies corresponding to musical notes on a keyboard from A3 to G#5

Frequencies								
220.0	233.1	246.9	261.6	277.2	293.7	311.1	329.6	349.2
370.0	392.0	415.3	440.0	466.2	493.9	523.3	554.4	587.3
622.3	659.3	698.5	740.0	784.0	830.6	880.0	932.3	987.8
1046.5	1108.7	1174.7	1244.5	1318.5	1396.9	1480.0	1568.0	1661.2

Procedure. On each trial, a fixation cross appeared for 500ms, before the tone was played through the headphones for one second. Participants were free to respond either during or after playback. Feedback was as in Experiment 1a; participants were given two response opportunities. Buttons were available on-screen in 3 horizontal rows of 12, and participants responded using the mouse. The buttons had not only the numerical label normally associated with absolute identification (i.e., 1…36), but also the corresponding piano key note (i.e., A3…G#5). Three of the six participants had some musical training; the other three had none at all.

In 8 of the 10 sessions, participants practiced for 6 blocks of 108 trials each. In

the first and last session however, participants only completed 4 blocks of 108 trials. Fewer experimental trials were completed in this first and last session because participants also completed a brief pairwise discrimination task. This task consisted of 2 blocks of 72 trials, during which participants were asked to discriminate between adjacent stimuli in the set. Adjacent tones were presented sequentially – the first tone was played for one second, followed by 500ms of silence, and then either the next higher or lower frequency tone in the stimulus set was played for one second. The participant was then asked to indicate which of the two tones was higher. This pairwise discrimination task simply confirmed that all participants were perfectly able to discriminate between adjacent stimuli, both before and after practice.

Results

The added pairwise discriminability task meant that the number of trials in Session 1 and Session 10 of Experiment 6 was not equal to those in other sessions. Since information transfer is sensitive to sample size, "pseudo sessions" of 540 trials each were used for analysis. Two participants were unable to complete all six blocks of the experiment within the allotted time frame in each session, and hence completed fewer trials (4104 and 3996 trials for each participant respectively, equivalent to 7 pseudo sessions of 540 trials) than other participants (from 5940 to 6264 trials, or 11 pseudo sessions). The lines for individual subjects in

Figure ***7*** reflect this imbalance in trial numbers.

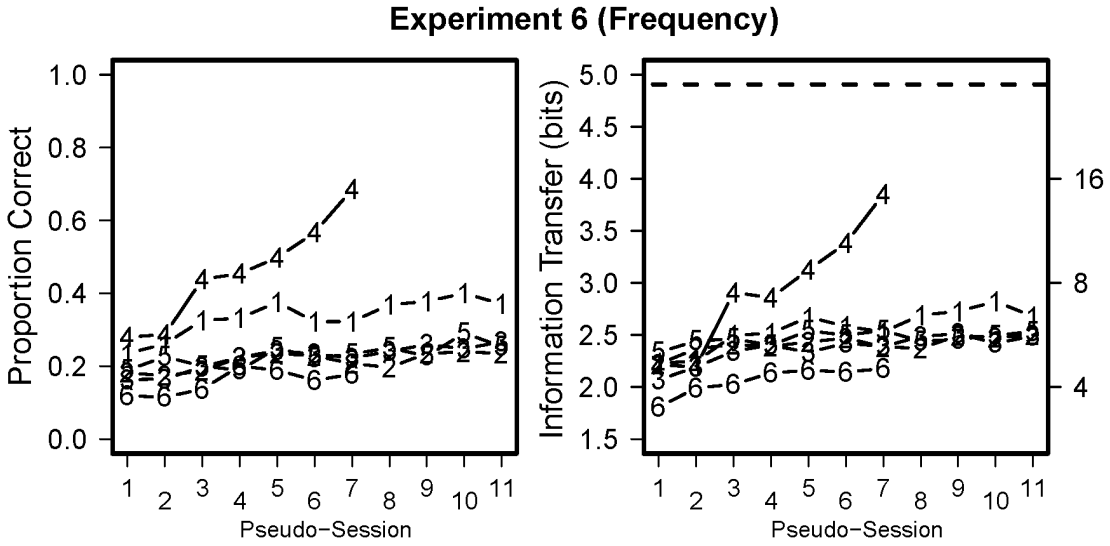


Figure 7. Proportion correct and information transfer for Experiment 6 (tones varying in frequency). The right hand axis on the information transfer graph is the equivalent number of stimuli that were perfectly identified (or 2^{bits}). Each pseudo session is equivalent to 540 trials. Two participants completed fewer trials than other participants and hence only have data for seven pseudo sessions. The dashed line indicates the maximum possible information transmission, $\log_2$ (*number of stimuli*)

Those participants who completed 11 pseudo sessions demonstrated a reliable improvement in information transfer from 2.21 to 2.59 bits (6.02 stimuli) and in accuracy from 19% to 30% (respectively: $F(2.47,7.42)=7.18$, $p=.02$; $F(1.38,4.13)=9.82$, $p=.03$). One participant was quite different from the others, and only this person exceeded Miller's (1956) bound of 7±2 stimuli, identifying the equivalent of 14.4 stimuli. This exceptional participant was one of three participants in this experiment who had several years of musical training (the other two such participants performed just like the three untrained participants).

Discussion

As expected, the level of performance in the initial session was a little better than that for tone intensity (Experiment 5b) but lower than for all our experiments with visual stimuli. In line with our hypothesis, the amount of improvement due to practice was also greater than for tone intensity, but less than that observed for comparable visual experiments. One remarkable participant showed a very large improvement with learning, even relative to the experiments with visual stimuli. The exceptional participant was the one with the most musical experience, and also the participant who began their musical training at the youngest age. These facts agree with findings from the absolute pitch literature, suggesting that early and lengthy musical training encourage the development of absolute pitch (e.g. see Takeuchi & Hulse, 1993). Newer research also suggests that absolute pitch ability exists at a baseline rate in the general population of people with little musical training (e.g. Ross, Olson & Gore, 2003). We are examining the relationship between absolute pitch and absolute identification, as well as the effect of practice on each, in experiments currently underway in our laboratory.

Experiment 7

We noted that better initial performance was correlated with greater improvement through practice. However, in Experiments 1-6, this correlation was observed across different stimulus manipulations. That is, some kinds of stimuli support better initial performance than others, and these also tend to support greater learning effects. In Experiment 7, we decouple initial performance level from any stimulus manipulations, by instead manipulating participants' motivation.

Method

Experiment 7 replicated Experiment 6, with a six new participants and only one methodological difference: participant reimbursement was contingent on performance. Correct and incorrect responses were rewarded differently (see Table 4), but we provided a minimum reimbursement of $150 for ten sessions. The maximum reimbursement actually achieved by a participant was just over $220.

Table 4. Method of reimbursement used in Experiment 7. The rows represent the first and second attempts made by the participant to name the stimulus, and the columns represent the accuracy of these attempts.

Rate of Reimbursement for Experiment 7			
	Correct	One Off	Two Off
1st Response	$0.05	$0.03	$0.02
2nd Response	$0.01	Nil	Nil

Results

Similar to Experiment 6, performance reliably increased across the ten sessions for both accuracy and information transfer (see Figure 8: accuracy: $F(1.47,7.35)=10.85$, $p=.009$; information transfer: $F(1.63,8.16)=14.06$, $p=.003$). In line with our hypothesis that motivation may be manipulated by monetary incentive, the average amount of improvement was larger ($t(8.8) = 2.39$, $p = .04$) in Experiment 7 than in Experiment 6 (an average of 0.38 bits for 11-session participants in Experiment 6, compared with 0.64 bits in Experiment 7), but the difference in initial performance was non-significant (2.14 bits 11-session participants in Experiment 6 to 2.15 bits in Experiment 7, $p=.87$).

A parametric test of the difference between the improvement seen in Experiment 6 and 7 is inappropriate, due to the exceptional participant from Experiment 6. Consequently we used a nonparametric Wilcoxon test, which takes account only of ordinal (rank) information, and so is not unduly distorted by the virtuoso participant. The results of this test supported the hypothesis that participants who received motivational reimbursements improved their performance more than those who did not ($W=29$, $p = .047$).

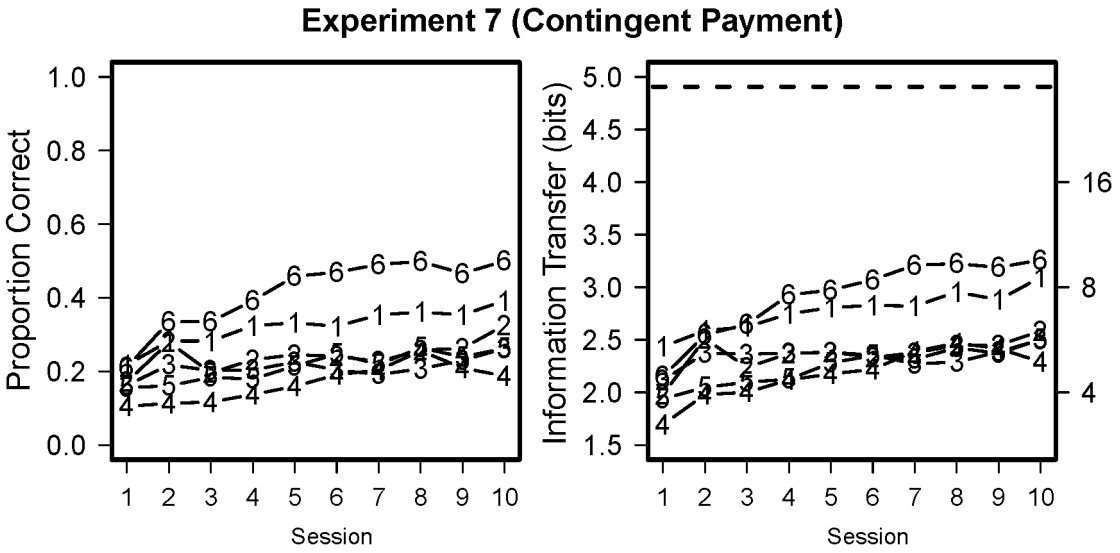


Figure 8. Proportion correct and information transfer for Experiment 7 (tones varying in frequency, using contingent payment methods). The right hand axis on the information transfer graph is the equivalent number of stimuli that were perfectly identified (or 2^{bits}). The dashed line indicates the maximum amount of information possible, $\log_2$ (*number of stimuli*)

Discussion

Participants who received extra motivation through monetary reimbursement based on response accuracy showed almost twice as much learning as those who received reimbursement independent of performance. This result is consistent with Rouder et al.'s (2004) suggestion that motivation is required for learning. Experiment 7 also has implications for the effect of stimulus modality on learning. If stimulus modality were the only determining factor for practice effects, we would have expected little difference between Experiments 6 and 7. Instead, when participants were suitably motivated we observed that they improved by slightly less than those in the visual modality experiments, but much more than in the two other auditory experiments.

While the amount of improvement observed in Experiment 7 compared with Experiment 6 emphasizes the importance of motivation, there was no significant

difference between initial (first session) performance levels in the two experiments. This makes it difficult to directly evaluate the hypothesis that initial performance predicts overall improvement. Future research could produce a more direct test of the hypothesis by experimentally manipulating the initial performance level. However, we note that our initial attempts at such experiments have proven unsatisfactory, because almost all manipulations that influence initial performance level involve manipulations of the stimuli, thus confounding the critical hypothesis with other hypotheses regarding stimulus-driven effects.

Regardless of the motivational manipulations in Experiment 7 however, no participant came close to performing as well as the one exceptional participant in Experiment 6 (who performed better than the majority of participants across all experiments). This finding speaks to the strength of individual differences in absolute identification performance, both in initial performance and amount of learning. We analyse these individual differences across experiments below.

Summary of Results

Table 4 contains a summary of information transmission results from the 10 conditions in our 7 experiments. Through these experiments we have shown that the learning observed for lines of varying length in Rouder et al. (2004) was not due to virtuoso participants or atypical methodological aspects of their design: Experiment 1 showed that learning was not due to the two-response method and Experiment 2 showed that learning was not due to external visual cues. Experiment 3 showed that the extended stimulus presentation time associated with lines compared with auditory stimuli was not required for learning. In Experiment 4 we showed that substantial learning also occurred for another visual stimulus – lines of varying inclination.

Experiment 5 showed very small practice effects for tones of varying intensity, consistent with the conventional wisdom about absolute identification (e.g., Shiffrin & Nosofsky, 1994) participants reached a low information transmission limit after a few sessions. Experiments 6 and 7 demonstrated that learning was possible for another auditory continuum, particularly when participants were well motivated. Participants practicing with tones of varying frequency were able to learn much more than those with tones of varying intensity, but not as much as those who practiced with most of our visual stimuli.

Table 5. Summary of Results. Note that all results are calculated based on pseudo session, or every 540 trials, for ease of comparison. Note also that for Experiment 6 the averages in brackets represent those participants who completed eleven pseudo sessions.

Experiment	Stimulus Continuum	Set Size	Average Information (bits)			
			First Session	Minimum	Improvement	Maximum
1a	Lines (Length)	30	2.38	2.37	0.88	3.25
1b	Lines (Length)	30	2.28	2.28	0.87	3.15
2a	Dots (Separation)	30	2.45	2.45	0.85	3.30
2b	Dots (Separation)	15	2.08	2.08	1.03	3.11
3	Dots (Separation)	30	2.39	2.39	0.64	3.03
4	Lines (Angle)	30	2.37	2.37	0.80	3.17
5a	Lines (Length)	16	2.35	2.35	0.74	3.09
5b	Tones (Intensity)	16	1.56	1.53	0.42	1.96
6	Tones (Frequency)	36	2.14 (2.21)	2.14(2.21)	0.59 (0.38)	2.73 (2.59)
7	Tones (Frequency)	36	2.15	2.15	0.64	2.78

Rouder et al.'s (2004) results were surprising because they violated two truisms of absolute identification: that practice has little effect on performance, and that there is a severe limitation in performance, equivalent to 7±2 stimuli. Our results confirm and generalize Rouder et al.'s observation that practice can have a substantial effect on performance. However, the last column in Table 5 shows that on *average*, participants did not greatly exceed Miller's limit of nine stimuli after 10 hours of practice. Indeed, the equivalent number of stimuli perfectly identified after practice, averaged across visual stimuli, for which performance was best, was 9.88 stimuli, not much above Miller's upper limit.

Individual subjects, however, tell a different story. Of the 58 participants that took part in all experiments, 22 exceeded Miller's limit. Indeed, two participants (in Experiment 1a and 6) reached a maximum rate of information transfer over 4 bits in their last session (16.1 and 17.5 stimuli respectively). These results are reminiscent of participant RM in Rouder et al. (2004), who was able to perfectly identify approximately 20 stimuli. Given that Rouder et al. looked at the effect of practice for only three participants, it is possible that their participants are best thought of as equivalent to the better performers in our experiments. Indeed, given that two of their three performers were authors, we expect their results are also consistent with our findings regarding improvements due to increased motivation. It seems, therefore, that Miller's (1956) magical number 7±2 may be best interpreted as not being too far wrong for the *average* participant, even if this is not true for some individuals.

General Discussion

The deeper question our work has provoked is: what produces differences in the ability to increase capacity in absolute identification? Table 4 shows that the

experiments in which there was a large amount of improvement with practice were also the same experiments in which performance during the very first practice session was high. Figure 9 shows that this result extends, at least approximately, to the individual-participant level. That is, participants who performed well in their first practice session – no matter which experiment they were in – also tended to be those who showed large learning effects. Although far from perfect, the correlation between initial performance and improvement was substantial both for experiments with larger set sizes ($r(39)=.609$, for experiments with 30 or 32 stimuli) and smaller set sizes ($r(15)=.653$, for experiments with 15 or 16 stimuli, both $p<.001$). Participants' levels of initial performance were highly correlated with the stimulus modality used for their experiment ($r(54)=.94$, $p<.001$), but a partial correlation confirmed that individual differences in initial performance still explained unique variance in the amount of improvement with practice, even after removing the effects of stimulus modality ($r(53)=.40$, $p<.01$).

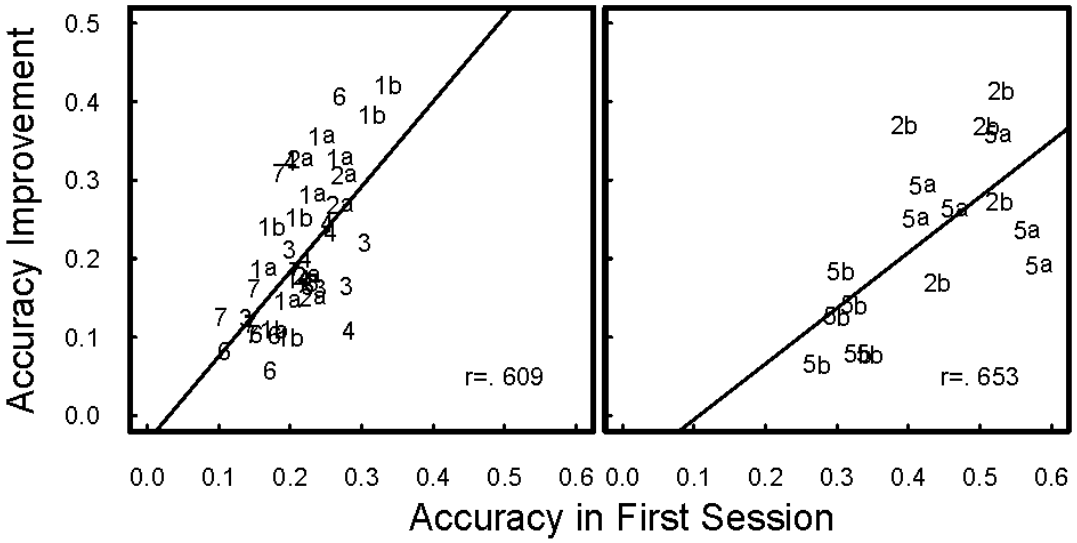


Figure 9. Improvement as a function of accuracy in the first 540 trials for experiments with larger set sizes (30 or 32 stimuli; left panel) and smaller set sizes (15 or 16 stimuli; right panel). Each point represents a single participant from a single experiment. The number denoting each participant on the graph is the experiment in which they took part.

The correlational analyses suggest that two important determinants of learning are (a) stimulus continuum (modality) and (b) individual differences between participants, at least partly caused by motivation, and that the effects of both factors are well described by performance during the first hour of experimental trials. Rast and Zimprich (2009) also found both strong individual differences and a positive correlation between participant's initial performance and learning rate in paired associate learning. This task bears some resemblance to absolute identification, where participants must learn stimulus-label associations (Siegel & Siegel, 1972). Note that the observed positive correlation between initial performance and learning need not have occurred. For example, a naïve expectation might have been a *negative* correlation, as higher initial performance leaves less room for improvement. Indeed, such a result seems assured in extreme cases where ceiling effects arise, such as for participants with almost perfect initial performance levels.

Our results do not uniquely identify the mechanism through which increased initial performance might be associated with greater overall improvement. However, several mechanisms seem likely candidates. For example, an exemplar model may naturally account for such improvement if information about the magnitude of a stimulus is stored *only* when the response is correct. A second possible explanation for the differences between experiments is that they depend on the pairwise discriminability of the stimulus sets, which might similarly vary between subjects. The Weber fractions for length and loudness are approximately 2% and 4.8%[2] respectively (Laming, 1986; Teghtsoonian, 1971), suggesting that people are less sensitive to changes in tones of varying loudness than lines of varying length. Such explanations seem implausible,

[2] Laming (1986) observed that the Weber fraction for pure tones improves as intensity increases according to the function $0.23A^{-0.14}$, where A is the amplitude. The magnitudes used in the current study (61dB-106dB) are sufficiently large to make Laming's function well approximated by the constant Weber fraction reported.

however, because in all of our experiments, stimulus separation was well above the Weber fraction, and research has shown that increasing separation between stimuli either has no effect at all (e.g., Pollack, 1951; Gravetter & Lockhead, 1973) or results in a quite small improvement in performance (e.g., Stewart et al. 2005; Lacouture, 1997).

Theoretical implications

Recent years have seen the development of several comprehensive models for performance in absolute identification (e.g., Petrov & Anderson, 2005; Kent & Lamberts, 2005; Stewart, Brown & Chater, 2005; Brown, Marley, Donkin & Heathcote, 2008). Our findings present severe challenges for these theories on several fronts, challenges which may require substantial re-development of the models. Such development is beyond the scope of this paper, and so we limit ourselves to delineating the problem, and providing an example of the direction that model development could take.

All theories of absolute identification respect the received wisdom in the field. No modern theories include any mechanism by which sustained practice can improve performance, and all theories take pains to treat all stimulus continua identically, as long as pairwise discrimination is perfect. Both of these assumptions are challenged by our results, and those reported by Rouder et al. (2004) – theories must predict learning with practice, and this learning should be different for different stimulus continua. The third major challenge for theoretical accounts is to accommodate the correlation we observed between initial performance and the amount of improvement with practice. It is not yet obvious to us how to develop a theory for absolute identification that accommodates our results in a natural way. However, as a proof-of-concept, we illustrate that it is possible to build learning effects into the SAMBA model for absolute identification (Brown et al., 2008). Similar illustrations are likely able to be constructed for other models.

Most theoretical accounts of performance in absolute identification agree that incorrect responses arise from two separate sources - systematic biases, and capacity limitations – and it is reasonable to posit that learning may improve performance by acting on either source. Contrast is one important systematic effect for the former type where decisions are biased away from stimuli observed a few trials earlier (e.g., if one observed a large-magnitude stimulus two or three trials previously, the current decision is likely to be biased towards smaller responses). Like all systematic biases, contrast reduces accuracy, but Triesman and Williams (1984) showed how contrast can be viewed as an adaptive mechanism that helps the observer track changes in a non-stationary stimulus environment. For example, SAMBA (Brown et al., 2008) attributes contrast effects to the re-direction of selective attention towards recently-seen magnitudes. This improves performance in a changing environment, by keeping attention directed towards relevant stimulus magnitudes. Although this mechanism is adaptive in general, and particularly when the stimulus set is unfamiliar, our experiments employed a fixed set of stimulus magnitudes for thousands of trials, making tracking unnecessary. In this case contrast impedes performance without any benefit, and so it would be rational to reduce contrast with practice.

Analysis of the data from Experiment 1a support this notion. Figure 10 illustrates sequential effects in the data from Experiment 1a, using an “impulse” plot (Ward & Lockhead, 1971). It shows average error as a function of the number of trials since stimulus presentation for data from the first and last sessions of Experiment 1, averaged over participants and over groups of ten adjacent stimuli (i.e., line #1 represents stimuli #1-#10, line #2 represents stimuli #11-#20 and line #3 represents stimuli #21-#30). The data from the first session of practice (left panel) show standard bias effects: assimilation of the responses towards the stimulus from the previous trial,

and contrast of responses away from stimuli from earlier trials. For example, when a small stimulus (line #1) was shown on the previous trial (lag=1), average errors were negative, meaning that responses tended to be smaller than the correct response (i.e., errors are biased towards the previously presented small stimulus). When the same small stimulus was presented a few trials previously (lag > 1), the data show contrast, where errors tend to be too large when the stimulus presented two or more trials ago was small (i.e., biased away from the previously presented small stimulus). Data from the final session of practice (right panel) are unusual - the magnitude of the contrast effect decreased markedly with practice, while the assimilation effect did not change much.

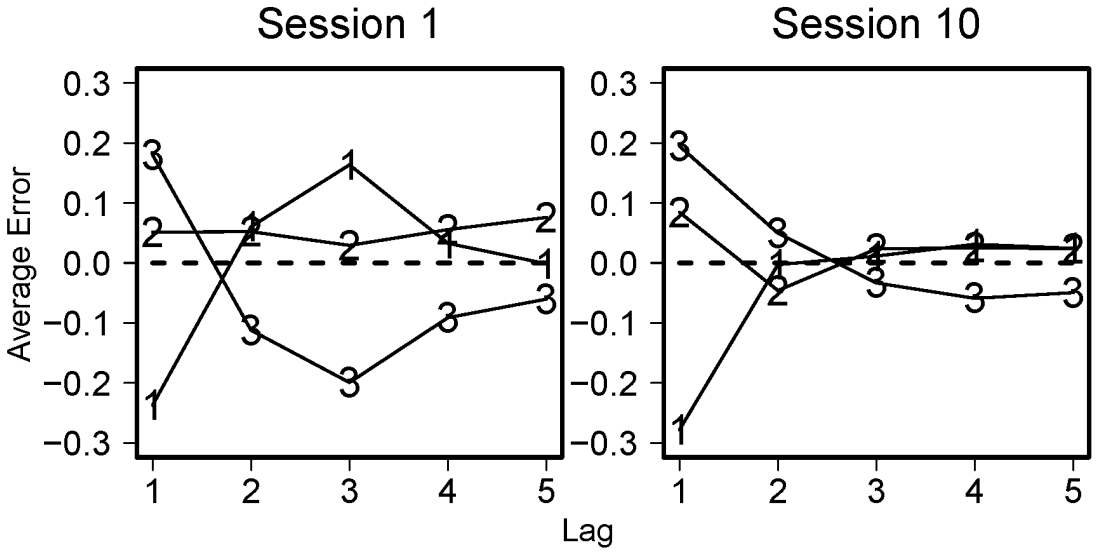


Figure 10. Impulse plots for data from Session 1(left panel) and Session 10 (right panel) in Experiment 1a. The different lines represent the magnitude of the stimulus presented 1..5 trials previously: line 1 = stimuli #1:#10, line 2 = stimuli #11:#20 and line 3 = stimuli #21:#30. The x-axis (lag) shows the number of trials since the occurrence of the stimulus used to condition the three lines.

Theoretical considerations, and the data, both suggest that one way to include learning effects in absolute identification is by reducing the magnitude of model parameters governing contrast, without altering assimilation. This approach fits naturally with the SAMBA model because SAMBA attributes contrast effects to a selective attention process, but assimilation effects to a more automatic, lower-level inertia in the decision process. To simulate this process in SAMBA, we began by setting all parameters at values estimated by Brown et al. (to fit a data set from Lacouture, 1997, parameters reported in Brown et al.'s Table 2). Then, to match the data from the first session of Experiment 1a we adjusted three parameters: we reduced the size of the assimilation parameter (D=.035), increased the size of the contrast parameter (M=.25), and we adjusted the rehearsal capacity (λ=.872) to match the overall accuracy level of

the data. The predicted impulse plot for the model using these parameters is shown in the left panel of Figure 11. To simulate the result of learning by reducing contrast magnitude, we steadily reduced the contrast parameter to *M*=0, over the course of learning, which removes almost all contrast effects from the model's predictions for the final session, as shown in the right panel of Figure 11.

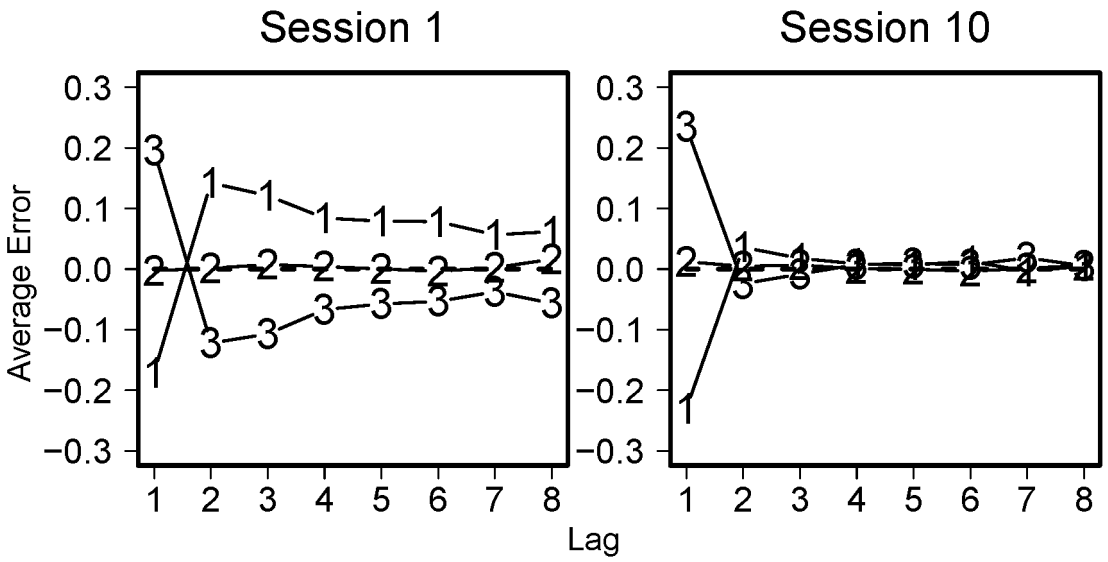


Figure 11. Impulse plots for SAMBA simulations for Session 1 (left panel) and Session 10 (right panel) with decreasing contrast.

When the effect of learning is modelled by the reduction of contrast magnitude, the impulse plots predicted by SAMBA match the data quite well. However, this way of modelling learning fails to capture the large improvement in accuracy. To match the large accuracy gains made by subjects, the model also needs to have its rehearsal capacity parameter changed with practice (from λ= .875 to 10). This version of the model, with both contrast and rehearsal capacity influenced by practice, matches both the impulse plots and the accuracy data from Experiment 1a. The left panel of Figure 12 shows the proportion of correct responses in Experiment 1a as a function of ordinal

stimulus magnitude separately for the first and last session of practice, and the right panel shows the same calculations for the predictions of SAMBA given the aforementioned parameter values. Although the model parameters were not adjusted to accommodate all effects (such as the tendency in the empirical data for better performance with small than large line lengths), SAMBA accounts well for the effect of practice. As in the data, the model predicts a substantial increase in performance over practice, and this increase is approximately equal in magnitude across the range of stimuli. SAMBA also predicts an increasing U-shape in this plot with practice, and the data appear to confirm this prediction.

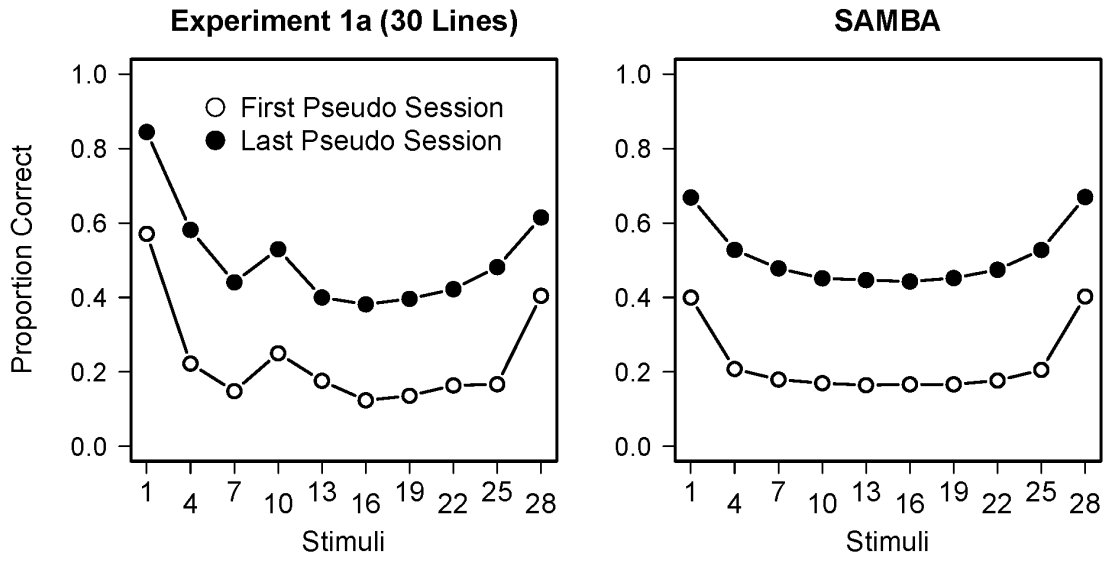


Figure 12. Response accuracy versus (rank) stimulus magnitude for Experiment 1a (left panel) and SAMBA's predictions (right panel). Open and filled symbols correspond to data from the beginning and end of practice, respectively. The data are averaged over participants, and over groups of three consecutive stimuli (e.g., the left-most point on each line represents average accuracy for stimuli #1, #2 and #3).

Similar accounts could be implemented in other comprehensive models of absolute identification, as they all include separate contrast and capacity parameters that can be manipulated as above. This approach advances theoretical understanding because it delimits the mechanisms by which practice improves performance, greatly constraining model development. However, there are three important questions that are left unaddressed:

1. By what mechanism(s) are rehearsal capacity and contrast magnitude changed by practice?
2. Why are there differences in the effect of practice when using different stimulus continua?
3. Why should pre-practice performance correlate strongly with the amount of improvement from practice?

Conclusions

Rouder et al. (2004) demonstrated that practice dramatically improved performance in absolute identification. We have shown that this effect generalises across most participants, and many different procedural manipulations. We also found reliable effects of some stimulus manipulations, a surprising correlation between initial performance and the gains from practice, and a dissociation between the effects of practice on assimilation and contrast magnitude. We showed that the fundamental result (improved accuracy with practice) as well as the dissociation, can be accommodated quite naturally within an existing comprehensive theory of absolute identification. The remaining findings stand as a challenge for the field: to develop a theory that naturally predicts improved performance and decreased contrast with practice, as well as providing a link between initial and final performance. A theory that provides such a link might then also explain the differences observed between stimulus continua,

because many of the differences in amount of learning between continua were captured by differences between initial performance on those continua.

Acknowledgements

This research was supported, in part, by Australian Research Council Discovery Project DP0881244 to Brown and Heathcote. Parts of the research reported herein contributed to a doctoral thesis for Dodds.

References

Braida, L. D., & Durlach, N. I. (1972). Intensity perception. II. Resolution in one-interval paradigms. *Journal of the Acoustical Society of America, 51*(2B), 483-502.

Brown, S.D., Marley, A.A.J., Donkin, C. & Heathcote, A.J. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review, 115*(2), 396-425

Garner, W. R. (1953). An informational analysis of absolute judgments of loudness. *Journal of Experimental Psychology, 46*(5), 373-380.

Gravetter, F., & Lockhead, G. R. (1973). Criterial range as a frame of reference for stimulus judgment. *Psychological Review, 80,* 203-216.

Hake, H. W., & Garner, W. R. (1951). The amount of information in absolute judgments. *Psychological Review, 58*(6), 446-459.

Hartman, E. B. (1954). The influence of practice and pitch distance between tones on the absolute identification of pitch. *The American Journal of Psychology, 67*(1), 1-14.

Kent, C. & Lamberts, K. (2005). An exemplar account of the bow and set-size effects in absolute identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 289 –305.

Lacouture, Y. (1997). Bow, range, and sequential effects in absolute identification: A response-time analysis. *Psychological Review, 60,* 121-133.

Lacouture, Y., & Lacerte, D. (1997). Stimulus modality and stimulus-response compatibility in absolute identification. *Canadian Journal of Experimental Psychology, 51*(2), 165-170.

Laming, D. (1986). *Sensory Analysis*. London: Academic Press.

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

Norwich, K. H., Wong, W., & Sagi, E. (1998). Range as a factor determining the information of loudness judgments: overcoming small sample bias. *Canadian Journal of Experimental Psychology, 52*(2), 63-71

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.

Pollack, I. (1951). Sensitivity to differences in intensity between repeated bursts of noise. *Journal of the Acoustical Society of America, 23*(6), 650-653.

Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America, 24*(6), 745-749.

Rast, P., & Zimprich, D. (2009). Individual differences and reliability of paired associates learning in younger and older adults. *24*(4), *Psychology and Aging*, 1001-1006

Ross, D. A., Olson, I. R., & Gore, J. C. (2003). Absolute Pitch does not depend on early musical training. *Annals of the New York Academy of Sciences*, 999, 522-526.

Rouder, J. N. (2001). Absolute identification with simple and complex stimuli. *Psychological Science, 12*, 318-322.

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.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379-423

Shiffrin, R., & Nosofsky, R. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review, 101,* 357-361.

Siegel, J. A., & Siegel, W. (1972). Absolute judgment and paired-associate learning:

kissing cousins or identical twins? *Psychological Review, 79*(4), 300-316.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, 112(4), 881-911.

Takeuchi, A. H., & Hulse, S. H. (1993). Absolute pitch. *Psychological Bulletin*, *113*(2), 345-361.

Teghtsoonian, R. (1971). On the exponents in Stevens' Law and the constant in Ekman's Law. *Psychological Review*, *78*(1), 71-80.

Thorndike, E. L. (1932). The Fundamentals of Learning. New York: Teachers College Press.

Triesman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91*(1), 68-111.

Ward, L. M., & Lockhead, G. R. (1971). Response system processes in absolute judgment. *Perception & Psychophysics, 9*(1),73-78

Weber, D. L., Green, D. M., & Luce, R. D. (1977). Effects of practice and distribution of auditory signals on absolute identification. *Perception and Psychophysics, 22*(3), 223-231

Chapter Two

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

Pennie Dodds[1], Christopher Donkin[2], Scott D. Brown[1], Andrew Heathcote[1], A. A. J. Marley[3]

[1]School of Psychology, University of Newcastle, Australia
[2]Department of Psychology & Brain Sciences, Indiana University, Indiana
[3]Department 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, Dosher & 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.

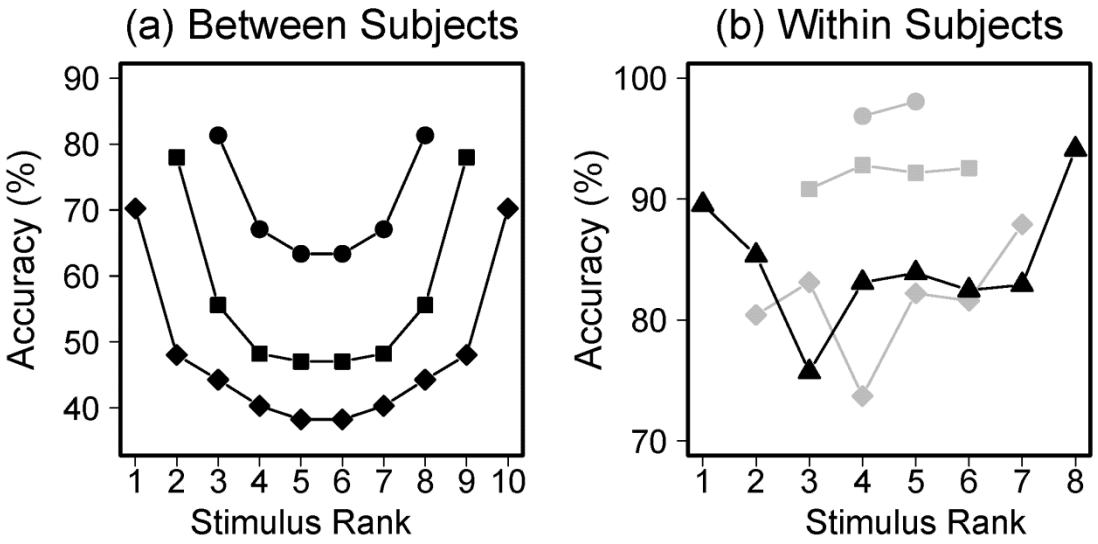


Figure 1. (a) Between-subjects data from Stewart et al.'s (2005) Experiment 1 and (b) within-subjects data from Kent and Lamberts' (2005) Experiment 2. Each line on each graph represents a different set size. All set sizes in Figure 1 (b) are in grey except set size n=8, for ease of comparison. Both 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 other[3], 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 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 stimuli[4].

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

[3] 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.

[4] 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.

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.

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^{o} to 6.7^{o}). 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, $F(1,77)=10.9$, $p=.001$).

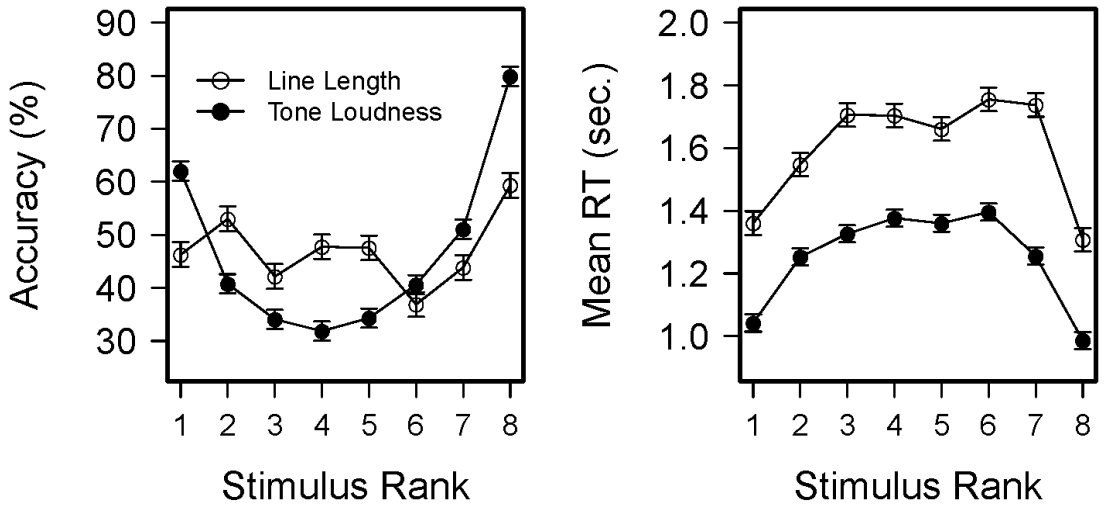


Figure 2. Accuracy and mean RT as functions of ordinal stimulus magnitude for Experiment 1. The lines represent different stimulus sets – line length or tone loudness. Error bars are 95% confidence intervals calculated in the repeated-measures manner described by Loftus and Masson (1994), separately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

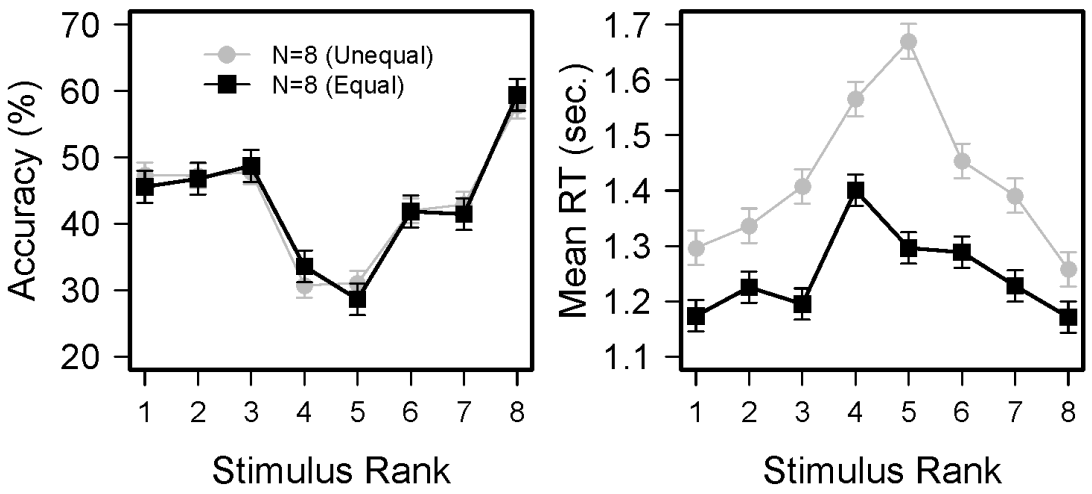


Figure 3. Accuracy and mean RT as functions of ordinal stimulus magnitude for Experiment 2. Each line represents a different section of the experiment. Grey lines represent the Section 2 (unequal presentation), black lines represent Section 3 (equal presentation). Error bars are as in Figure 2.

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.

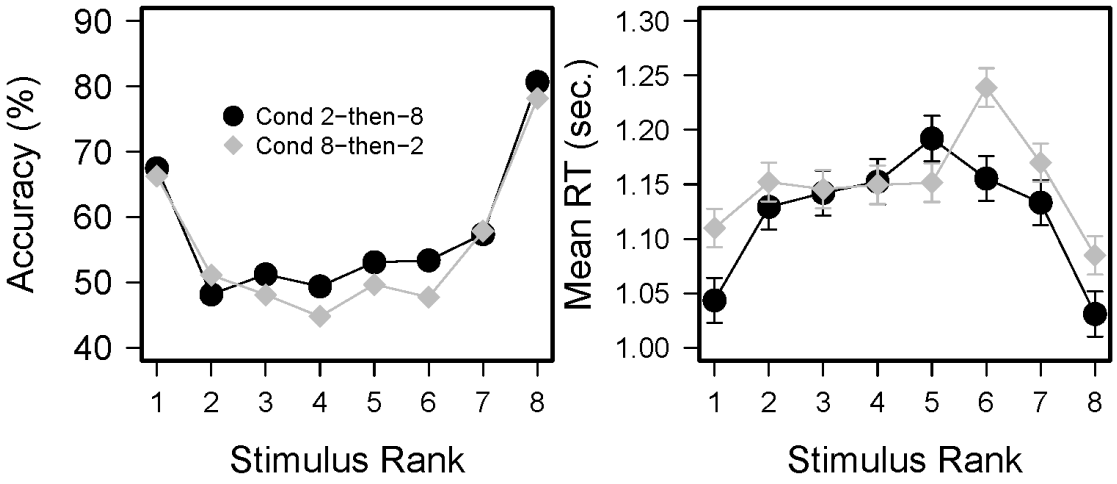


Figure 4. Accuracy and RT as a function of stimulus magnitude for Experiment 3. Note that there are two conditions: some participants experienced the N=2 set size first, then N=8 (condition 2-then-8) and others experienced the reverse order (condition 8-then-2). Error bars are as in Figure 2. Note: error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

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.

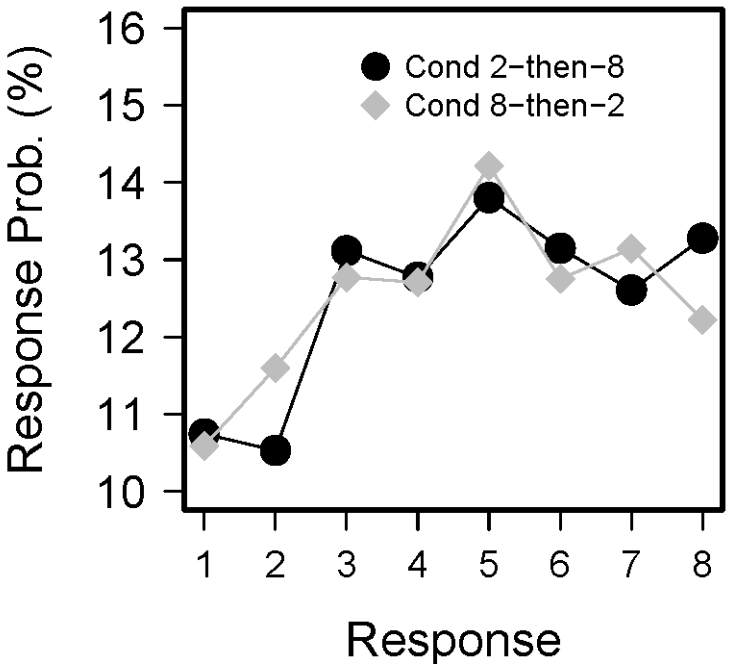


Figure 5. Response probability for Experiment 3

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., Dosher, 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.

SECTION TWO: THEORETICAL IMPLICATIONS

Included Papers

Brown, S.D., Marley, A.A.J., Dodds, P., & Heathcote, A.J. (2009). Purely relative models cannot provide a general account of absolute identification. *Psychonomic Bulletin & Review, 16, p.583-593*

Dodds, P., Brown, S. D., Zotov, V., Shaki, S., Marley, A. A. J. & Heathcote, A. (2011). *Absolute Production and Absolute Identification.* Manuscript submitted for publication

Dodds, P., Donkin, D., Brown, S.D., Heathcote, A. (2010) Multidimensional scaling methods for absolute identification data *In S. Ohlsson & R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society. Portland, OR: Cognitive Science Society.*

Dodds, P., Rae, B. & Brown, S. D. (2011). *Perhaps Unidimensional is not Unidimensional.* Manuscript submitted for publication

Chapter Three

Purely Relative Models Cannot Provide a General Account of Absolute Identification

Scott D. Brown[1], A.A.J. Marley[2], Pennie Dodds[1] & Andrew Heathcote[1]

[1]School of Psychology, University of Newcastle, Australia
[2]Department of Psychology, University of Victoria, Canada

Correspondence should be addressed to:
Scott Brown
Psychology Building
University of Newcastle
Callaghan NSW 2308
Australia

Phone: +61 249215760
Email: scott.brown@newcastle.edu.au
Web: http://www.newcl.org/

Word Counts
Abstract: 142
Body: 6406 (including footnotes & acknowledgments)
Figure Captions: 139
References: 17
Appendix: 760

Abstract

Unidimensional absolute identification – identifying a presented stimulus from an ordered set – is a common component of everyday tasks. Laboratory investigations have mostly used equally spaced stimuli and the theoretical debate has focused on the merits of *purely relative* vs. *purely absolute* models. Absolute models incorporate substantial knowledge of the complete set of stimuli, while relative models allow only partial knowledge and assume that each stimulus is compared to recently observed stimuli. We test and refute a general prediction made by relative models, that accuracy is very low for some stimulus sequences when stimuli are unequally spaced. We conclude that, although relative judgment processes may occur in absolute identification, a model must incorporate long term referents in order to explain performance with unequally spaced stimuli. This implies that purely relative models cannot provide a general account of absolute identification.

Absolute identification requires participants to identify which stimulus has been presented from a pre-specified set. In general, people are unable to accurately identify more than about 8-10 stimuli that vary on a single psychological dimension, which is surprising when comparative judgments with the same stimuli (i.e., judging whether one stimulus is less than, equal to, or greater than another stimulus) are completely accurate. For over 20 years, theories of absolute identification have been divided along a continuum from purely absolute accounts to purely relative accounts (see Brown, Marley, Donkin & Heathcote, 2008, and Stewart, Brown & Chater, 2005, for reviews). Absolute models assume some form of memory for the magnitude of each stimulus in the set – a set of long term "referents" that represent the stimuli. Relative models make a more parsimonious assumption by representing the stimuli using a limited set of partial information. Usually, relative models assume that the only long term memory is for a single scale factor related to the "spacing" of the stimuli, that is, to the magnitude *differences* between adjacent stimuli. The relative approach has proven successful in magnitude estimation tasks (e.g., Luce & Green, 1974; Marley, 1976), and the superficial similarity between the tasks suggests that the same approach may work in absolute identification.

The theoretical debate has progressed mainly by pairwise comparison of particular absolute and relative models, for example: Marley and Cook (1984) vs. Laming (1984); Petrov and Anderson (2005) vs. Stewart et al. (2005); and Stewart et al. (2005) vs. Brown et al. (2008). There have been one or two attempts at a more general comparison, but these have proven less diagnostic than hoped (see, e.g., Brown, Marley & Lacouture, 2007; Stewart, 2007; or Experiment 2 of Stewart et al., 2005). Here, we present a classwise comparison based on key differences in the way absolute and relative models map from the stimulus to the response space. Rather than relying on

small differences in quantitative goodness-of-fit, we identify a qualitative failure of relative models, caused by their core structure. In particular, we show that relative models make very strong and surprising predictions for experiments using unequally spaced stimuli. We then test these predictions with a new experiment that addresses a potential limitation of past research.

We focus on absolute identification experiments with unequally spaced stimuli presented with feedback, which means that participants are informed of the correct response after each trial. Feedback is almost always presented in numeric format (e.g., as a digit on a computer screen), and so researchers have used the term *numeric feedback* (Holland & Lockhead, 1968, p. 412). The numeric nature of feedback is important in our discussion of relative models, especially of the relative judgment model (RJM, Stewart et al., 2005). In fact, we show that relative models – including the RJM – are unable to account for certain aspects of data from experiments with unequally spaced stimuli. Although it is not the model described by Stewart et al., an extended version of the RJM can account very accurately for unequally spaced designs[5]. However, the extension contradicts the very core assumptions of the relative account of absolute identification, transforming the *relative* judgment model into an *absolute* judgment model, or at least into a hybrid *absolute-relative* judgment model.

Absolute vs. Relative Stimulus Representations

Absolute and relative models of absolute identification assume fundamentally different psychological representations. All absolute models include a flexible long-term memory representation of the entire set of stimuli used in an experiment. For example, Marley and Cook (1984) assume end anchors and an attention mechanism that

[5] Stewart et al. used this extended model to fit Lockhead and Hinson's (1986) unequal spacing data, but did not specify the nature of the extension.

together yield a long-term representation of the stimulus context and, indirectly, of stimulus magnitude. Petrov and Anderson (2005) posit explicit "anchors" which provide referents for the magnitude of each stimulus. Theories based on Lacouture and Marley's (1995) bow mapping (including Brown et al., 2008, and Lacouture & Marley, 2004) use both end anchors and a referent for each stimulus.

By contrast, a fundamental property of relative models is that they explicitly deny the use of memories for stimulus magnitudes. Instead, they use only magnitude *differences* between stimuli presented on successive trials, and assume that equal stimulus differences are mapped to equal differences on a response scale. Relative models have enjoyed considerable success, and have been able to account for almost all of the data accounted for by the more complex absolute theories (see, Stewart et al., 2005; Stewart, 2007). However, our analyses suggest that this success is a product of the way in which researchers have traditionally designed their experiments: almost always using designs where the stimuli are equally spaced and the feedback respects this equal spacing. This matches the assumption underlying relative accounts, but runs the risk that they will not generalize to absolute identification in the real world, where stimuli are often not equally spaced. There have been isolated investigations into the effects of unequally spaced stimuli (Lockhead & Hinson, 1986; Lacouture, 1997). However, these experiments have always used a within-subjects design to compare equal and unequal spacing conditions. This may have prompted participants to take particular note of the stimulus structure and encouraged them to use an absolute, rather than relative, processing mode – whether or not that mode was their default. Our experiments address this possibility by manipulating unequal spacing conditions between-subjects.

The representations used by relative accounts of absolute identification make a powerful and surprising prediction, that unequally spaced stimuli should result in very poor accuracy for certain trial sequences. On the other hand, absolute accounts predict that data from experiments with unequally spaced stimuli should not be radically different from standard data. To illustrate the point, consider the relative judgment models of Laming (1984), Holland and Lockhead (1968), and Stewart et al. (2005), and, for simplicity of the example, ignore sequential effects. These models depend critically upon a single estimate for the difference between adjacent stimulus magnitudes. This "spacing" estimate is used to scale the psychological difference between the current and previous stimulus into a difference in response units. The resulting estimate of the response difference between the current and previous stimulus is then added to the numeric feedback for the previous trial[6]. This numeric feedback informs the participant of the correct response for the previous stimulus, and so when the estimated response difference between the previous and current stimulus is added, a response can be generated for the current stimulus.

When the stimuli used in an experiment are unequally spaced, this process breaks down in the obvious manner. The single estimate used for the spacing between adjacent stimuli cannot capture all of the different spacings that exist between different stimuli. The relative model is forced into a compromise when scaling from stimulus differences to response differences, using some average estimate of the spacing between stimuli. This average estimate leads to errors whenever the current and prior stimuli are separated by spacings that are different from the average estimate. We develop the above argument more formally in the Appendix. There, we set out a very basic model

[6] Laming's (1984) model differs from the other two in that it assumes the response scale is the log of the numeric responses. This approach retains the core problems of relative models for unequally spaced stimuli when the numeric responses are 1,…,N.

that captures the core elements of relative judgment, but includes no extra components, such as random variability or sequential effects. We show that when stimuli are unequally spaced, the basic model predicts very low accuracy for certain combinations of current and prior stimulus magnitudes, regardless of the values given to the model's parameters. Below, we test this prediction using data from an experiment with unequally spaced stimuli, replicating Lockhead and Hinson's (1986) design. The simple model we analyze in the Appendix does not include many of the extra components used in cutting-edge relative models, so our analyses will not apply perfectly to those accounts. Therefore, we also show that the leading relative model (the RJM) cannot account for our data, or those from one of Lacouture's (1997) unequal spacing experiments. These analyses confirm that the same problems are observed in cutting edge relative models as are found in the basic architecture analyzed in the Appendix.

Methods

Participants

Introductory psychology students from the University of Newcastle took part in the study, receiving course credit as compensation: ten participants in the low spread condition, and eight in each of the other two conditions.

Stimuli

There were three spacing conditions: low spread, even spread and high spread. In each condition the stimuli were three 1000Hz tones of different intensities. The range of tone intensities was different in each condition, as illustrated in Figure 1. In the even spread condition the tones were equally spaced at 79dB, 82dB and 85dB. Stimuli in the other conditions were identical, except that in the low spread condition, stimulus #1 was

made less intense (73dB), and in the high spread condition stimulus #3 was made more intense (91dB).

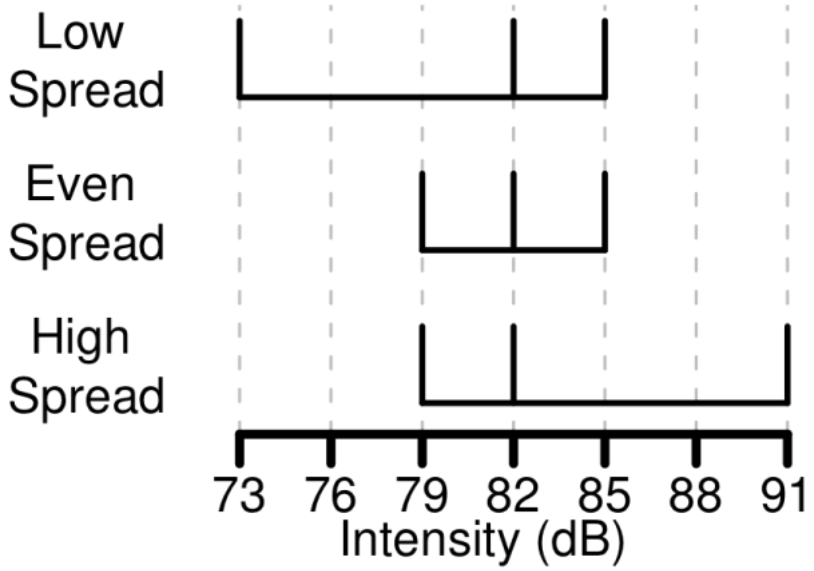


Figure 1. Schematic illustration of the stimuli used in the three different conditions.

Procedure

Each participant was randomly assigned to either the low, even or high spread condition. Each condition had three phases: digit identification, practice, and a test phase. The digit identification block was 90 trials in length, during which participants responded to a series of electronically pre-recorded numbers (1, 2 or 3). They were asked to press the corresponding number key on a regular keyboard; each number was played via headphones thirty times, in random order. This phase was intended to examine baseline reaction times for unambiguous stimuli, so that differences in mean response times for the three different response buttons (and fingers) could be identified. During practice, each of the three tones was played once, in ascending order of intensity. Each tone was labeled with the number 1, 2 or 3, which appeared on screen while the tone was played. Participants were required to press the corresponding key to

continue. For example, *"This is tone number 1, if you think you have heard this tone, press 1 to continue"*. The test phase had 10 blocks. In each block, each stimulus was presented 30 times, with the order of the 90 trials being randomized. On each trial, a visual cue (+) was displayed for 500ms, then the stimulus was played for 1000ms, and the participant had up to 20 seconds to respond. If no response was made, the next trial was presented and a missing value recorded. If a response was incorrect, the correct answer was displayed on the screen for 1000ms. If the response was correct, "*Correct*" was displayed on screen for 1000ms. Participants were required to take a minimum 30 second break between each block.

Results

Response times faster than 180msec or slower than 5sec were removed from the analysis, which accounted for fewer than 1% of trials in each condition. Results from the digit identification block showed there were no substantial differences in response speed across stimuli #1-#3, on average response times were 642msec, 678msec and 635msec respectively, and this pattern was maintained within the three experimental conditions. Figure 2 illustrates the absolute identification results. Results in the even spread condition were typical of traditional absolute identification tasks. Mean response time (top row, middle panel) was slower for the middle stimulus than the edge stimuli, although the mean difference was slight (59msec). Response probabilities (bottom row, middle panel) show the correct response was most frequent for each of stimuli #1-#3 – 78%, 79% and 87%, respectively. There was a slight asymmetry, with the softest stimulus identified less accurately, and more slowly, than might be expected relative to the loudest stimulus.

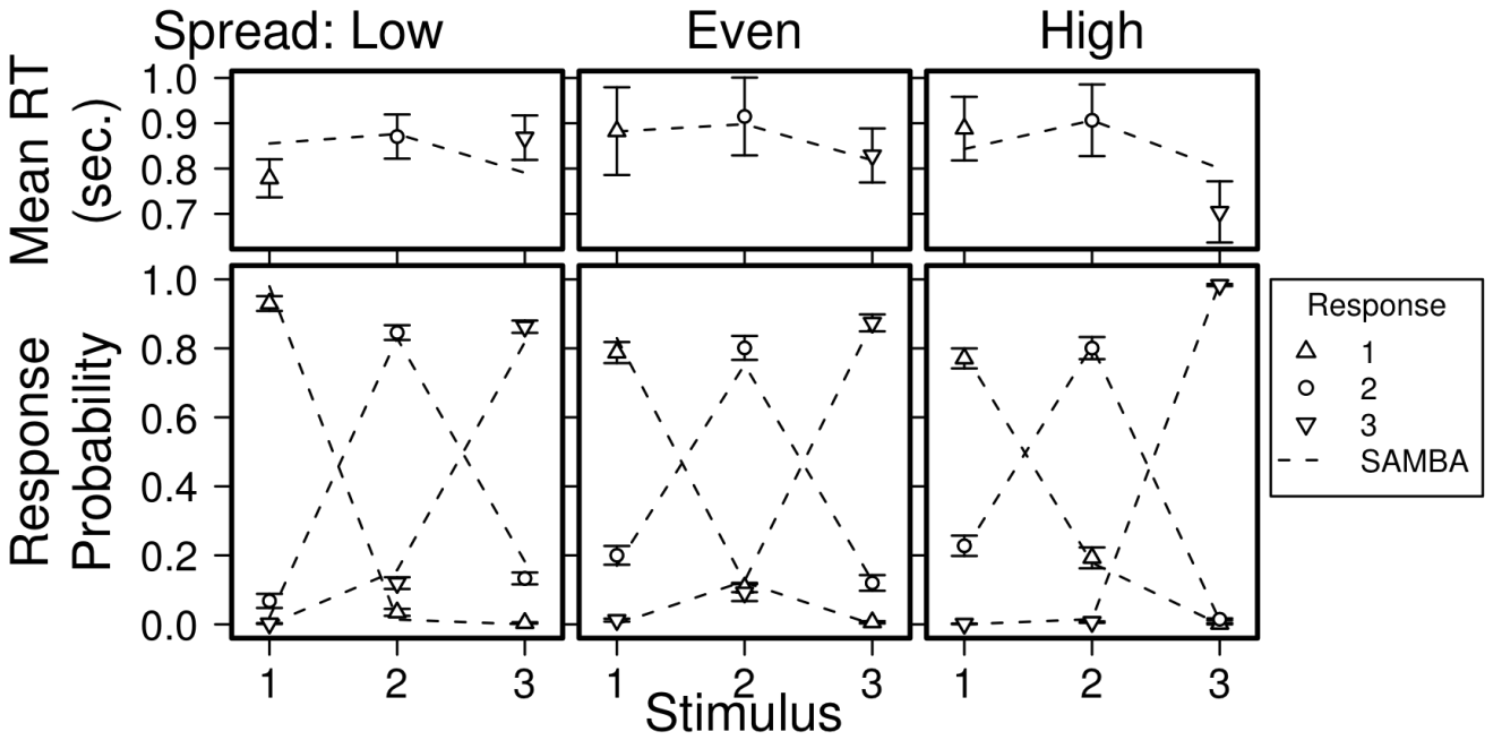


Figure 2. Mean response time (top row) and response probabilities (bottom row) for the low spread, even spread and high spread conditions. Triangle, circle and inverted triangle symbols depict data associated with responses #1, #2 and #3, respectively. Error bars are +/-1 standard error, calculated across participants, assuming normal distributions of means in the population. The dashed lines join predictions from the SAMBA model, discussed later.

Figure 2 replicates the key aspects of Lockhead and Hinson (1986). In the low- and high-spread conditions, responses to the privileged stimuli (the ones with greater separation) are more accurate than in the even spread condition ($t(16)=3.7$, $p=.001$ and $t(14)=4.4$, $p=.001$ respectively). Mean response times for the privileged stimuli were also faster, although the differences were not significant ($t(16)=0.9$, $p>.05$ and $t(14)=1.4$, $p>.05$ respectively). These advantages are unsurprising, as in each case the stimulus itself is different, either much louder or much softer.

What is more interesting is that the remaining two stimuli in the low- and high-spread conditions are more often confused than in the even spread condition, even though these two stimuli are physically identical across pairs of conditions. For example, stimuli #1 and #2 are physically identical in the even spread and the high spread conditions (79dB and 82dB in both cases), yet they are more often confused in the high spread condition than the even spread condition. In the even spread condition, response #1 is given on 11% of presentations of stimulus #2, and but this rises to 19% in the high spread condition, and the difference is significant ($t(14)=2.6$, $p=.014$). Similar patterns occur (with smaller magnitudes) for the other identical stimulus/response pairs.

An Absolute Account of the Data

Theories using absolute processes naturally account for data from unequally spaced stimuli because they include complete knowledge of the stimulus set, including long-term memories for the magnitudes of all stimuli in the set. When the spacing of the stimuli changes, so do these referents. The tracking process that carries out these changes may be specified in great detail (e.g., Petrov & Anderson, 2005; Treisman & Williams, 1984), or not (e.g. Brown et al., 2008), but nevertheless all absolute models include the necessary components. We use Brown et al.'s model (SAMBA) to illustrate.

SAMBA assumes that the magnitude of a stimulus is estimated in a noisy and error-prone fashion, which is then compared against long-term memories (referents) for each stimulus. When the physical spacing of certain stimuli is small, relative to the average spacing of stimuli in the entire set, so too is the difference between their referents. Since decisions are based on comparison with these referents, greater confusion is predicted between stimuli that are closer together, relative to the overall context of the experiment – just as observed in the data.

In Figure 2, the dashed lines join predictions from SAMBA, for both response times (top row) and response probabilities (bottom row). SAMBA's account of the data is very parsimonious – exactly the same parameter values are used to generate predictions for all three experimental conditions. The different predictions arise without parameter changes because the different stimuli in the three conditions provide different long-term referent values. These referents capture the critical qualitative patterns in both response times and response probabilities. The quantitative fit to the data is quite good, with all predicted response probabilities falling within .05 of the corresponding observed probabilities (root-mean-squared error, RMSE=.026).

SAMBA's predictions were generated by adjusting the parameters used by Brown et al. (2008) to fit data from the equally spaced condition of Lacouture's (1997) experiment. To fit the current data set, four parameters were changed. One parameter was adjusted to fit the overall level of accuracy (η=16); larger values of η endow the model with improved memory for the context of the experiment, allowing more precise estimates of stimulus magnitudes. A second parameter was adjusted to fit the overall level of response times (C=447); larger values of C correspond to more caution in evaluating evidence for the different responses. Finally, two anchor values (L and U) were changed to accommodate the asymmetry in the data; these anchor values describe

the range of stimuli that the observer sets as relevant for this experiment. Response accuracy is maximized if the range is set identical to the range used in each particular stimulus condition, but observers typically do not quite manage this. We fixed *L* to be 6dB quieter than the quietest tone in each condition, and *U* to be 3.3dB louder than the loudest in each condition. An even better fit to the data – particularly the response time asymmetry – could have been obtained by allowing differences in the anchors between conditions. Such differences are plausible, given the between subject manipulation, but our arguments do not rely on small differences in quantitative fit, and so the extra complexity is not necessary.

SAMBA estimates stimulus magnitudes using a selective attention mechanism based on Marley and Cook's (1984) rehearsal model. The details are in Brown et al. (2008), but the important point is that the averages of these magnitude estimates serve as referents. Magnitude estimates are expressed as ratios in the interval [0,1], with zero representing the lower anchor (*L*) and one representing the upper anchor (*U*). With the parameters above, in the even spread condition the average magnitude estimates for the three stimuli are {.4, .59, .78}, and these estimates capture the even spacing of the physical stimuli. In the low spread and high spread conditions, the average magnitude estimates are {.29, .71, .85} and {.29, .43, .85}, respectively. The latter two sets of estimates capture the relevant three-to-one stimulus spacings without the need for changes in parameters between conditions.

The Relative Account

Relative models make the strong prediction that response accuracy for certain stimulus sequences will be very low when stimuli are unequally spaced. For example, consider the relative models proposed by Laming (1984) or Stewart et al. (2005). Both models depend critically on a memory for the average spacing between adjacent

stimulus magnitudes (λ in Stewart et al.'s model, β in Laming's). Throughout this paper, we will use the symbol Z_i to represent the physical magnitude of the stimulus presented on trial i, measured on a logarithmic scale. The symbol S_i is used for the rank of that stimulus within the entire set of stimuli experienced by a participant. In the even spread condition of our experiment, we used three stimuli with physical magnitudes 79dB, 82dB, and 85dB. Relative accounts of absolute identification operate using the knowledge that 3dB separates adjacent stimuli, as follows. Suppose that the stimulus presented on the previous trial was Z_{n-1}=79dB (S_{n-1}=1) and the stimulus presented on the current trial is Z_n=82dB (S_n=2). The core elements of a relative model would operate by:

1. Estimating the magnitude difference between the current and previous stimulus (in this case, 82dB-79dB = +3dB difference).

2. Transforming the difference estimate into the numerical response scale, using the knowledge that adjacent stimuli are separated by 3dB, so that the +3dB difference is transformed to a difference of +1 response.

3. Converting the response difference into a response by adding it to the correct response from the previous trial, which is known by feedback. Thus, the response on the current trial would be the +1 difference added to the previous correct response (1), yielding the response 2 (which is correct).

When the stimuli are unequally spaced, this process breaks down. Our high spread condition used three stimuli with intensities 79dB, 82dB and 91dB – the loudest stimulus is much louder than before, but the other two are unchanged. Participants performed quite well in this condition, with better than 84% accuracy for each of the three stimuli. However, consider the relative judgment account of the same trial sequence as above, when stimulus Z_{n-1}=79dB (S_{n-1}=1) is followed by stimulus Z_n=82dB

(S_n=2), the magnitude difference is the same as before, +3dB. However, if the observer's long-term memory is based on the *average* difference between adjacent stimuli, they will use λ=6dB. This causes the observed magnitude difference to be transformed into a numerical response difference of only +½. When this response differences is added to the numeric feedback from the previous trial (1) the model predicts that the response given for the current trial should be equally likely to be 1 (incorrect) as 2 (correct). Manipulating λ can solve this particular problem, for example by using λ=3dB. However, this simply shifts the problem to other stimulus sequences (e.g., then *all* trials in which stimulus #2 follows stimulus #3 are classified incorrectly). This type of reasoning is formalized in the Appendix.

Figures 3 and 4 illustrate that this exact problem arises even in the fit of a much more complicated relative model, the RJM of Stewart et al. (2005). In this section, we focus on the RJM as described by Stewart et al.'s text and equations. Personal communication has revealed that Stewart et al. actually implemented a different version of their model, at least when dealing with experiments using unequally spaced stimuli. We call that model the *extended RJM*, and consider it carefully in the next section. Figure 3 shows that the global fit of the RJM is quite good, with RMSE=.042, which is in the same ballpark as SAMBA's fit (RMSE=.026). When fitting the RJM, we adjusted four parameters: one for the scaling of stimulus differences to response differences (λ); one for the effect of the prior trial on the current decision (α_1); a variance parameter (σ); and a decision threshold (χ_1). We had to allow the RJM to have different parameter values for the equally-spaced (λ=0.786dB, α_1=.312, σ=.208, χ_1=0.702 and χ_2=4-χ_1) and unequally spaced (λ=2.016dB, α_1=.087, σ=.092, χ_1=1.36 and χ_2=4-χ_1) conditions. The different values of the spacing parameter, λ, reflect the very different stimulus spacing

conditions in the equal versus the unequal spacing conditions[7]. These extra parameters (eight, as opposed to the four used by SAMBA) provide the RJM with some extra flexibility, which may concern some readers; however we were unable to find a common set of parameters that gave a reasonable fit to all three conditions. We also explored even greater parameter freedom for the RJM, by allowing independent parameters for the two response thresholds (χ_1 and χ_2) – this version of the model performed only marginally better than the symmetric version described above. Note that the RJM does not make predictions for response times, so Figure 3 shows only response probabilities.

[7] Our parameter estimate of λ=0.786dB agrees with the corresponding estimate from Stewart et al.'s (2005) fit to Lockhead and Hinson's (1986) experiment. When converted to units of dB, they found λ=0.746dB.

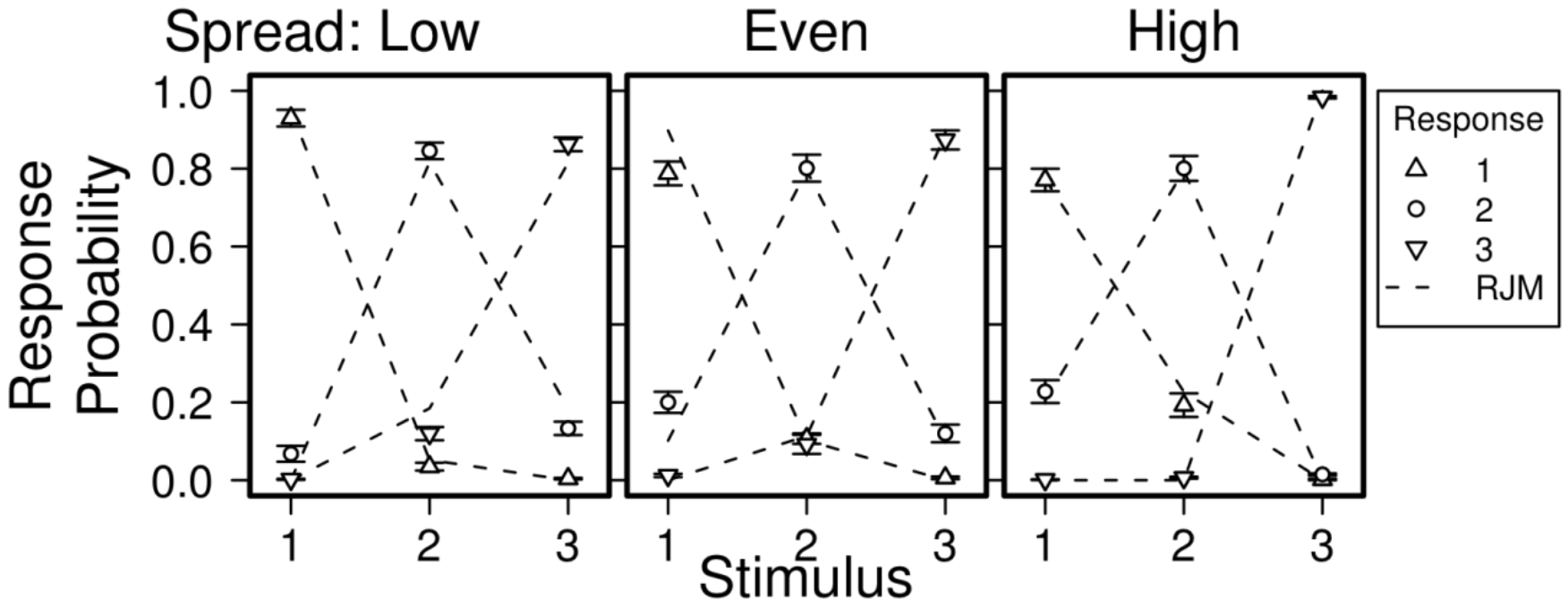


Figure 3. Data (points) and predictions (dashed lines connecting predicted values) from the RJM. For an explanation of the format, see Figure 2.

The previous discussion suggests that relative accounts predict very low accuracy for particular stimulus transitions, such as between stimuli #1 and #2 in the high-spread condition and stimuli #2 and #3 in the low-spread condition. Figure 4 graphs the accuracy associated with each stimulus (shown using different symbols) conditional on the previous stimulus (given by the x-axes). The three columns of Figure 4 show these graphs separately for the low-spread, even-spread and high-spread conditions. The top row shows just the data, the second row shows corresponding predictions from SAMBA, and the bottom row shows the predictions made by the RJM.

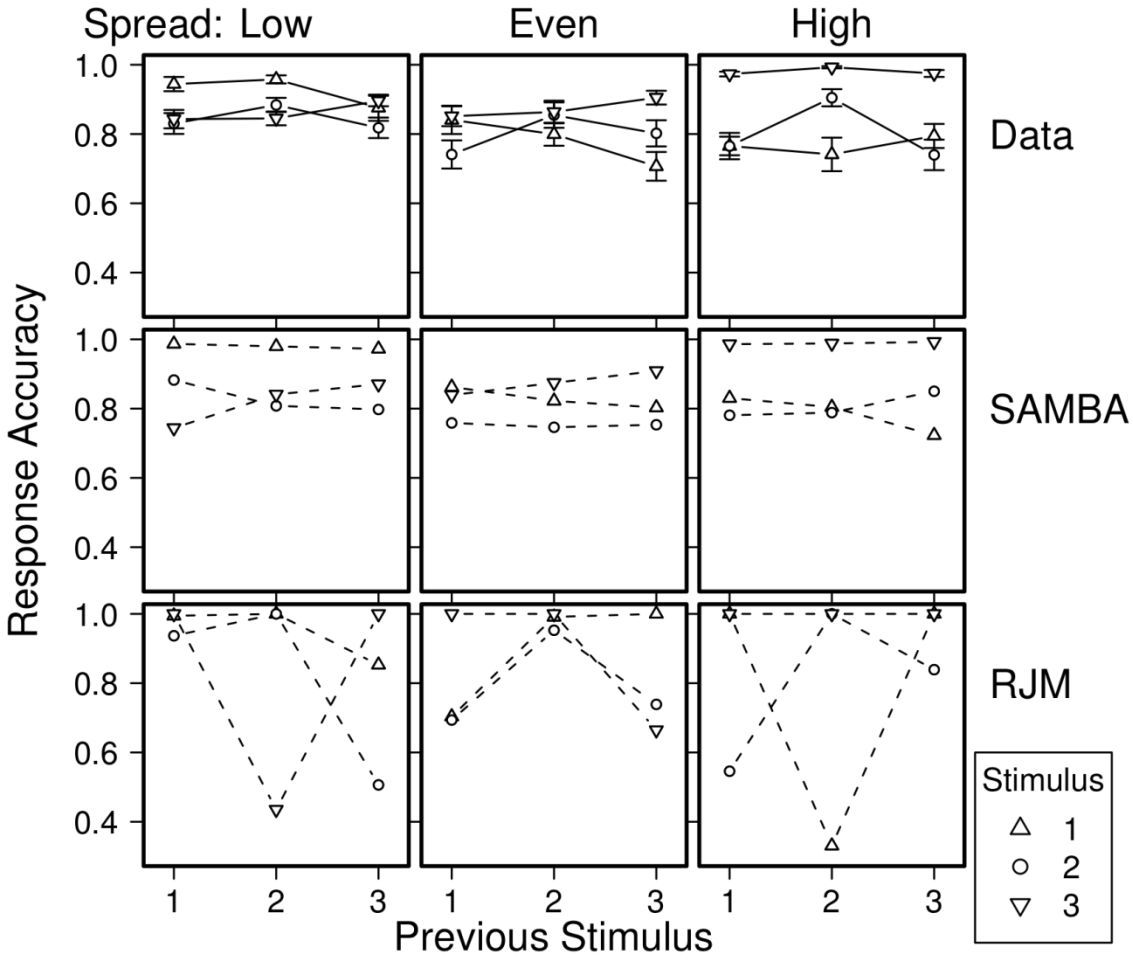


Figure 4. Response accuracy for each stimulus (shown by different symbols – see legend) conditioned on the previous stimulus (x-axis). Top row shows data, lower two rows show predictions from SAMBA and the RJM. The three columns correspond to the low spread, even spread and high spread experimental conditions.

The top row of Figure 4 show that participants performed quite well on *all* stimulus transition sequences – even the very worst accuracy was still 71% (when stimulus #1 followed stimulus #3 in the even spread condition). SAMBA's predictions, shown in the second row, match the data quite well (RMSE=.059), and the greatest mis-match between the data and SAMBA's predictions is only .12. In contrast, the predictions for the RJM, on the third row, are very different from the data. Just as expected, the predicted accuracy for some stimulus transitions is around 50%. The overall RMSE for RJM's fit to the sequential data is more than three times that of SAMBA (.19), as is the greatest mis-match between the sequential data and predictions (.41). These analyses demonstrate that the apparently adequate account of the data provided by the RJM in Figure 3 was really a consequence of averaging together large over-predictions for some conditional accuracy values together with large under-predictions for others. We tried to remedy this mis-fit by adjusting the free parameters of the RJM solely to optimize the fit shown in Figure 4, ignoring the overall mean response probabilities shown in Figure 3. This analysis resulted in a slight improvement in fit, but not enough to change the conclusions to be drawn, nor did it change the predictions of extremely poor performance for certain stimulus transitions. When the model was endowed with almost double the number of free parameters (an extra two for asymmetric response criteria, plus independent free parameters for all three conditions) and when all of those parameters were adjusted to optimize fit for Figure 4, the overall RMSE for the RJM was still double that of SAMBA (at .12) and the worst mis-fit was still very large (.31).

Lacouture (1997)

Lacouture (1997) also studied absolute identification with unequally spaced stimuli. He used a larger stimulus set, which has the consequence that relative models are less able to trade off under-prediction and over-prediction of the conditional data in order to provide an apparently adequate fit to the unconditional data. In one of his simplest conditions, he used a standard design with ten lines of increasing length that were equally log-spaced except for a large gap between the central pair of lines that was 6 times as large as the other gaps – using arbitrary units[8] for log-length, the lines' lengths were: {1, 2, 3, 4, 5, 11, 12, 13, 14, 15}. The data from this condition were very similar to those observed under standard conditions, except for improved response accuracy and latency for stimuli adjacent to the large gap. Figure 5 illustrates these data following Lacouture's analysis: plotting accuracy conditioned on the *response*, rather than the customary conditioning on the *stimulus*. Similar effects are observed with either analysis, but they are somewhat clearer in the response-conditioned version.

[8] The RJM is insensitive to arbitrary linear transformations of the psychological stimulus magnitudes, though the numerical value of some parameters in data fits may depend on the particular representation selected.

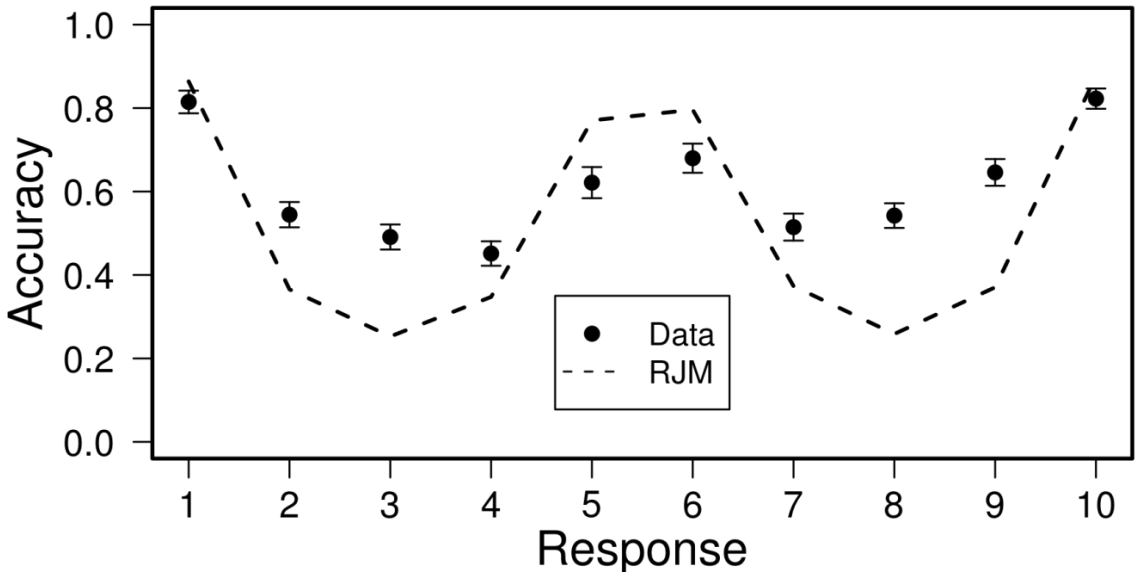


Figure 5. Response accuracy from Lacouture's (1997) large central gap condition (points), and predictions from the RJM (dashed lines). Error bars show +/-1 standard error, assuming binomial distributions.

Donkin, Brown, Heathcote and Marley (in press) demonstrate that SAMBA fits these data well (RMSE=.08), while simultaneously accounting for the associated response times, and for both types of data for Lacouture's other stimulus spacing conditions, all using the same set of parameters. Our analyses (see Appendix) show that the core architecture that underlies relative models makes inappropriate predictions for the choice data. To confirm that these problems are not limited to the basic relative architecture, we also fit the RJM to Lacouture's data. We optimized the RJM's parameters to fit only the data of Figure 5 (σ=0.074, λ=1.75 and *C*=.136; we could not obtain a better fit by adjusting the five sequential effect parameters $\alpha_{1\text{-}5}$). Other parameter settings allow the RJM to capture the accuracy values for responses 2-4 and 7-9 somewhat better, but always at the expense of far worse predictions for other responses. As expected from our analytic results, the RJM fits the data very poorly (RMSE=.18).

The relative account of Lacouture's (1997) data fails in exactly the manner predicted by our analysis in the Appendix. There is a tension in the model between transforming the small spacing between stimuli #1-#5 and #6-#10 (just one stimulus spacing unit) to numerical differences on the response scale, and transforming the large gap between stimuli #5 and #6 (six stimulus spacing units) to a numerical difference on that scale. The RJM settles on a compromise solution, estimating the spacing parameter at λ=1.75 spacing units. Of course, this compromise fails for certain stimulus transitions. For example, it makes inappropriate predictions whenever the current and previous stimuli lie on opposite sides of the large gap (i.e., when the current stimulus is between #1 and #5 and the prior stimulus was between #6 and #10, or vice versa).

These predictions are confirmed by the predicted response probabilities from the RJM fits – for example, when the stimulus given on the previous trial was the largest one (#10) and the current stimulus is the smallest (#1), the RJM always predicted an incorrect response (#3). Lacouture's participants did not show such behavior. Stimulus #10 was followed by stimulus #1 a total of 21 times, but not once did this elicit response #3. Instead, 17 responses were correct (#1), and the other four were all just one response away (#2). Similar patterns are observed for many other stimulus sequence pairs that involve either very large or very small jumps between successive stimuli, and these result in near chance prediction of the conditional accuracy values by the RJM (RMSE=.44). In contrast, SAMBA fits these same values with RMSE=.17 (Donkin et al., in press), with the misfit due mostly to a failure to capture the asymmetry in the data due to the responses to stimuli #4 and #5 being less accurate than those to stimuli #6 and #7).

Rescuing the Relative Account

The analyses above suggest that purely relative accounts of absolute identification must fail when stimuli are unequally spaced. In this section, we present two ways by which the relative account can better address data from unequally spaced stimuli. However, a side effect of both approaches is an increase the amount of long-term stimulus magnitude information used by the model. In each case, this changes the theoretical account from "purely relative" to either "purely absolute" or a hybrid account which falls somewhere in between the two poles.

Mapping the numeric feedback to stimulus magnitude

In our analysis of the RJM above, Z_n and Z_{n-1} are the physical magnitudes of the current and previous stimuli, measured on a log scale. The difference between these magnitudes is scaled to a difference on the response scale by the parameter λ. Finally, this response-scale difference is added to the feedback given to the participant on the previous trial. This feedback is invariably numeric (one of the digits 1, 2, … , N). For example, in the low-spread condition of our experiment the physical stimulus magnitudes were 73dB, 82dB and 85dB. When stimulus #3 was given, the feedback provided to the subject after their response was the label #3, *not* the physical magnitude of the stimulus (85dB). Stewart et al. (2005) extend the RJM to accommodate unequally-spaced data from Lockhead and Hinson's (1986) experiment by assuming that the feedback provided to the model about the correct answer for the previous trial (i.e., the label #1, #2 or #3) is transformed by the observer back into a physical stimulus magnitude (e.g., they assume that the observer transforms the label #3 back to the magnitude 85dB, or some representation of that).

There are two problems with this extended RJM. The first problem is that the extension is never mentioned in print. The reader naturally assumes the conventional definition of "feedback": the numeral associated with the correct response. Stewart et al. (2005) reinforce this assumption in three ways. Firstly, they cite Holland and Lockhead's (1968) model – which explicitly uses numeric feedback – as a basis for their own. Secondly, Stewart et al. carefully define a different symbol for feedback than the one used for stimulus magnitude. Thirdly, and most explicitly of all, the text above Stewart et al.'s Equation 8 clearly uses numeric feedback (equivalent in this case to stimulus rank S_{n-1}) while simultaneously using psychological magnitudes – different

from stimulus ranks – for stimulus differences (written as $A\ln(r)S_{n-1}$, which is equal to Z_n). This clearly shows that the model Stewart et al. describe is the one we have implemented, not the "extended RJM". It is surprising that the unusual definition of feedback required for the extended RJM is not discussed by Stewart et al.. This omission is particularly surprising because the extended RJM makes a very powerful assumption about the psychological processes in question – that the observer can somehow transform numeric feedback about the stimulus into information about absolute stimulus magnitude.

The second, and more serious, problem with Stewart et al.'s (2005) assumption is that it violates the very heart of their work. On p.892, Stewart et al. write[9]: "What is admitted to the decision process on trial n is not some representation of the magnitude of S_n but a representation of the difference between S_n and S_{n-1}". Allowing the assumption that the feedback label (F_{n-1}) can be transformed by the observer into the stimulus magnitude (Z_{n-1}) perfectly solves the problem of fitting the data for unequally spaced stimulus sets. However, making this assumption directly contradicts the core of their model – that stimulus magnitudes are not admitted to the decision process. On a deeper level, assuming that the observer can transform numerical feedback into a stimulus magnitude is equivalent to assuming that the observer is able to rely on long-term referents that encode the absolute magnitude of each stimulus used in the experiment. This assumption goes against the very core of *all* purely relative accounts

[9] Recall that the RJM is insensitive to arbitrary linear transformations of the psychological stimulus magnitudes, except for the numerical estimates of some parameters. So even though Stewart et al. make this statement in terms of *ranks* (S_n and S_{n-1}) for the equally spaced case, it also applies to (log) stimulus magnitudes (Z_n and Z_{n-1}) in the unequally spaced case.

of absolute identification. Even if the assumption that feedback labels are replaced by stimulus magnitudes can be motivated in some way (e.g., by assuming optimization of performance via learning), the resultant effect is still a code of the absolute magnitude of each stimulus in the experiment.

The use of absolute referents might be justified as an exceptional case for Stewart et al.'s (2005) account, appropriate for Lockhead and Hinson's (1986) experiment because of its within-subjects design. When Lockhead and Hinson's participants experienced the unequally spaced conditions, they may have noted the difference from the equally spaced condition, and stored this information in the form of a set of long-term referents that accurately capture the spacing of the stimuli. Thus, a within-subjects design for unequally spaced experiments may encourage participants to adopt an absolute, rather than relative, approach, mirroring the approach taken by Stewart et al. to modelling these data. Although this is implicitly the theoretical approach taken by Stewart et al., they do not acknowledge that this necessarily makes their model absolute. Of course, the same explanation does not apply to our between-subjects experiment, nor to situations beyond the laboratory where absolute identification is accomplished with unequally spaced stimuli.

Judgment relative to the last two stimuli

A variant of relative models (implemented in the RJM by Stewart, 2007) assumes that sometimes the magnitude of the current stimulus is judged relative to the stimulus that occurred *two* trials previously. If the model uses either the previous or next-to-previous stimulus as a referent, depending on which is closer in magnitude to the current stimulus, a better fit can be obtained to data from both of the experiments we analyze above. The improvement in fit comes about by avoiding the problematic

stimulus sequences described earlier. For example, in Lacouture's (1997) experiment, relative models have difficulty when there is a very large difference between the current and previous stimulus, such as when a stimulus from the smaller sub-group of lines (stimuli #1..#5) follows a stimulus from the larger sub-group of lines (stimuli #6..#10), or vice versa. In Lacouture's experiment, 50% of trials fit this description. On these trials, a model allowed to use the two-back stimulus as a referent can avoid the problematic situation half of the time, because the two-back stimulus has a 50% chance of being from the same sub-group as the current stimulus. Thus, a model using either the prior or two-back stimulus as a referent strikes a problematic stimulus sequence on only one quarter of trials. The model still makes unreasonable predictions for that 25% of trials, but the global average fit is much improved.

The relative account could be even further improved by allowing the access to the previous three stimuli as referents (or four, or five …). However, the core problem would still remain for particular trial sequences. When given a run of stimuli all from the same sub-group of lines (e.g., several trials in succession all using stimuli #1..#5) a relative model will predict very low accuracy if the next stimulus is drawn from the other subgroup. Such runs of stimuli are sufficiently common in data from large experiments that they cannot be ignored.

Discussion

Relative models of absolute identification (e.g., Laming, 1984; Stewart et al., 2005) have explicitly denied the use of long-term memories for stimulus magnitudes. Instead, they are based on a more parsimonious representation of the stimulus magnitudes which uses just a single value that maps differences in stimulus magnitudes

to differences on the response scale. The scaled difference between the stimulus magnitude on the current and previous trial (plus the distortions due to previous differences) is added to the numeric feedback from the previous trial to generate a value on the response scale. We have shown that this approach fails when the spacing between stimuli is unequal. We also explored ways in which theories based on relative judgment can be modified to alleviate the observed problem. These solutions allow relative models to fit the data very well, but they do so by violating the core assumption of relative accounts. Our analyses (see Appendix) show that these problems are not simply due to poor parameter estimates, or to the particular details of the detailed relative model we tested – instead, the problem arises from the core architecture that underlies all relative judgment models.

The success of absolute models and the failure of relative models is due to the fact that the former have a quite complete representation of the stimulus magnitudes and a flexible mapping from the stimulus to the response space, whereas the latter have a representation of differences in stimulus magnitudes and a restrictive mapping from the stimulus to the response space. The long-term referents used in absolute models allow them to flexibly represent different stimulus magnitudes. On the other hand, relative models have explicitly denied such long-term memory elements. Without such referents, relative models are forced to use a greatly simplified stimulus-to-response mapping based on the assumption of a linear relationship between psychological stimulus magnitudes and the numeric feedback values (1, 2, 3, … N). With this assumption, relative models succinctly summarize the stimulus-to-response mapping with just one parameter, for the spacing between psychological stimulus magnitudes from stimuli with adjacent responses. This summary works very well when the

experimenter uses a design with N equally spaced stimuli and the numeric feedback set (1, 2, 3, … N), but breaks down for other stimulus spacing's with the same feedback set. Absolute models do not use such a limited framework. For example, SAMBA does not treat the responses as numbers that can be added and subtracted, but rather as independent labels applied to response accumulators.

It may be possible to escape the above limitation of relative models by using a non-numerical mapping of the type used in SAMBA. A simple version would map the difference between the current and prior stimulus magnitudes, (Z_n-Z_{n-1}), to some response label that was not necessarily a real number. This response label could then be combined with an appropriate transform of the numeric feedback for the previous trial (F_{n-1}) in a cognitive operation that mimics the mathematical operation of addition.

Unfortunately, the same problems we have identified above apply even to such extensions of relative models. In order to appropriately accommodate unequally spaced stimuli, such a model would still require the additional assumption that participants can transform the labels they are given as feedback (e.g., 1, 2, or 3) into stimulus magnitudes (e.g., 58dB, 60dB, 66dB). This transformation can be accomplished by assuming that a long-term referent is maintained for the magnitude of each stimulus, but of course that makes the model absolute, rather than relative. Equivalently, the transformation can be accomplished via a look-up table that remembers the correct response associated with each pair of possible values of {F_{n-1} , (Z_n-Z_{n-1})}. As with all the other versions of relative models that can successfully accommodate data from unequally spaced designs, this is just an absolute referent model by another name. All modifications to relative models that allow them to operate with unequally spaced stimuli work by including in the model a representation of the absolute magnitudes of

the stimuli. This representation can be incorporated in many forms, such as in the look-up table above, or in an assumed transformation between numeric feedback and stimulus magnitude, or even in the location of response criteria. In all cases, the modification includes in the model a very complete representation of the stimulus magnitudes, which runs counter to the basic tenets of relative judgment.

Our results make it clear that relative judgment based on a single scale factor and numeric feedback cannot provide a general account of absolute identification. However, it is possible that absolute identification is accomplished, at least in some cases, via a cognitive process of relative judgment that relies on a set of absolute referents. Indeed, the SAMBA model incorporates just such a relative process, although it was not used in any of the fits presented here, and was required to account for only one of the many benchmark phenomena fit by Brown et al. (2008). Similarly, the extension of the RJM to unequally spaced designs discussed above uses relative judgment in addition to a set of long-term memories for stimulus magnitudes. The success of these hybrid models is interesting, and deserves further investigation, particularly given the strong case that has been made for the general importance of relative judgment in cognition (Chater & Brown, 2008). However, our main point remains – that *purely* relative processes are insufficient to provide a general account of absolute identification.

Acknowledgements

Brown and Heathcote were supported by Australian Research Council Discovery Project DP0881244 and Marley was supported by Natural Science and Engineering Research Council Discovery Grant 8124-98 to the University of Victoria.

References

Brown, S.D., Marley, A.A.J., & Lacouture, Y. (2007). Is absolute identification always relative? *Psychological Review, 114*(2), 528-532.

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.

Chater, N., & Brown, G.D.A. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science, 32*(1), 36-67.

Donkin, C., Brown, S.D., Heathcote, A.J. & Marley, A.A.J. (in press). Dissociating speed and accuracy in absolute identification: The effect of unequal stimulus spacing. *Psychological Research.*

Holland, M. K., & Lockhead, G. R. (1968). Sequential effects in absolute judgments of loudness. *Perception & Psychophysics, 3,* 409-414.

Lacouture, Y. (1997). Bow, range and sequential effects in absolute identification: A response-time analysis. *Psychological Research, 60*(3), 121-133.

Lacouture, Y., & Marley, A. A. J. (1995). A mapping model of bow effects in absolute identification. *Journal of Mathematical Psychology, 39*(4), 383-395.

Lacouture, Y., & Marley, A. A. J. (2004). Choice and response time processes in the identification and categorization of unidimensional stimuli. *Perception and Psychophysics, 66*(7), 1206-1226.

Laming, D. (1984). The relativity of absolute judgments. *British Journal of Mathematical & Statistical Psychology, 37*(2), 152-183.

Lockhead, G. R., & Hinson, J. (1986). Range and sequence effects in judgment. *Perception and Psychophysics, 40*(1), 53-61.

Luce, R. D., & Green, D. M. (1974). The response ratio hypothesis for magnitude estimation. *Journal of Mathematical Psychology, 11*(1), 1–14.

Marley, A. A. J. (1976). A revision of the response ratio hypothesis for magnitude estimation. *Journal of Mathematical Psychology*, *14*(3), 252-254.

Marley, A. A. J., & Cook, V. T. (1984). A fixed rehearsal capacity interpretation of limits on absolute identification performance. *British Journal of Mathematical and Statistical Psychology*, *37*, 136-151.

Petrov, A. A., & Anderson, J. R. (2005). The dynamics of scaling: A memory-based model of category rating and absolute identification. *Psychological Review, 112*(2), 383-416.

Stewart, N. (2007). Absolute identification is relative. *Psychological Review*, *114*, 533-538.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112*(4), 881-911.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91*(1), 68-111.

Appendix

We examine the performance of a canonical relative judgment model for absolute identification with correct feedback that is intended as an abstraction of the major assumptions of the RJM (Stewart et al., 2005) and the theoretical frameworks of Laming (1984) and Holland and Lockhead (1968). One absolute model of absolute identification (SAMBA) also includes a local judgment component (see Brown et al., 2008, p.403-404) that shares some characteristics with relative judgment models. This component does not suffer from the problems outlined below, as it operates on absolute knowledge (including referents for stimulus magnitude estimates).

Consider the "large central gap" condition from Lacouture's (1997) experiment. This was an absolute identification experiment using 10 lines whose psychophysical magnitudes (Z_i) can be represented using arbitrary units as {1,2,3,4,5,11,12,13,14,15}. That is, there were 10 stimuli, with a gap equal to five missing stimuli in the middle. We examine the performance of a simplified, deterministic relative judgment model for this experiment. We require of this model that:

1. There is a parameter $\lambda>0$ that transforms psychophysical differences to numerical differences on the response scale.
2. On each trial, *n*, a response magnitude estimate M_n is produced according to $M_n=F_{n-1}+(Z_n-Z_{n-1})/\lambda$, where Z_n and Z_{n-1} are the log physical magnitudes of the current and previous stimuli and F_{n-1} is the numeric feedback for the previous trial.
3. The magnitude estimate M_n is partitioned into a response by comparison with a set of cut-points $C_0<C_1< C_2< \ldots < C_9<C_{10}$, with $C_0\rightarrow-\infty$ and $C_{10}\rightarrow\infty$. Response *j* is given if and only if $C_{j-1}<M_n<C_j$. (For simplicity of exposition,

we ignore the case where $M_n=C_i$ for some *i*. In any case, this event occurs with probability measure zero in any probabilistic relative judgment model with a continuous response scale.)

A more complete model might include extra components, such as variability in several parameters, and influences from earlier stimulus differences such as $Z_{n-3} - Z_{n-2}$. We do not concern ourselves with these details, as they serve only to decrease model performance. In particular, the magnitude estimate M_n usually has zero-mean noise added to it. By considering just the noise-free estimate, we restrict ourselves to considering the most probable response for any given sequence of stimuli.

We require that the model should produce reasonable predictions for absolute identification data. For any combination of Z_n and Z_{n-1}, the most probable response should be the correct one. In our noise-free model, this means that we require $C_{i-1}<i<C_i$ when *i*, $i \in \{1,\ldots,10,\}$ is the correct response for the stimulus presented on trial *n* .

Lemma 1: $C_{i-1}<i<C_i$ for i=1,..,10

Proof of Lemma 1: Consider the case of a repeated stimulus, where the stimulus with rank i is presented on both the current trial (n) and the previous trial (n-1). The resulting magnitude estimate will be M_n=i, regardless of the value of λ. To ensure that M_n=i is converted into the correct response #i (the integer i), we require that $C_{i-1}<i$ and $C_i>i$. ■

Lemma 2: $\lambda < 4/3$

Proof of Lemma 2: Consider the case Z_n=5 and Z_{n-1}=1. Then $M_n=1+4/\lambda$. For the correct response #5 to be issued, we require $C_4<M_n<C_5$. From Lemma 1, we know that $C_4>4$, so $M_n>4$. After re-arranging, and with our assumption that $\lambda>0$, we arrive at $\lambda < 4/3$, as required. ■

Theorem: If $Z_{n-1}=11$ and $Z_n=5$, the predicted response will be incorrect.

Proof: There is a magnitude difference of -6 units between these stimuli, so the resulting magnitude estimate is $M_n=6-6/\lambda$. Invoking Lemma 2 gives that $M_n<1.5$. Invoking Lemma 1 gives therefore that $M_n<C_2$, so the predicted response is either #1 or #2. Both of these are very different from the correct response #5 (the integer 5). ■

Various other inconsistencies can be obtained in a similar manner, and these inconsistencies can be made arbitrarily large by considering designs with more stimuli and more unequal spacing. The intuition for the problem is that a small value of λ is required to manage the gaps between closely spaced stimulus magnitudes (1-5, and 11-15) but a large λ is required to manage the large gap (5-11).

The approach to fixing this problem taken by Stewart et al. (2005), and discussed in the text above, is to transform the numerical feedback on trial n-1 to the absolute psychological magnitude of the stimulus presented on trial *n-1*. Thus, with R denoting this mapping, the above example requires $R(F_i)=Z_i$, for any i. Step 2 of the relative judgment process is then be replaced by:

2'. On each trial, *n*, a magnitude estimate is produced according to $M_n=R(F_{n-1})+(Z_n-Z_{n-1})/\lambda$, where R is defined above such that $R(F_{n-1})=Z_{n-1}$.

With this adjustment, Lemma 2 does not hold, allowing the modified model to fit the data.

Chapter Four

Absolute Production and Absolute Identification

P. Dodds[1], S. D. Brown[1], V. Zotov[2], S. Shaki[3], A. A. J. Marley[4], A. Heathcote[1]

[1] School of Psychology, University of Newcastle, Australia
[2] DRDC – Toronto
[3] Ariel University Center of Samaria, Israel
[4]Department of Psychology, University of Victoria, Canada

Word Count: 4223
Abstract: 166
References: 22
Tables: 1
Figures: 9

Address correspondence to:
Pennie Dodds
School of Psychology
University of Newcastle
Callaghan NSW 2308
Australia
Ph: (+61)2 4921 6959
Email: Pennie.Dodds@newcastle.edu.au

Abstract

In an absolute identification task, participants are shown a stimulus set (e.g., lines varying in length) each associated with a unique label. Later participants are asked to recall the corresponding label when presented with a stimulus. We studied absolute production, a closely related paradigm requiring the inverse response: our participants were shown a label and asked to produce the corresponding line length. Absolute identification has a long history of study, culminating in comprehensive and detailed quantitative models of performance. These models have proven difficult to distinguish on the basis of identification data, so we investigate whether an extension to absolute production is a fruitful avenue for future development. We demonstrate similarities in the data obtained from the two paradigms, and illustrate how a core theoretical element, common to all identification theories, provides a basis for a theoretical account of production. We develop a mechanistic process based on iterative identification and response adjustment, applicable across models, which allows current identification models to make predictions about production data.

Absolute identification (AI) requires participants to make a stimulus magnitude judgment. Participants are first provided with a set of stimuli that vary in magnitude on a single dimension (e.g., lines varying in length or tones varying in intensity) and each stimulus is given a label. Almost always, the labels are integers, "1" for the smallest stimulus, "2" for the next stimulus, and so on. The participant is then presented on each trial with a randomly selected stimulus from the set and asked to produce the associated label. Despite its seeming simplicity, performance in the task is usually quite poor, with perfect identification limited to 7±2 stimuli (Miller, 1956, but see Rouder, Morey, Cowan & Pflatz, 2004; Dodds, Donkin, Brown & Heathcote, 2011a).

Absolute identification exhibits many complex and interesting phenomena; two that are important here are sequential effects and edge effects. Sequential effects refer to the tendency for decisions to be predictably influenced by stimuli from previous trials. Usually, the response on the current trial is biased towards the stimulus from the previous trial ("assimilation"), but away from stimuli from earlier trials ("contrast"). These phenomena are usually illustrated with impulse plots (Ward & Lockhead, 1970; see Figure 1 for an idealized example) and they provide powerful constraints for theoretical accounts of AI.

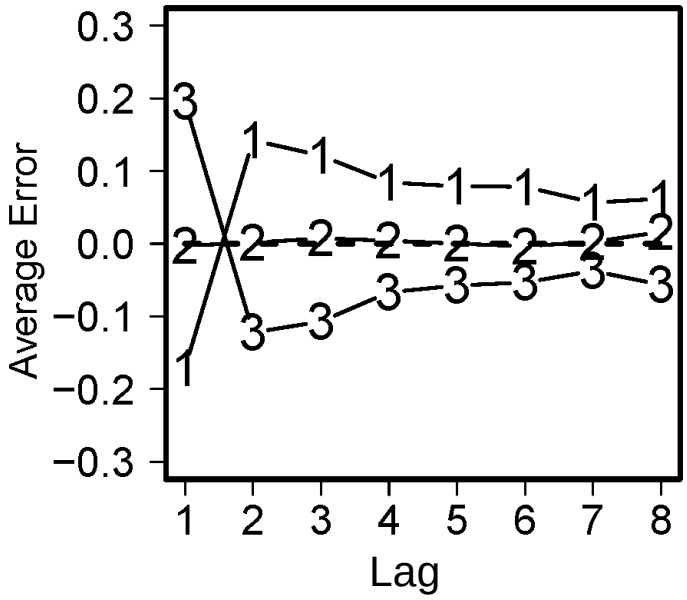


Figure 1. An idealized impulse plot. *Lag* on the x axis refers to the number of trials previous to the current trial on which either a small, medium, or large stimulus was presented (in the figure, lines 1, 2, and 3, respectively).

Edge effects (or bow effects) refer to the tendency for stimuli at the ends of the stimulus range to be identified faster and more accurately than those in the middle of the range. These phenomena are resistant to almost all experimental manipulations, including changes to stimulus spacing (Lacouture, 1997), set size (Stewart, Brown & Chater, 2005), practice, and modality (Dodds et al., 2011a; but see Dodds, Donkin, Brown, Heathcote & Marley, 2011b, for a recently-discovered exception). Figure 2 illustrates a bow effect plot for accuracy from Stewart, Brown and Chater (2005).

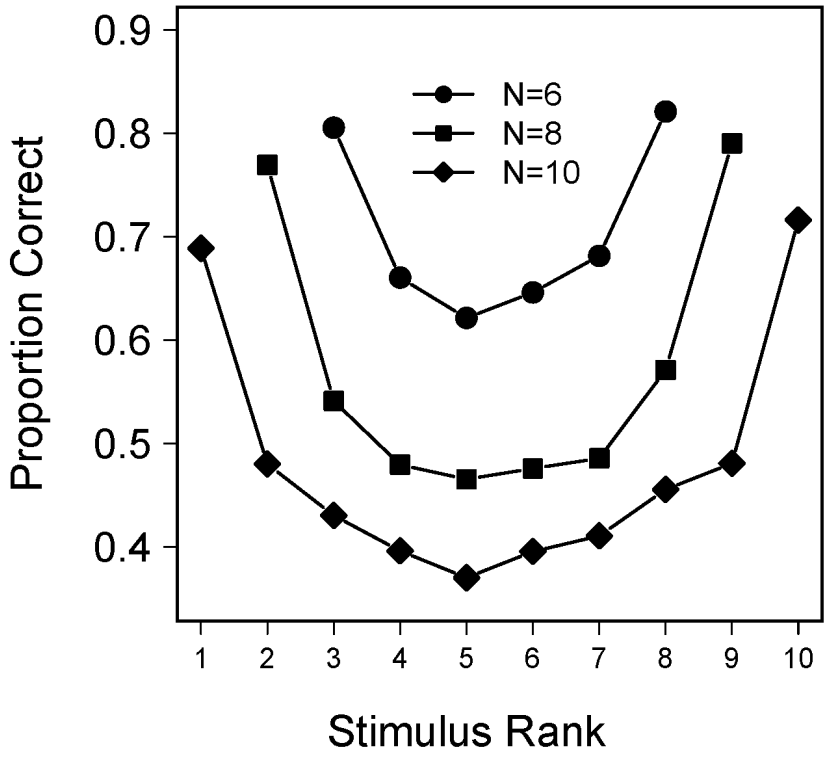


Figure 2. Bow plots for proportion correct, from Stewart, Brown and Chater (2005). The figure shows the resistance of the bow effect to manipulations of the number of stimuli in the set: 6, 8 or 10 stimuli, shown by circles, squares and diamonds, respectively.

The rich variety of reliable empirical phenomena in AI has spurred the development of comprehensive quantitative models of performance. One example, SAMBA (Brown, Marley, Donkin & Heathcote, 2008), combines components of three prior modelling approaches to simultaneously account for many of the standard effects in AI. The first stage of SAMBA involves selective attention, and results an internal estimate of the stimulus magnitude, judged relative to the smallest and largest stimuli.

Later stages of SAMBA transform this magnitude estimate into an identification response.

Like some other current models of AI, SAMBA makes predictions about identification accuracy, the latency of responses, sequential effects on both of these measures, and several other dependent variables. However, it has proven difficult to discriminate between these models on the basis of existing data (e.g., Petrov & Anderson, 2005; Brown et al., 2008; Stewart et al., 2005; Kent & Lamberts, 2005; Brown, Marley, Dodds & Heathcote, 2009). We propose that examination of tasks similar to AI might shed light on the underlying cognitions, as long as the new tasks can be shown to share core elements with AI. *Absolute production* (AP) may provide such an opportunity. The participants' task in AP is the inverse of AI – they are required to produce a stimulus (e.g., a line) when prompted with a label, rather than producing a label when prompted with a stimulus. Similar stimulus production tasks have been used in studies of semantic categorization (Rosch, 1973), memory distortion (Zangwill, 1937), prototype representation (Busemeyer & Myung, 1988), magnitude scaling (DeCarlo & Cross, 1990) and classification (Zotov, Jones & Mewhort, 2011). Related work in psychophysics (e.g., Petrusic, Harrison & Baranski, 2004) has found end effects similar to those in absolute identification, including decreased Weber fractions for extreme stimuli. Relationships between AP and classification tasks similar to AI have revealed good correspondence between production and estimation tasks (DeCarlo & Cross, 1990) and between classification accuracy and category production (Zotov et al., 2011). The continuous nature of AP responses potentially provides benefits over the discrete responses collected in AI experiments. In particular, continuous responses map

more directly onto a crucial element in all AI models – the internal magnitude representation.

Zotov, Shaki and Marley (2010) present the first data to directly examine links between AI and AP. Using lines varying in length, their participants performed a memory and a perception based AP task. Before each task, participants were presented with nine lines of different lengths, labelled 1...9 from shortest to longest. The memory task was absolute production: participants were presented with a randomly selected numeral from 1 to 9, and asked to produce the associated line. The perception task was intended as a control condition, to measure the variability in responses when long term memory was not required. In the perception condition, participants were presented with a randomly selected line from the same set, and asked to reproduce it very soon after it had been removed. Importantly, in both tasks, the response was the inverse of the standard AI task; participants produced (memory task) or reproduced (perceptual task) lines in response to labels. Zotov et al. found that performance for the memory condition, but not the perception condition, closely resembled typical AI performance, consistent with the hypothesis that AI and AP share underlying performance mechanisms.

We propose three extensions to Zotov et al.'s (2010) work. First, we address a methodological issue regarding the stimulus set, which allows for better comparison of our results against standard absolute identification data. Second, we extend Zotov et al.'s data analyses to more fine-grained measures of performance. Finally, we develop a theoretical account of absolute production based on an iterative adjustment mechanism that can be applied generally to existing models of absolute identification.

Experiment

In their absolute production experiment, Zotov et al. (2010) used linear spacing between their stimuli: the difference in length between adjacent lines was constant. Weber's Law asserts, however, that the just noticeable difference between adjacent stimuli is proportional to stimulus magnitude, and so traditionally stimuli are spaced logarithmically in absolute identification tasks: the ratio of the lengths of adjacent lines is constant. Some investigations have also used power-spaced lines, to respect Stevens Law (e.g., Rouder et al., 2004), but linearly spaced stimuli are almost never used. To examine whether the unusual stimulus spacing affected Zotov et al.'s (2010) results, we manipulate stimulus spacing as a between-subjects factor. We also manipulate the stimulus probe as a within subjects factor – each participant was asked in one half of a 2-hour session to reproduce a line from long term memory when prompted with a label, and in the other half asked to reproduce a line that had been shown immediately previously ("memory" and "perception" conditions, respectively). Zotov et al. also manipulated an element of the production method (allowing participants to either start with a random-sized line, or a zero-length line). They concluded this made little difference to the data, so we have used only one of these conditions (zero-length start line).

Participants

Twenty participants from the University of Newcastle took part in this task. The task took about two hours to complete, for which each participant was reimbursed $30.

Stimuli

Stimulus spacing was manipulated between-subjects. In each condition, the stimuli were nine line lengths. In one condition, the lines had lengths from 50 to 850 pixels in 100 pixel increments. In the other condition, lines had lengths from 100 to 900 pixels in 32% increments (100, 132, 174, 228, 300, 394, 520, 684, 900 pixels). The lines were black on a white background. Each was labelled with a number from one through to nine, in order of increasing magnitude. Each participants took part in one line length condition (equal spacing or log spacing). Monitors were 19 inch LCDs with resolution set to 1280 x 1024. Participant viewing distance was not constrained, but was approximately 60 cm from the screen.

Procedure

Participants took part in two different tasks, a perceptual task and a memory task. The order of the tasks was randomized, and they were completed consecutively in a single two-hour testing session (Figure 3 summarizes the experimental conditions). Before each task, all stimuli were presented to participants with labels. The perceptual and memory tasks were identical except for the stimulus that was presented to participants. In the perceptual task, a randomly selected line was presented. In the memory task, a randomly selected label was presented. In each case, the stimulus was presented for one second, followed by a 200ms blank screen. For the perception condition, this was thought to be sufficient time for the decay of visual short-term memory (e.g. Westwood, Heath & Roy, 2003). Then participants were asked to draw on-screen, using a mouse, the line that was associated with the stimulus (namely, a label or a line). The location of the response line was jittered within 10 pixels of the screen centre on every trial. To produce a line, participants used a mouse to enlarge a 3-pixel

by 3-pixel black dot that appeared in the middle of the screen. To enlarge the dot, participants clicked on one edge of the dot and dragged it to the right. Dragging the edge to the right extended the dot in both directions creating a horizontal line; the line was 3 pixels high. When satisfied with their production, the participant clicked a button labelled "Confirm".

No response deadline was imposed. Feedback was presented for one second after production had ceased by displaying the correct stimulus line. The next trial began after another 200ms. Participants took part in 7 blocks of 82 trials in each condition for a total of 1134 trials per testing session. Short breaks were provided in the middle of each block, and at the end of each block. An extended break was provided in the middle of the two conditions.

Experimental Task
(Within Subjects – Task Order Counterbalanced)

Line Spacing **(Between Subjects)**	**Perceptual Task**	**Memory Task**
Log	*Task*: Perceptual Task *Spacing*: Log Spacing	*Task*: Memory Task *Spacing*: Log Spacing
Linear	*Task*: Perceptual Task *Spacing*: Linear Spacing	*Task*: Memory Task *Spacing*: Linear Spacing

Figure 3. Experimental conditions: Line spacing (between subjects) and Experimental task (within subjects).

Results

Data from one participant in the log-spaced condition were lost due to computer failure. For the remaining participants, produced magnitudes were very close – on average – to the correct lengths in all four conditions. Average produced line lengths (calculated across participants), deviated by an average of only 9.6 pixels from the actual response. Figure 4 shows average deviation from the correct response, and illustrates a clear difference between the memory and perception conditions. For both linear and log spacing groups, in the perception condition there was a tendency to overestimate the size of small stimuli and underestimate the size of large stimuli. This tendency was not evident in the memory condition for either group of participants.

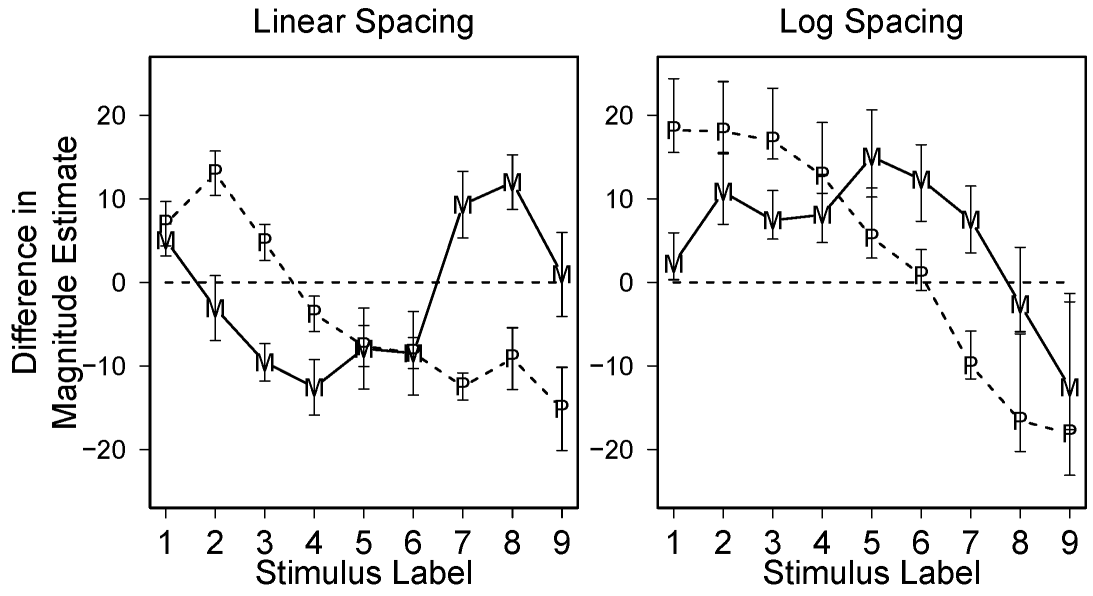


Figure 4. Mean produced line lengths for each stimuli in the Log and Linear conditions, separately for each of the memory (M) and perception (P) tasks. Error bars show Loftus and Masson (1994) standard error across participants for each condition.

Categorised Responses

One way to compare AP data with AI data is to categorise the produced magnitudes according to which prototype stimulus they fall closest to. When categorized this way, production responses can be identified with stimulus labels and therefore treated as identification responses. Figure 5 shows the accuracy of these pseudo-identification responses for each condition, plotted against stimulus magnitude. A three-way mixed design ANOVA (stimulus spacing x experimental task x stimuli) showed that there was no main effect of the spacing conditions (p=0.16). The data for the memory condition are quite similar between the linear and log spacing groups, and also exhibit one of the fundamental phenomena from AI: a bow shape, in which the edge stimuli elicit better performance than central stimuli.

We tested the statistical reliability of this bow shape using linear contrasts that compared the difference in proportion correct for the two shortest lines (1 and 2) against that for the two longest lines (8 and 9), using the mixed ANOVA error term for the effect of stimulus (all p-values reported for linear contrasts are two-tailed). This was used in anticipation of contrasting gradients: that is, a bow shaped curve should result in opposite gradients between stimuli 1 and 2, and stimuli 8 and 9. These contrasts confirmed the bow shaped curve, demonstrated by a significant difference in probability correct between stimulus 1 and stimulus 2, and stimulus 8 and stimulus 9, for the memory task in both linear and log spacing conditions ($F(1,136) = 21.8$, $p < .001$; $F(1.136) = 9.68$, $p = .002$). There was no evidence for a bow in the data from the perception task in the log spaced condition ($p = .9$), but there was a marginally significant difference for the linear condition ($p = .06$). These results are consistent with the hypothesis that the memory condition elicits performance similar to AI.

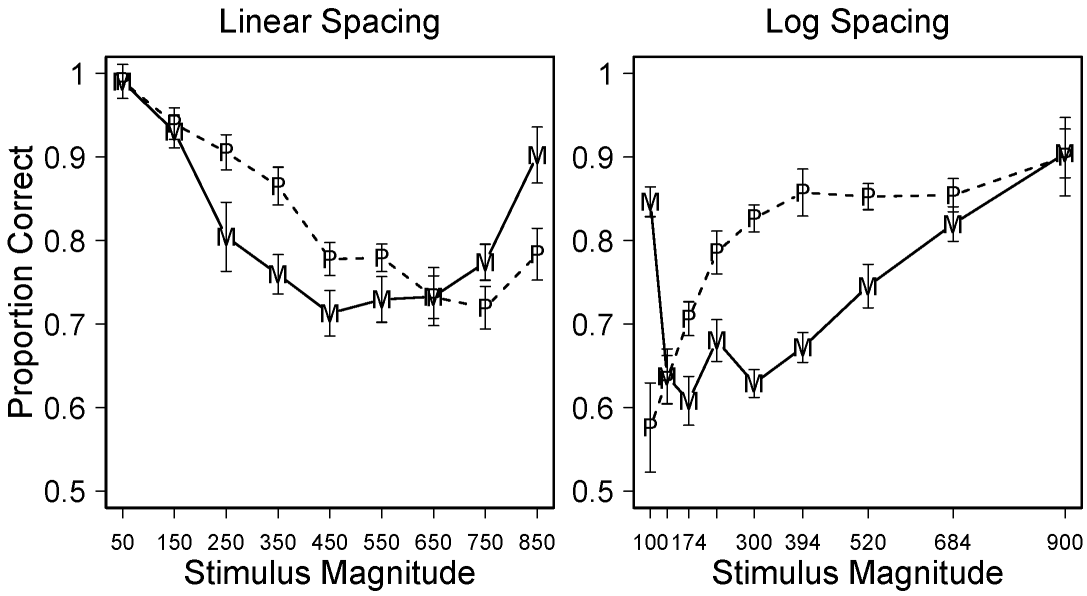


Figure 5. Mean proportion correct for linear and log spacing conditions for each of the memory (M) and perceptual (P) tasks. Error bars are as reported in Figure 4.

Sequential Effects

Returning to the raw data, two other benchmark phenomena from AI are often examined together using impulse plots (see the earlier discussion of Figure 1). In a typical absolute identification task, responses are biased *toward* the stimulus from the previous trial (assimilation) and *away* from stimuli experienced 2-6 trials previously (*contrast*).

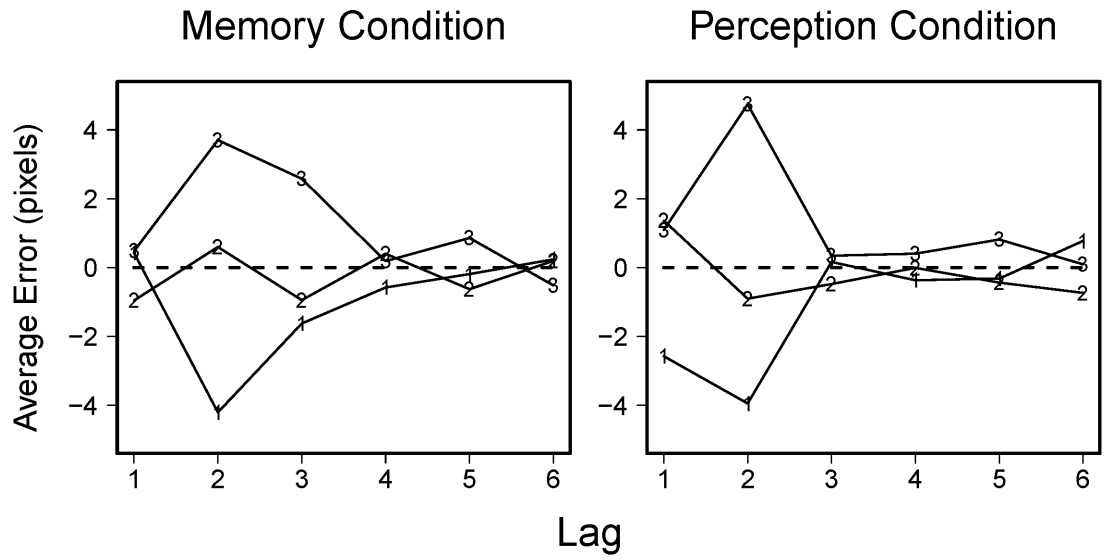


Figure 6. Impulse plots for the memory and perception tasks, for the log and linear spaced conditions combined.

Figure 6 shows impulse plots for the absolute production responses, separately for the memory and perception conditions, collapsed over stimulus spacing. We did not observe the standard AI pattern of sequential effects. Rather, there is little systematic response bias at *lag = 1*, and then assimilation at *lag = 2*. The assimilation decays by *lag = 3* in the perception condition, but persists to at least *lag = 3* in the memory condition. There was no evidence of contrast effects at any lag. These sequential effects are intriguing, but should be interpreted with some caution due to the very high overall task accuracy. As discussed regarding Figure 5, mean error was only a few pixels, and the bias effects in Figure 6 are similarly small, around 4 pixels at their strongest.

The difference between the sequential effects in our data and those in standard absolute identification paradigms (see Figure 1) is striking, but we also note that similar sequential effects have been observed in identification tasks with non-standard designs. For example, Ward and Lockhead (1971) collected data in an absolute identification

paradigm in which participants were not given feedback, and they found two effects similar to ours: no evidence of contrast effects, and also assimilation effects persisting for several trials. Dodds et al. (2011a) also found no contrast effects in the data collected from participants *after* extensive task practice (thousands of trials). In the current experiment, of course, participants were provided with feedback, and did not experience thousands of trials of practice, so it is unclear exactly how our results from absolute production should be related to the earlier results from identification.

Variability of Response Estimates

One of our goals is to relate participants' magnitude estimates more directly to the internal magnitude representations assumed by most major theories of absolute identification (although Stewart et al., 2005, assume an internal representation of the response magnitude, rather than the stimulus magnitude). A key property of all such theoretical representations is their variability. They predict, or assume, that the variability of the internal representation across repeated presentations of the same stimulus is larger for stimuli in the middle of the range than near the ends of the range. It is this assumption that is primarily responsible for producing the bow effects in the models. A major advantage of AP over AI is that it yields the possibility of directly studying such variability in the (internal) representations through the distribution of produced stimuli (here, line lengths) to each label.

To check the properties of the variability in our AP data, we calculated standard deviations separately for each subject and each line length, and separately for each of the four experimental conditions. Figure 7 shows these data averaged over subjects. Note that we trimmed the upper and lower 5% of the data before calculating standard

deviation, to reduce sensitivity of variance measures to extreme outliers produced by some subjects.

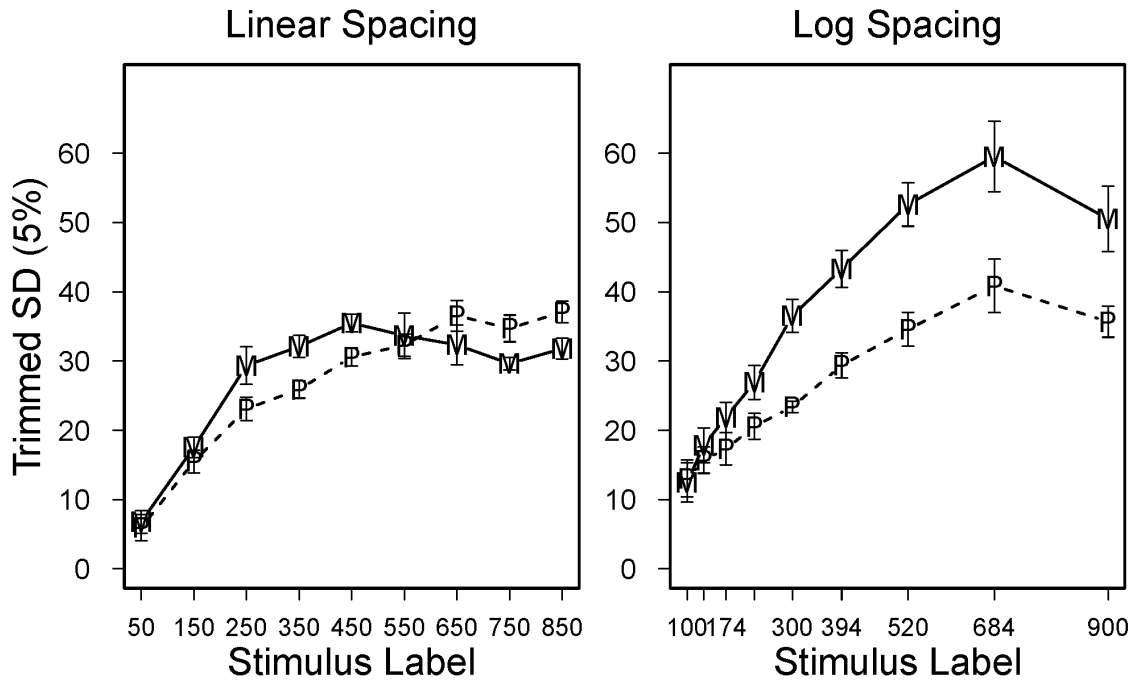


Figure 7. Mean trimmed standard deviation for each of the linear and log spacing conditions and the memory (M) and perception (P) task. Error bars show standard error across participants for each condition.

Mixed ANOVA results showed no significant difference in performance for linear vs. log spacing conditions (p=.43). The same type of linear contrasts applied to pseudo-identification responses (i.e., produced line lengths, categorized as described earlier) confirmed a bow shaped curve in trimmed standard deviation for the memory task in the log spaced condition ($F(1,136) = 5.55$, $p = .02$) but not for the remaining conditions (ps >.1). The non-significant linear contrast for trimmed standard deviation in the linear-spaced group's memory condition is most likely due to the unexpected up-

tick in variability for the largest stimulus label: in that condition, variability in responses decreases steadily from stimulus #5 to #8, but then increases for the largest stimulus (#9). This up-tick might well be due to the higher probability of long outlier responses to the largest stimuli – in support of that notion, we note that further analysis showed that the linear contrast in the linear-spaced group's memory condition approached significance as the proportion of trimming in the standard deviation calculation increased, but the linear-contrasts in the perception conditions remained non-significant. Two participants in particular may have been responsible for outlying data – these two participants showed performance rates considerably lower than other participants (49% and 57%, compared to an average performance of 79%).

Discussion

In several important ways, the absolute production (AP) data in the memory condition closely resembled data from standard absolute identification (AI) experiments. Bow shaped curves were found for accuracy, and for estimates of the standard deviations of line length responses. Just as in typical AI, there was improved performance for stimuli at the edges of the stimulus range, and greater variability in responses, and poorer overall performance, for those in the centre.

Consistent with results from Zotov et al. (2010), data from the AP perception condition were less similar to typical AI performance – accuracy changed monotonically with stimulus magnitude, and the standard deviation measures showed smaller bow effects. As might be expected given the consistent results, there were generally small differences between stimulus sets that were linearly spaced (as used by Zotov et al.) and logarithmically spaced (as is typical in AI).

The data produced by subjects in our perceptual conditions were more complex than would be expected if those responses were influenced only by psychophysical variability in perceived stimulus magnitude. Rather, we suspect that much of the variance in those data might be due to response adjustment mechanisms, which we model below.

Identification and Production: A Point of Contact for Theoretical Accounts

Our experiment suggests some similarities, and also some clear differences, between absolute production and absolute identification. Therefore, although absolute production responses show promise in providing a new source of constraint for modelling, an account of the differences is also required. As a first step, we attempt to quantitatively link absolute production data with the internal magnitude representation of an AI model (SAMBA: Brown et al., 2008). All current models of AI assume such an internal representation – of either stimulus or response magnitude – and these estimates share key properties between the models, most importantly greater variability of the estimates for central than for edge stimuli. Thus, even though our analyses below concern just one model, the results likely have more general implications.

The internal magnitude representation generated by SAMBA in response to a stimulus measures magnitude against a long-term memory for the range of stimuli encountered in the experiment. The magnitude representation is later used to produce an identification response, but this estimate might also be directly compared to responses in an absolute production task. We examine correspondence between the internal magnitude representations from SAMBA and the magnitudes produced by participants in the memory conditions of our experiment. SAMBA's internal magnitude representation takes values on the unit interval, so for comparison with data we re-

scaled its values using the inverse of the logarithmic function used to represent the stimuli in SAMBA's inputs. To more tightly constrain the model, we set several of SAMBA's parameters to values estimated from previous data (those of Lacouture, 1997, and Brown et al., 2008). The precise values of these fixed parameters were not important for the current analyses because they primarily influence data patterns not of interest here, such as response time effects. The free parameters we adjusted to fit the current experiment are reported in Table 1. Different parameters were used for the log-spaced and linear-spaced conditions, as these data came from different groups of participants.

Table 6. SAMBA parameters

SAMBA Parameter	*Linear Spacing*	*Logarithmic Spacing*
Lower Anchor Position (L)	50 pixels	25 pixels
Upper Anchor Position (U)	950 pixels	990 pixels
Perceptual Noise (σ_P)	.005	.008
Rehearsal Capacity (n)	250	150

Figure 8 compares the predictions from SAMBA's (transformed) magnitude representation against the production data. The top two panels show that, when the magnitude estimates from SAMBA are categorized into identification pseudo-responses, as we did previously, SAMBA reproduces response accuracy results from the

memory conditions. The lower two panels show that SAMBA also very accurately captures the variability in those magnitude estimates.

Our intent with the perceptual AP task was to provide a referent for the memory AP task. We hypothesised that the perceptual task data would be less influenced by the context of the stimulus set than the memory task data, allowing us to identify any extra AI-like variability in the memory data. Consistent with added variability, the memory task was performed worse than the perceptual task, with lower categorisation accuracy and higher response variability. Perhaps surprisingly, the perceptual task also showed some clear context effects, including improved performance for end stimuli. We are not aware of other perceptual AP data, or theoretical accounts of this task, so it is not clear what should be made of these results. Given our aim of using production to understand AI, we leave further consideration of perceptual AP to future work.

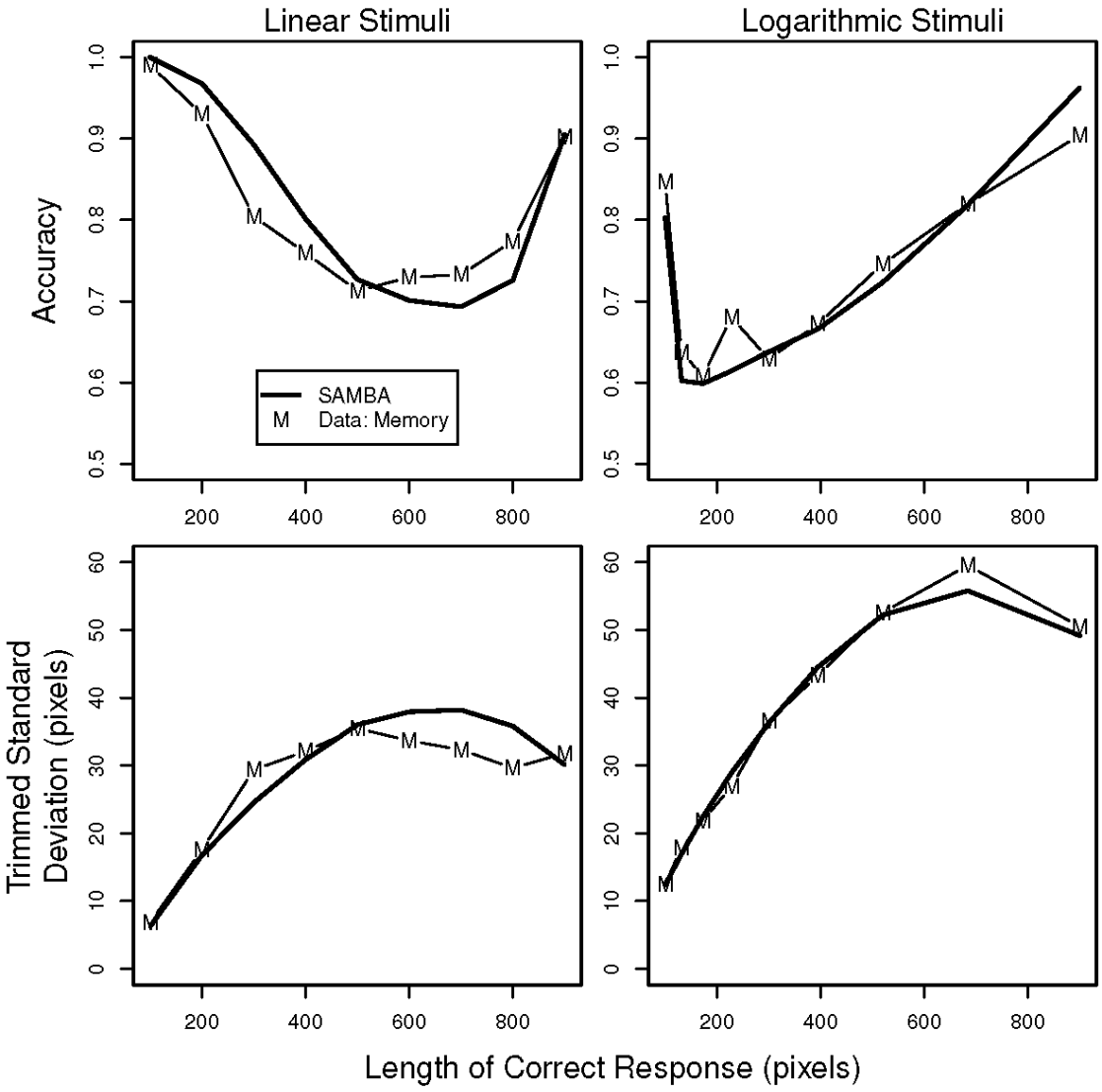


Figure 8. The lines with symbol *M* reproduces memory condition data on response accuracy (from Figure 5) and variability (from Figure 7). Solid black lines show predictions from an absolute identification model (SAMBA).

A Candidate Response Mechanism

Our analyses suggest that the internal magnitude representation from at least one model (SAMBA) provides a link between absolute identification and absolute production, at the conceptual level. This link potentially holds for all comprehensive models of absolute identification, because all such models assume an internal magnitude representation, of either stimulus or response magnitude, with similar properties to SAMBA's. However, missing from our simulations above is a detailed process describing how the internal magnitude representation might be transferred to a physical magnitude produced by participants. We propose a response mechanism that applies generally across identification models to aid the future goal of discriminating theories on the basis of their identification architecture alone.

The response mechanism we propose is based on iterative identification and refinement. For concreteness, suppose an observer in the memory condition of our experiment, with linearly spaced stimuli, was prompted with the stimulus label #7, in which case their goal is to produce a line 650 pixels in length. Our proposed process begins with the production of an initial guess. For our simulations, we assumed an arbitrary initial guess uniformly distributed between the smallest and largest stimulus. This initial guess is then submitted for identification using the standard SAMBA AI model. Although we use SAMBA to model the identification process, any other identification model could be used.

Continuing our example, suppose the initial guess corresponded to a line of length 320 pixels, and the identification process classified this as stimulus #4. The observer then deduces that the initial guess magnitude must be made longer, because the identified label (#4) is smaller than the goal label (#7). An estimate of the required

adjustment to the physical line length can be obtained from the numerical difference between the goal and identified labels, after appropriate scaling. We assume that participants make an adjustment equal to some proportion (a parameter) of the magnitude difference suggested by the goal and identified labels. The scaling parameter for the adjustment procedure is the only free parameter of the response process. We estimated its value at .022 for the subjects in the linearly spaced group and at .037 for subjects in the logarithmically spaced group. Iterative adjustments followed by identifications continue until the candidate magnitude is identified with the goal label.

This response production process also allows AI models to predict the number of adjustments that participants make. Further, AI models that predict AI response times (such as SAMBA; Brown et al. 2008, and EGCM-RT; Lamberts, 2000) will automatically make predictions for response times in absolute production experiments, based on the summed response times for the successive identification processes. Figure 9 shows SAMBA's predictions for the distribution of produced line lengths, along with the data. The predictions and the data are shown using box plots to illustrate that the model does a reasonable job of capturing many aspects of the data – not just the mean.

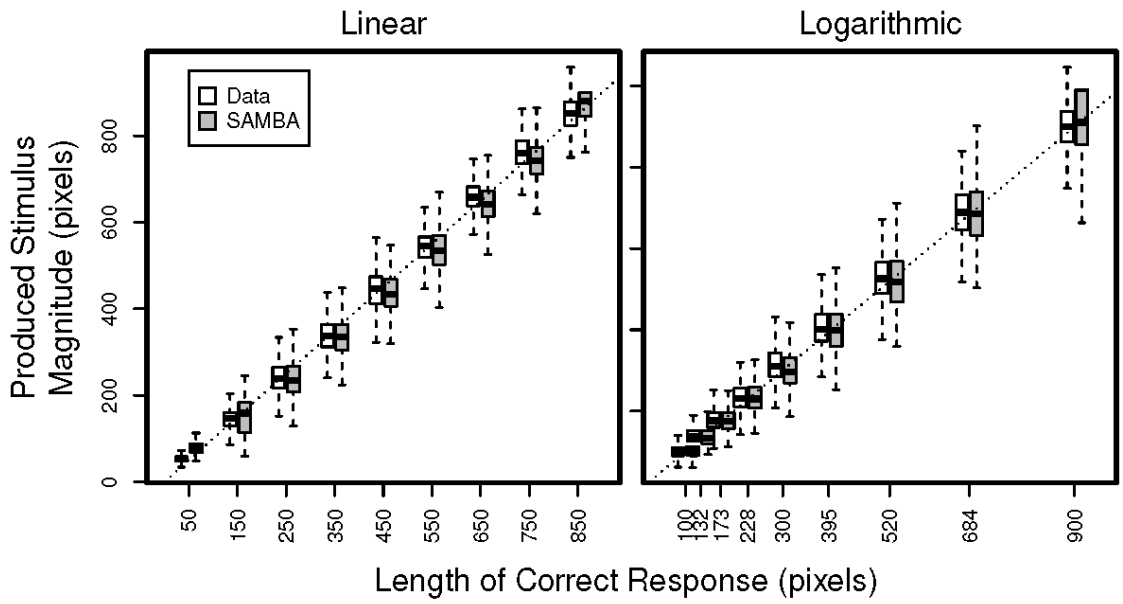


Figure 9. Fits of a iterative version of SAMBA to absolute production data. Box plots show distributions of the response magnitudes over all subjects and trials (data – unfilled) and over an equal number of pseudo-trials simulated from the iterative version of SAMBA (filled). The dotted line is $y=x$.

The proposed response production mechanism offers an insight into why the sequential effects observed in our AP data were so different from those observed in standard AI data. Assimilation effects in AI have sometimes been explained by appeal to response processes. For example, SAMBA explains assimilation by assuming slow decay in the response production mechanism (ballistic evidence accumulators). The radical change we have proposed to the response production mechanism for absolute production will naturally change the predicted sequential effects. For example, SAMBA's account of assimilation effects is mediated by correlations between starting values for response accumulators between trials. The repeated identification processes

required in our production model will remove these correlations, eliminating SAMBA's previous prediction of assimilation at *lag = 1*. It is an open question as to whether those predicted changes agree with data.

Apart from its relationship with absolute identification, our investigation suggests that absolute production is an interesting task in its own right. This task parallels cognitive components of everyday activities such as drawing and construction. Methodologically, one of the most attractive aspects of absolute production is that it avoids the artificial coarseness induced in data when identification (classification) is required. This benefit potentially allows more direct access to the cognitions underlying the psychological representation of magnitude, which is an important and active research question of its own. For example, investigation of the internal representation of magnitude dates back at least to Teghtsoonian's classic (1971) theory of a common magnitude representation for all stimulus types, and has a very active counterpart in modern research on the mental representation of numbers and the number line (Dehaene & Brannon, 2010).

References

Brown, S. D., Marley, A. A., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review , 115* (2), 396-425.

Busemeyer, J. R., & Myung, I. J. (1988). A new method for investigating prototype learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 3-11.

DeCarlo, L. T., & Cross, D. V. (1990). Sequential effects in magnitude scaling: Models and theory. *Journal of Experimental Psychology: General, 119,* 375-396

Dehaene, S. & Brannon, E. M. (2010). Space, time,and number: a Kantian research program. *Trends in Cognitive Sciences, 14,* 517-519

Dodds, P., Donkin, C., Brown, S. D., & Heathcote, A. (2011a). Increasing capacity: Practice effects in absolute identification. *Journal of Experimental Psychology: Learning, Memory and Cognition , 47* (2), 477-492.

Dodds, P., Donkin, C., Brown, S. D., Heathcote, A., & Marley, A. A. (2011b). Stimulus-specific learning: Disrupting the bow effect in absolute identification. *Attention, Perception & Psychophysics , 73* (6), 1977-1986.

Kent, C. & Lamberts, K. (2005). An exemplar account of the bow and set-size effects in absolute identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 289 –305.

Lacouture, Y. (1997). Bow, range, and sequential effects in absolute identification: A response-time analysis. *Psychological Research , 60* (3), 121-133.

Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review , 107,* 227-260.

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within subjects designs. *Psychonomic Bulletin & Review , 1* (4), 476-490.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *The Psychological Review , 63* (2), 81-97.

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.

Petrusic, W. M., Harrison, D. H., & Baranski, J. V. (2004). Long-term memory for elementary visual percepts: Memory psychophysics of context and acquisition effects. *Perception & Psychophysics*, 66, 430-445.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*, 328-350.

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.

Stewart, N., Brown, G. D., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review , 112* (4), 881-911.

Teghtsoonian, R. (1971). On the exponents in Stevens' Law and the constant in Ekman's Law. *Psychological Review*, *78*(1), 71-80.

Ward, L., & Lockhead, G. R. (1970). Sequential effects and memory in category judgments. *Journal of Experimental Psychology*, *84*, 27-34

Ward, L., & Lockhead, G. R. (1971). Response system processes in absolute judgment. *Perception & Psychophysics , 9* (1B), 73-78.

Westwood, D. A., Heath, M., & Roy, E. A. (2003). No evidence for accurate visiomotor memory; systematic and variable error in memory-guided reaching. *Journal of Motor Behavior , 35* (2), 127-133.

Zangwill, O. L. (1937). An investigation of the relation between the processes of reproducing and recognizing simple figures, with special reference to Koffka's trace theory. *British Journal of Psychology*, *27*, 250-276.

Zotov, V, Jones, M. N, & Mewhort, D.J.K.M. (2011). Contrast and assimilation in categorization and exemplar production. *Attention, Perception, & Psychophysics*, *73*, 621-639.

Zotov, V., Shaki, S., & Marley, A. A. (2010). Absolute production as a - possible - method to externalize the properties of context dependent internal representations. In A. V. Bastianelli (Ed.), *Fechner Day 2010. Proceedings of the 26th Annual Meeting of the International Society for Psychophysics* (pp. 203-209). Padua, Italy: The International Society for Psychophysics.

Chapter Five

Multidimensional Scaling Methods for Absolute Identification Data

P. Dodds[1], C. Donkin[2], S.D. Brown[1], A. Heathcote[1]

[1]School of Psychology, University of Newcastle
[2]Department of Psychological and Brain Sciences, Indiana University

Abstract

Absolute identification exposes a fundamental limit in human information processing. Recent studies have shown that this limit might be extended if participants are given sufficient opportunity to practice. An alternative explanation is that the stimuli used – which vary on only one physical dimension – may elicit psychological representations that vary on two (or more) dimensions. Participants may learn to take advantage of this characteristic during practice, thus improving performance. We use multi-dimensional scaling to examine this question, and conclude that despite some evidence towards the existence of two dimensions, a one dimensional account cannot be excluded.

Keywords: absolute identification; unidimensional stimuli; multidimensional scaling; MDS; learning

A typical Absolute Identification (AI) task uses stimuli that vary on only one physical dimension, such as loudness, brightness, or length. These stimuli are first presented to the participant one at a time, each uniquely labeled (e.g. #1 through to *n*). The participant is then presented with random stimuli from the set, without the label, and asked to try and remember the label given to it previously.

This seemingly simple task exhibits many interesting benchmark phenomena. The one of most concern for the current paper is the apparent limitation in performance. The maximum number of stimuli that people were previously thought to be able to perfectly identify was only 7±2 (Miller, 1956). Performance was thought to improve slightly with practice and then reach a low asymptote (Pollack, 1952; Garner 1953).

This finding was particularly surprising given that this limit appeared to be resistant to practice (Garner, 1953; Weber, Green & Luce, 1977), and was generally consistent across a range of modalities (e.g. line length: Lacouture, Li & Marley, 1998; tone frequency: Pollack, 1952; Hartman, 1954; tone loudness: Garner, 1953; Weber, Green & Luce, 1977). In addition, this limitation appears to be unique to unidimensional stimuli. For example, people are able to remember hundreds of faces and names, and dozens of alphabet shapes. It is generally accepted that this is because objects such as faces, names, and letters vary on multiple dimensions. Performance generally increases as the number of dimensions increase (Eriksen & Hake, 1955). This makes intuitive sense when one considers the individual dimensions on a multidimensional object. For example, if people are able to learn to perfectly identify 7 lengths, and 7 widths, they could potentially learn to identify 49 rectangles formed by a combination of lengths and widths.

Despite decades of research confirming this limit in performance for unidimensional stimuli, more recent research has suggested that we may be able to significantly increase this limit through practice (Rouder, Morey, Cowan and Pfaltz, 2004; Dodds, Donkin, Brown & Heathcote, submitted). For example, given approximately 10 hours of practice over 10 days, Dodds et al.'s participants learned to perfectly identify a maximum of 17.5 stimuli (out of a possible 36), a level significantly beyond the 7±2 limit suggested by Miller (1956). From 58 participants that took part in a series of AI tasks, 22 exceeded the upper end of Miller's limit range (nine stimuli).

Other Stimulus Dimensions

The results from Dodds et al. (submitted) were not limited to the identification of lines varying in length. Dodds et al. also used a wide range of other stimuli, and found similar learning effects. For example, dots varying in separation, lines varying in angle and tones varying in pitch all demonstrated similar results. Participants learned to perfectly identify a maximum of 12.6 stimuli using dots varying in separation, 10.4 using lines varying in angle and 17.5 using tones varying in frequency, all exceeding Miller's (1956) upper limit of 9 stimuli.

The learning effects from Rouder et al. (2004) and Dodds et al. (submitted) may be attributed to the type of stimuli employed. The existence of severe limitations in performance is unique to unidimensional stimuli, and since multiple dimensions are commonly associated with improved performance (Eriksen & Hake, 1955) it may be argued that the stimuli vary on multiple dimensions. Tones varying in frequency for example, are generally viewed as multidimensional. While Dodds et al. employed *pure* tones, leaving the stimuli to vary on only one *physical* dimension (wavelength), our perception of loudness increases as a function of increasing frequency. Therefore as

frequency increased, participants would perceive the tones as being of different loudness, creating a greater number of perceived dimensions. This is not an uncommon phenomenon, as a similar effect is found in colour perception. Different colours are generated by a manipulation which is *physically* unidimensional (wavelength change), but the psychological representation of colour is generally considered to consist of three dimensions (e.g., MacLeod, 2003). Therefore it may be possible that the internal psychological representation of different line lengths used in both Rouder et al. (2004) and Dodds et al. (submitted) varied on more than one dimension.

In order to examine this theory using the same stimuli employed by Dodds et al. (submitted), we use Multidimensional Scaling (MDS) methods to examine the structure of similarity ratings generated using these stimuli. MDS refers to a broadly used range of statistical techniques, designed to allow the examination of relationships between objects of interest. Given a matrix of proximity data, MDS uncovers a spatial arrangement of objects in a manner that best reconstructs the original proximity data. For example, given a matrix of data with the distances between *n* cities, MDS analysis would present a spatial 'map' that would arrange the cities in the most likely location, given the distances provided by the data. Because we use subjective "similarity ratings", rather than actual measured distances, we employ non-metric MDS, which does not assume a linear mapping between similarity ratings and distances.

Typically, MDS is employed after one has already assumed the number of dimensions on which the stimuli might vary. In the current experiment however, we use MDS to determine the number of dimensions that best describe Dodds et al.'s (submitted) stimuli.

Method

Participants

The 27 participants, recruited from an introductory psychology course at the University of Newcastle, Australia, took part in exchange for course credit.

Stimuli

Stimuli were 16 lines varying in length (Figure 1). See Table 1 for pixel lengths. Lines were 11 pixels in width and were black, presented on a white background. Stimuli were log spaced, and were separated by a distance substantially greater than the Weber fraction for length (2%; Laming, 1986; Teghtsoonian, 1971).

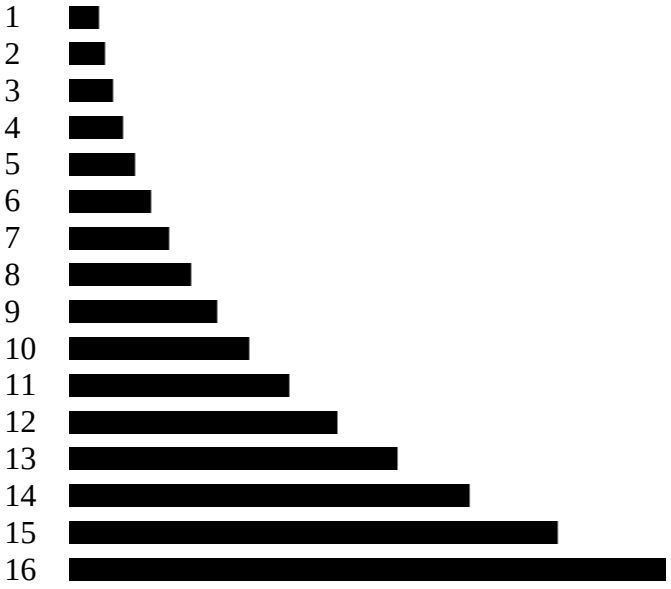


Figure 1. Unidimensional stimuli (line lengths) used in the Experiment. On any single trial, two of these stimuli were presented consecutively. All possible pairs of stimuli, including identical stimuli, were presented twice during the Experiment.

Table 1. Pixel lengths of the 16 lines used as stimuli

Pixel Lengths							
15	18	22	27	33	41	50	61
74	90	110	134	164	200	244	298

Procedure

Participants were instructed to rate the similarity of two stimuli that appeared on a computer monitor, on a scale of 1 to 100. On each trial, a single line would appear on the screen for 1 sec, followed by another line for 1 sec. The position of each line was jittered randomly on every presentation. After the two stimuli had been removed from the screen, a slider panel appeared at the bottom of the screen, allowing the participant to move a scrolling bar along a scale of 1 to 100 (where 1 = *dissimilar* and 100 = *similar*). Every possible pair of stimuli from the set, including identical pairs were presented twice. This resulted in 8 blocks of 64 trials, or a total of 512 trials (i.e., where n=16 stimuli and r=2 replications, number of trials = rn^2). A mandatory 30 sec break was taken between each block.

Each participant was given five practice trials at the beginning of the experiment, where they were asked to complete an identical task to the one above, with the exception that the stimuli were circles varying in diameter. The purpose of the practice trials was only to familiarize the participant with the response method. Different stimuli were used to prevent additional exposure to experimental stimuli.

Results

The main objective of our analysis is to determine whether the stimuli used by Dodds et al. (submitted) are represented internally by one or multiple dimensions. Initial

descriptive analysis suggested that the data were consistent with a one-dimensional explanation: Figure 2 shows the average similarity ratings across participants, plotted as function of stimulus magnitude for each stimulus in the rating pair. Note that identical stimuli are rated as very similar (along the central diagonal), and rated similarity decreases monotonically with the rank-distance between the stimuli (at the left and right corners).

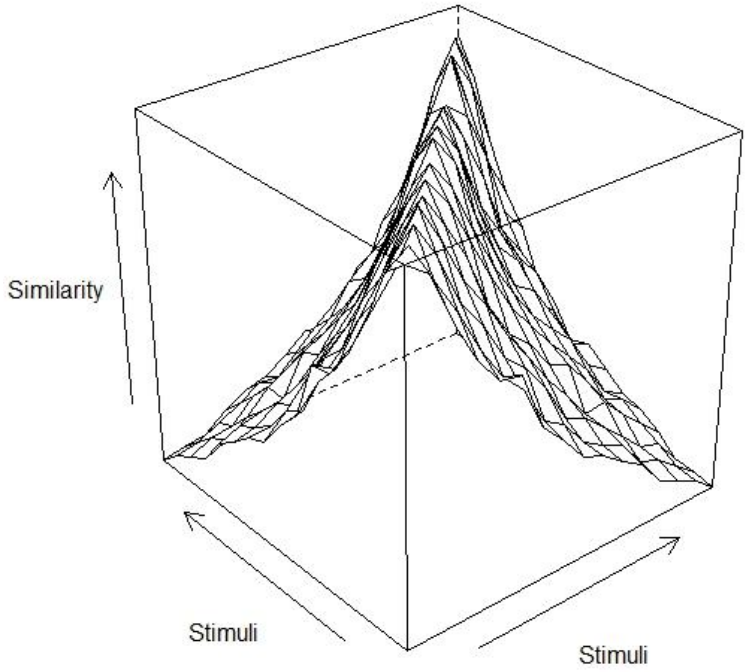


Figure 2. 3D structure of similarity ratings of all 16 stimuli.

Although Figure 2 indicates that the similarity ratings are consistent with a 1D psychological representation, they could nevertheless hide very subtle effects in the data, or large effects for individuals that average out in the group. In order to test this, we calculated non-metric MDS analyses for individual data. Each participant's data were transformed into a single symmetric dissimilarity matrix by subtracting the average similarity rating for each pair of items from 100 and averaging across reversed presentations (e.g., stimulus pair #1-#7 with stimulus pair #7-#1). This matrix was submitted for MDS analyses using both 1D and 2D representations for the data.

Deciding which of the 1D and 2D MDS analyses provides the best account of the data is not trivial. Various ad hoc methods have been used, including examining a goodness of fit measure, or examining the spatial arrangement the points in proximity plots. We applied both methods to our data. In MDS, goodness of fit between the reconstructed and observed dissimilarity matrices is typically measured by sum-squared error, which is called the *stress* value. Smaller stress indicates a better fit; however the MDS models are *nested* meaning that stress must always decrease as more dimensions are included. This means that stress must always be smaller for the 2D than the 1D model. Statistical tests on the magnitude of decrease in stress are not easily constructed, because the key properties of non-metric MDS make it difficult to assume a distributional model for the data. Figure 3 graphs the average stress value, across participants, for MDS fits with dimensions from 1 to 10 (a *scree plot*).

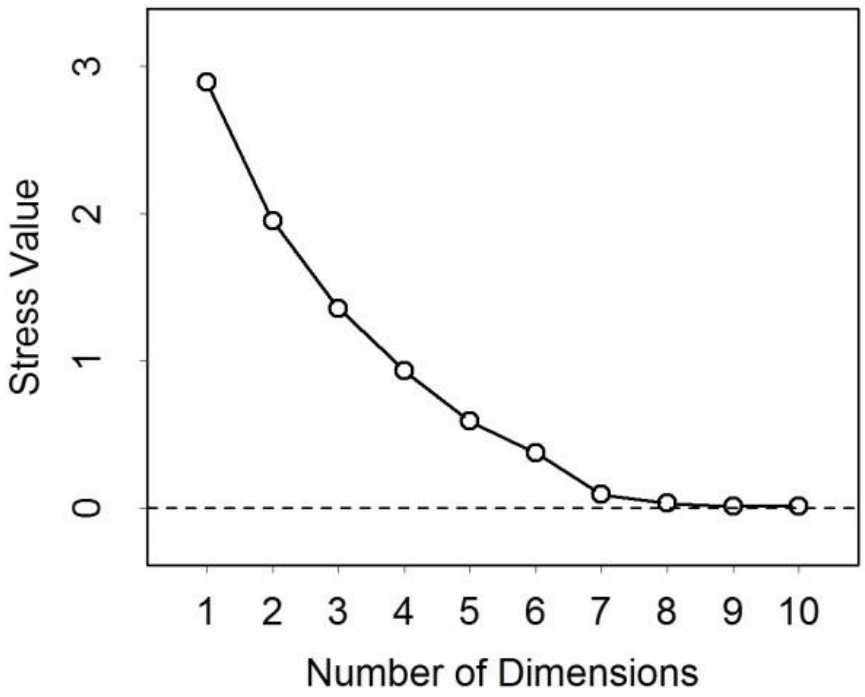
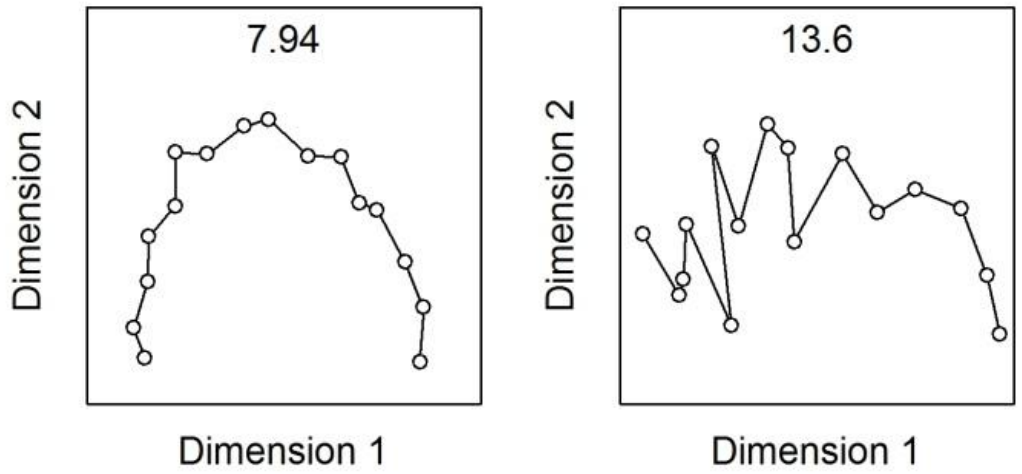


Figure 3. Scree plot showing the decrease in stress value as the number of dimensions increase.

Some authors recommend determining the number of dimensions from a scree plot by finding its "elbow"; a sharp drop in stress value, followed by a relatively flat continuation. Such a pattern could suggest that the latter dimensions fail to provide sufficiently better fit to warrant adding more dimensions to the model. Unfortunately, this method fails to provide any insight into the number of dimensions that best describe the stimuli, as there is no obvious elbow in the scree plot. This is a common problem (e.g., Grau & Nelson, 1988; Lee, 2001). In addition, the use of such methods has been criticized as placing unreasonable emphasis on a numerical measurement. Such methods to determine dimensionality are often used to the exclusion of other, more meaningful aspects of analysis, such as simply the interpretability of results (Shepard, 1974).

A more appropriate method to determine whether a two dimensional model provides a sensible description of the stimuli might be to examine the spatial relationship between objects in the purported 2D psychological space. This can be investigated with a “proximity plot”, where each of the points provided in the similarity matrix are physically arranged in a manner that best satisfies the distances (or similarities) provided in the original data. Figure 4 shows two examples of these proximity plots, for two participants, from MDS analyses with two dimensions.

The philosophy of using MDS to recover internal structure relies on the assumption that, if the psychological representation of the stimuli was truly two dimensional, these 2D MDS proximity plots should reconstruct the internal representation. Because of the nature of the models under consideration (e.g. of categorization and absolute identification), this internal representation should have some relatively smooth and systematic shape. On the contrary, if the internal representation of the stimuli is truly one dimensional, these 2D MDS proximity plots should illustrate the 1D structure (a straight line) possibly along with some meaningless noise.



a b

Figure 4. Two proximity plots of individual fits of a two dimensional model. Each of these graphs is the resulting proximity plot from a single participant in the Experiment. Each point represents a single stimulus in 2D space. Lines connect adjacent stimuli in the set. The value at the top of each graph is the stress value, a goodness of fit measure.

However, these interpretations of the proximity graphs are only appropriate when examining the results of metric MDS analyses (using true, quantitative distances). In the current case, where non-metric MDS analyses must be used, patterns that may normally suggest a two dimensional internal representation, might actually arise from data that are truly *one dimensional*. This problem stems from the monotone transformations allowed by non-metric MDS, between the observed similarity data and the internal psychological distances (as noted originally by Shepard, 1974). Since non-metric MDS analyses only preserve the rank order of the similarity ratings, leaving the

exact similarity values to vary in systematic ways that best suit the data, there is considerable flexibility in the spatial arrangements that might arise from a single underlying dimension. Therefore both Figure 4a and Figure 4b could be construed as evidence favouring a single underlying dimension. Whilst the two proximity plots demonstrate distinctly different patterns, both provide evidence to suggest that our stimuli vary on only a single dimension.

Even though smooth C- or U-shaped proximity plots are *consistent* with one dimensional internal representations, they are also consistent with two dimensional internal representations – that is, truly C- or U-shaped underlying structures. We attempt to resolve this ambiguity using a simulation study comparing MDS outputs from 1D and 2D fits to truly 1D data, in the presence of noise. These simulations provide a metric for interpreting the stress values from our fits to data.

Simulation Study

We investigated this problem of dimensionality with a simulation study. We generated synthetic data from a similarity matrix that was truly one dimensional (the rated distance between each stimulus was a linear function of their ranked difference in the set). We scaled this generating similarity matrix to be as similar to the observed data as possible; we used 16 stimuli, with maximum and minimum similarity ratings of 95.91 and 6.88, respectively. Similarity between stimuli *i* and *j* could then be set as:

$$sim_{max} - (sim_{max} - sim_{min})*(abs(i-j)/15)$$

From this true similarity matrix, we generated synthetic data sets that matched the characteristics of the real data. Noise was added to the matrix using a normal distribution with standard deviation 12.18, and sampled similarity values outside

[0,100] were truncated. These settings resulted in synthetic similarity matrices that were nearly identical to the human data, on average, for the range and variance of similarities, and also for the variance of similarity values across participants, conditioned on each stimulus pair.

We generated 1000 such matrices, and fit each with MDS using both 1D and 2D settings. The lower panel of Figure 5 shows the difference in stress values between these two fits for each simulated data set (negative values indicate a better fit for 2D than 1D).

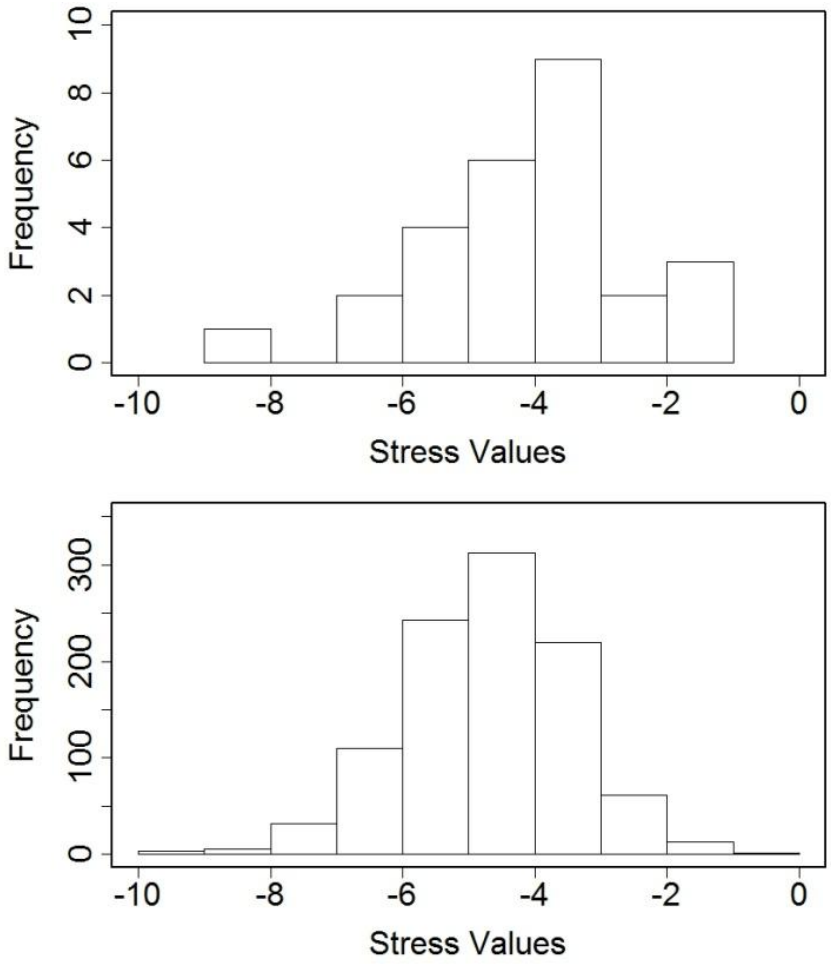


Figure 5. Difference in stress values for between 2D and 1D fits of the original data (top panel) and the true 1D data (bottom panel)

The upper panel of Figure 5 shows the difference between 2D and 1D stress values for the fits to our human data. The important thing to take from these graphs is that the decrease in stress generated by moving from a 1D to a 2D fit is about the same for our human data as it is for our synthetic data. Since the synthetic data were generated by a truly 1D process, this means that the stress values calculated for our human data are entirely consistent with a 1D account. This provides further support to

the evidence provided by the MDS analysis of our own data – that our stimuli may vary on only a single dimension.

Discussion

The purpose of the current experiment was test line-length stimuli commonly used in AI and always assumed to be unidimensional (e.g., Dodds et al., submitted; Rouder et al., 2004; Lacouture & Marley, 1995; Lacouture, 1997). Dodds et al. found that contrary to previous research, their participants were able to substantially improve their performance at the task when given significant practice. Although the stimuli used in their experiment varied on only one physical dimension, the results were more reminiscent of experiments using multiple dimensions, where it is more common to find substantial improvement with practice.

Although the stimuli used in Dodds et al. (submitted) varied on only one physical dimension, it is possible that they may vary on multiple psychological dimensions. In order to examine how many psychological dimensions underpin these stimuli, we used two methods; 1) using MDS techniques we examined similarity data taken using these same stimuli and 2) compared the structure of our data to simulated one dimensional data. MDS proximity graphs suggested that the stimuli may vary on a single dimension, and our simulation study provided further support for this, showing that these fits could be consistent with a one dimensional data generating process, when noise is added.

When examining individual proximity graphs taken from MDS analysis assuming two dimensions, a C (or U) shaped pattern often emerged, which is commonly assumed to provide evidence towards a 2D solution (Shepard, 1974). While this may be appropriate for a metric MDS analysis, the monotonic transformations unique to *non-*

metric MDS allow some flexibility in the position of the objects in the final proximity graph. Despite this difference required in interpretation of metric vs. non-metric proximity graphs, it is possible that the two types of proximity graphs generated by our data (Figure 4) were genuinely representative of one vs. multiple dimensions, and that the action of specifying the number of dimensions to examine, forces the model to fit, sporadically producing evidence for and against a two dimensional solution. In support of a one dimensional solution however, our simulated data demonstrate a similar structure to our original similarity data, suggesting that the stimuli used in Dodds et al. (submitted) vary on only a single dimension.

Therefore it appears that the interpretation of MDS output for the number of underlying dimensions in the data is difficult. While we were able to gather evidence using a variety of techniques to suggest that our data were consistent with a single dimension, MDS could not provide a definitive answer. Lee (2001) showed that it is possible to reliably determine dimensionality from MDS analysis, but only when the determination is between larger numbers of dimensions. Like us, he found much poorer reliability when the choice was between lower numbers of dimensions. Hence, the task of choosing between a low number of dimensions remains very subjective, and users should take care not be misled by "overfitting", where a complex model imitates data from a simpler underlying data generating process. Furthermore, in the case of determining dimensionality, one should take care not to focus solely on quantitative results such as the stress value, but also take into consideration the pattern of data in the original similarity matrix (such as in Figure 2) or even simply the interpretability of results (Shepard, 1974).

Both the MDS analysis of the similarity data for Dodds et al.'s (submitted) lines of varying length and our simulation study were consistent with a 1D psychological representation. This finding makes it less likely that the substantial improvement with practice observed by Rouder et al. (2004) and Dodds et al. (submitted) in absolute identification of line lengths was due to participants learning to take advantage of a multi-dimensional psychological representation. This finding may also extend to the other stimuli that Dodds et al. employed. Similar learning effects to that of lines varying in length suggest that modality, or specifically, the number of dimensions that stimuli vary within, cannot be the sole cause of the improvement in performance. Hence, investigation of alternative explanations for the improvement they observed seems warranted.

References

Dodds, P., Donkin, C., Brown, S. D., Heathcote, A. *Practice Effects in absolute identification*. Manuscript submitted for publication

Eriksen, C. W., & Hake, H. W. (1955). Multidimensional stimulus differences and accuracy of discrimination. *Journal of Experimental Psychology, 50*(3), 153-160.

Garner, W. R. (1953). An informational analysis of absolute judgments of loudness. *Journal of Experimental Psychology, 46*(5), 373-380.

Grau, J. W., & Nelson, D. G. K. (1988). The distinction between integral and separable dimensions: evidence for the integrality of pitch and loudness. *Journal of Experimental Psychology: General, 117*(4), 347-370.

Hartman, E. B. (1954). The influence of practice and pitch distance between tones on the absolute identification of pitch. *The American Journal of Psychology, 67*(1), 1-14.

Lacouture, Y. (1997). Bow, range, and sequential effects in absolute identification: A response-time analysis. Psychological Research, 60, 121-133.

Lacouture, Y., & Marley, A. A. J. (1995). A mapping model of bow effects in absolute identification. Journal of Mathematical Psychology, 39, 383-395.

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.

Laming, D. (1986). Sensory Analysis. London: Academic Press.

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modelling. *Journal of Mathematical Psychology, 45*, 149-166.

MacLeod, D. I. A. (2003). New dimensions in color perception. *Trends in Cognitive Sciences, 7*(3), 97-99.

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

Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America, 24*(6), 745-749.

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.

Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, *39*(4), 373-421.

Teghtsoonian, R. (1971). On the exponents in Stevens' Law and the constant in Ekman's Law. Psychological Review, 78(1), 71-80.

Weber, D. L., Green, D. M., & Luce, R. D. (1977). Effects of practice and distribution of auditory signals on absolute identification. *Perception and Psychophysics, 22*(3), 223-231

Chapter Six

Perhaps Unidimensional is not Unidimensional

Pennie Dodds, Babette Rae and Scott Brown

University of Newcastle, Australia

Counts
Abstract: 140
Body: 3958
References: 22
Figures: 2
Tables: 2

Address correspondence to:
Pennie Dodds
School of Psychology
University of Newcastle
Callaghan NSW 2308
Australia
Ph: (+61)2 4921 6959
Email: Pennie.Dodds@newcastle.edu.au

Abstract

Miller (1965) identified his famous limit of 7±2 items based in part on absolute identification – the ability to recognize stimuli which differ on a single dimension, such as lines of different length. An important aspect of this limit is its independence from perceptual effects and its application across all stimulus types. Recent research however, has identified several exceptions. We investigate an explanation for these results which can reconcile them with Miller's work. We find support for the hypothesis that the exceptional stimulus types have more complex psychological representations, which can therefore support better identification. Our investigation uses data sets with thousands of observations for each participant, which allows the application of a new technique for identifying psychological representations: the structural forms algorithm of Kemp and Tenenbaum (2008). This algorithm supports inferences not possible with previous techniques, such as multi-dimensional scaling.

Absolute identification (AI) is the fundamental task of identifying stimuli that vary only on one physical dimension. For example, tone frequency (e.g. Hartman, 1954; Pollack, 1952), tone loudness (e.g. Garner, 1953) or line length (e.g. Lacouture, 1997). In a typical AI task, stimuli are first presented to the participant one at a time, each with a unique label. In the test phase, the participant is then presented with randomly selected stimuli from the set and asked to recall the associated labels.

Miller's (1956) classic paper investigated limits in both short term memory and in AI, and found that 7 ± 2 was not only the number of chunks that can be held in short-term memory, but was also the number of items people could learn to perfectly identify in such a unidimensional stimulus set. The upper limit of Miller's range (nine stimuli) is particularly surprising because it is resistant to many experimental manipulations, including extensive practice (e.g. Weber, Green & Luce, 1977), the number of stimuli in the set (e.g. Garner, 1953) and stimulus spacing (e.g. Braida & Durlach, 1972). Most importantly, this limit appeared to be a fundamental aspect of human information processing rather than a sensory limitation, because the same limit applied to a wide range of stimulus modalities (from electric shocks to saltiness: e.g., Lacouture, Li & Marley, 1998; Pollack, 1952; Garner, 1953)

Despite this longstanding assumption that uni-dimensional stimuli are unable to be learned beyond an upper limit, recent work has identified exceptions. One of Rouder, Morey, Cowan and Pfaltz's (2004) participants was able to learn to perfectly identify 20 line lengths. Dodds, Donkin, Brown & Heathcote (2011) reported related learning effects not only for line lengths, but also for dot separation, line angle, and tone frequency. These findings contradict Miller's theory of a small upper limit to memory processing capacity. This could represent an important finding because a small, or null,

effect of learning has been included as a central element of many theoretical accounts of AI and memory (including: Stewart, Brown & Chater's, 2005; Petrov & Anderson's, 2004; Marley & Cook's, 1984; and Brown, Marley, Donkin & Heathcote's, 2008). If Dodds et al.'s (2011) and Rouder et al.'s (2004) results are taken at face value, they might imply that supposedly fundamental capacity constraints can be altered by practice.

There is, however, an alternative explanation. The number of stimuli that can be reliably identified increases exponentially as the number of dimensions increase (Eriksen & Hake, 1955; Miller, 1956; Rouder, 2001), at least when those dimensions can be perceived independently ("separable" dimensions: Nosofsky & Palmeri, 1996). For example, people are able to identify hundreds of faces, names and letters, all of which vary on multiple dimensions. Or, if an observer could perfectly identify say, seven line lengths and also seven angles, they might be able to identify 49 different stimuli with these *combined* features, such as circle sectors. With an additional assumption, this line of reasoning might reconcile the learning effects observed by Rouder et al. (2004) and Dodds et al. (2011) with the long-standing results of Miller (1956). The extra assumption that is required is that some stimulus sets which vary on just one physical dimension might nevertheless invoke a more complex psychological representation. As with physically multi-dimensional stimuli, more complex psychological representations support richer percepts, perhaps allowing multiple ways to estimate the magnitude of a stimulus and hence better identification.

The stimuli used in AI always vary on just one physical dimension, but this does not guarantee that the corresponding psychological representations are uni-dimensional continua. For example, perceived hue is represented either on a circle or a disc

(Shepard, 1962; MacLeod, 2003) and the psychological representation of pitch is a helix (Bachem, 1950) even though the corresponding physical stimuli vary on only one dimension (wavelength, in both cases). In Dodds et al.'s (2011) and Rouder et al.'s (2004) studies, it might have been that those exceptional observers who learned to identify stimuli beyond Miller's limit managed this feat by constructing more complex psychological representations for the unidimensional stimuli. If these observers had access to percepts on dimensions that are even partially independent, this could explain their improved performance without challenging Miller's long-standing hypothesis that performance on any *single* dimension is severely limited.

Examining Psychological Representation

In the absence of additional evidence, there is an unsatisfying circularity to this argument. The only evidence that suggests that these physically unidimensional stimuli have more complex psychological representations, is that those same stimuli can be learned. The only tested prediction from the hypothesised complex representation is that those same stimuli can be learned well. One method of independently probing psychological representation is to use multidimensional scaling (MDS; Cox & Cox, 1993; 2001). MDS determines relationships between objects by examining estimates of the perceived similarity of pairs of the objects. In some cases, such as with colour, MDS techniques are able to reliably infer the complex psychological representation extracted from apparently unidimensional stimuli. This success presumably depends on the clear and consistent form of the representation across different people – allowing data to be averaged across subjects. In turn, the consistency of the psychological representation across subjects is probably an upshot of the basic physiology of the retina. In less clear-

cut cases MDS is not always sensitive to subtle or inconsistent changes in the form of psychological representations.

Dodds, Donkin, Brown and Heathcote (2010) collected similarity ratings for line lengths, which was one of the stimulus types that Rouder et al. (2004) and Dodds et al. (2011) identified as an exception to Miller's (1956) limit. Dodds et al. (2010) found that MDS was not reliably able to distinguish between one- and two-dimensional representations. The problem is that it lacks a framework for inference about these arrangements. This means that, if one wishes to recover the number of dimensions that best represent a relationship between objects, the conclusions are based on subjective judgements. Lee (2001) investigated this problem in detail and found that, for one- or two-dimensional representations, MDS correctly identified the number of dimensions only 14% of the time.

A recent advance in estimating the structure of psychological representations provides an alternative to MDS. Kemp and Tenenbaum (2008) developed an algorithm that to infer the structure of psychological representations based on relational data. Their method is based on a universal grammar for generating graphs, and the generality of those graphs allows the algorithm to represent structures as varied as trees, hierarchies, and points in vector spaces (as in MDS). An important benefit of Kemp and Tenenbaum's algorithm is that it includes a coherent framework for inference, allowing probabilistic comparison of different structural forms based on penalized likelihood, where the penalty term depends on structural complexity.

We use Kemp and Tenenbaum’s (2008) algorithm to investigate the psychological representation of the stimuli used in AI experiments. We limited our search to undirected graph structures only, on the assumption that the similarity of two

stimuli should not depend on the order of comparison (or, if it did, that this dependence was not of primary interest). We also limited our search to just two of Kemp and Tenenbaum's forms – the chain and ring (see Figure 1). Chain structures are the standard assumptions for AI stimuli: one-dimensional continua, where the psychological distance between stimuli is found by summing the distance from one neighbour to the next, and the next again, and so on. Ring structures represent just a small increase in complexity from chains, capturing the additional property that stimuli near one end of the set might be perceived to have something in common with stimuli at the extreme other end. This kind of relationship is found in both of the well-known cases of physically unidimensional stimuli having multidimensional psychological representations: long wavelength light has a perceived hue (red) which is similar to the hue perceived for short wavelength light (violet); similarly, the lowest frequency note in an octave (A) is perceived as similar to the highest (G#).

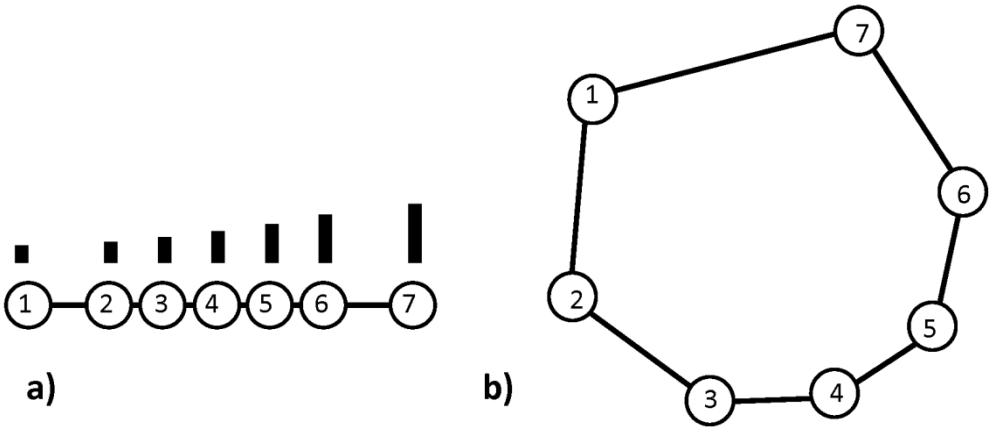


Figure 1. An illustration of a (a) chain structure and a (b) ring structure for lines of varying length. Note that in the chain structure stimuli 1 and 7 are far apart, while in the ring structure they are much closer.

Data

A direct way to investigate psychological structure relies on similarity estimates obtained by direct interrogation: participants are presented with two stimuli and asked to rate their similarity on some scale. Such ratings have many problems. Firstly, there is a severe limit on sample size, because participants find it difficult to give many repetitions of these responses. Secondly, the numerical similarity ratings provided by participants depend on the experimenter's choices. For example, different ratings would be provided if the observers are asked to rate similarity from 1-10 or from 0-100, or on a Likert scale, and the precise nature of this dependence is unclear. Even more troubling, it is unclear whether similarity ratings obtained by this method are based on the particular psychological representation of interest: the one underlying AI performance. To circumvent all three problems, we replace similarity judgments with confusion matrices calculated from many thousands of AI trials. These confusion matrices encode how often each pair of stimuli are confused with each other (e.g., when stimulus A is presented, what is the probability that it is identified as stimulus B?). Our assumption is thus that the probability of confusing two stimuli is monotonically related to their similarity.

We calculated confusion matrices using the data from four AI experiments reported by Dodds et al. (2011; see Table 1). In all four experiments, participants were given extensive practice over a series of 10 sessions, leading to around 5,000 observations per participant. Each experiment included five or six participants. Three of the experiments used a smaller number of stimuli (15 or 16) allowing for unconfounded comparison between different stimulus types. These three experiments included the only one in which participants did not exceed Miller's (1956) limit of 7±2 stimuli (tone

intensity) and two in which they did (line length and dot separation). The other experiment used 30 line lengths. We included this experiment because it showed some of the greatest improvement in performance with practice. We are wary of direct comparison between smaller and larger set size experiments because of the varying statistical reliability of the data sets. The larger set sizes resulted in one quarter as many observations contributing to each element of the confusion matrix – as few as three observations per matrix element. The penalized complexity used in Kemp and Tenenbaum's (2008) algorithm means that noisier data lead to a preference for simpler structures – a bias towards identifying chain structures, in our case. This might extend to our smaller set size experiments, because even with 5,000 observations in the smaller set sizes, the average number of observations contributing to each confusion matrix element was between 16 and 20. We return to this point in the Discussion.

We also analysed data from a new AI experiment using tone frequency. For this experiment, we gave six musically- trained participants practice with a set of 36 pure sine tones varying in frequency – their frequencies matched the fundamental frequency of the standard piano notes from A3 to G#6. The procedure for this experiment was similar to the procedure outlined for Experiment 6 in Dodds et al. (2011), except that responses were labelled not only with a number (1-36) but also the corresponding piano note name. There were 10 learning sessions providing 4860 identifications per participant.

Results

Table 1. Data sets used from Dodds et al. (2011)

Experiment*	Stimuli	Set Size
1a	Line Length	30
2b	Dot Separation	15
5a	Line Length	16
5b	Tone Intensity	16

* Note: Experiment refers to the experiment number as listed in Dodds et al. (2011)

Confusion matrices were constructed for individual participants, for a) their entire 10 hours of practice, b) the first 5 sessions of practice and c) the last 5 sessions. Data for individual participants were used as opposed to averaged data because of the small number of participants and large individual variation (see Table 2 for variation in accuracy). As described by Kemp and Tenenbaum (2008), feature data were simulated from the confusion matrices.

We used v1.0 (July 2008) of the Matlab implementation of the structural forms algorithm (obtained from the first author's website). For each confusion matrix, we identified the best chain and best ring structure, and recorded their penalized likelihoods. In all cases, we used default values for the algorithm's parameters. We made one modification to the algorithm, for numerical stability, restricting the search over edge lengths to disallow lengths that were extremely close to zero (smaller than e^{-10}). Note that this restriction still permits edge lengths of precisely zero, because adjacent stimuli can be collapsed into single nodes using the rules of the graph grammar. Our restriction only disallows extremely small, but non-zero, separation

between stimuli. In Table 2, we report differences in penalized log-likelihood between the chain and ring structure fits. To put the likelihood results in statistical perspective, differences in log-likelihood can be used to approximate the posterior probability that one model out of the pair (chain or ring) was the data generating model. This approximation should be interpreted with some care, as it relies upon some strong assumptions - for example, that the data generating model was one of the pair under consideration (e.g., Raftery, 1995). Nevertheless, using this interpretation, a difference in log-likelihood of two units corresponds to about three-to-one odds in favour of one model over the other, and a difference of six units in log-likelihood to better than twenty-to-one odds.

Whole Data Sets

We report the analyses for smaller set sizes (15 or 16 stimuli) separately from the larger set sizes (30 or 36 stimuli). This allows cleaner comparison within each group because the number of data per entry in the confusion matrices are comparable: about 18 observations per entry for the small set sizes, and about 4 for the large set sizes.

Small set sizes. Small set size experiments included those that used 16 tones of increasing loudness, 16 line lengths and 15 dots varying in separation. Participants who practiced tone loudness did not improve their performance much with practice, and their confusion matrices were also unanimously better described by chain structures than ring structures (see Table 2, where positive log-likelihood differences imply support for chain structures over ring structures). These results are consistent with Miller's (1956) original hypothesis that AI is subject to a severe capacity limit when the stimuli really are unidimensional.

In comparison, the confusion matrices for some of the participants who practiced 16 line lengths and 15 dot separations exhibited were better described by the ring structure than the chain. In these two experiments, the ring structure was deemed more likely for only about half of the participants (5 of 11; see Table 2). The support for a ring structure is even more surprising when considering the data used in these experiments: our use of confusion matrices rather than similarity ratings. For example, if asked for a similarity rating, a participant might rate the extreme edge stimuli as very similar, but they still might be very unlikely to confuse those stimuli in an identification experiment. This presumably biases our results towards the chain structure, and yet several participants were still better described by ring structures.

Those five participants for whom the ring provided a better description in these experiments also demonstrated higher initial identification performance, and more improvement with practice. At the beginning of practice (first session), their mean accuracy was 54%, compared with 44% for the participants better described by chain structures, and over the course of practice, those subjects identified as having ring-like representations improved their identification performance by 32% compared with 29% for the chain-like participants. We are hesitant to calculate inferential tests on these differences due to the very small number of participants (five in one group, six in the other).

Large set sizes. Table 2 shows accuracy and log-likelihood differences for experiments with 30 line lengths and 36 tone frequencies. Four of the twelve participants demonstrated greater likelihood for a ring structure compared to the chain structure. As with the smaller set size experiments, those who demonstrated a ring structure demonstrated greater improvement in performance (Mring = 0.36) compared

to those that demonstrated a chain structure (Mchain = 0.22) and also greater pre-practice performance (Mring = 0.36, Mchain = 0.22).

Table 2. Accuracy and log-likelihood values for each participant in each of the five experiments.

Experiment (Stimuli)	Participant	Initial Accuracy	Improvement in Accuracy	Overall Log-likelihood Difference *	Early Log-likelihood Difference *	Late Log-likelihood Difference *
Tone Loudness (16)	1	0.34	0.1	8.357	19.326	12.25
	2	0.3	0.19	20.092	18.705	25.209
	3	0.33	0.12	12.508	26.442	17.106
	4	0.27	0.09	24.007	27.554	20.384
	5	0.34	0.08	26.205	23.19	22.165
	6	0.31	0.13	19.078	19.527	17.521
Line Lengths (16)	1	0.58	0.22	-2.542	0.12	-5.116
	2	0.51	0.38	-3.423	0.171	-4.711
	3	0.41	0.31	7.041	12.733	4.816
	4	0.4	0.28	3.344	9.77	-0.223
	5	0.46	0.30	5.143	13.69	-2.046
	6	0.57	0.24	-3.365	-3.874	-4.798
Dot Separation (15)	1	0.44	0.2	7.852	10.378	8.621
	2	0.51	0.37	-0.877	4.088	-5.481
	3	0.53	0.27	4.472	7.229	2.468
	4	0.39	0.37	5.703	11.38	-12.56
	5	0.53	0.41	-2.658	2.36	-1.536
Line Length (30)	1	0.21	0.26	24.556		
	2	0.18	0.1	13.405		
	3	0.29	0.47	-1.426		
	4	0.2	0.1	24.292		
	5	0.31	0.41	57.852		
	6	0.17	0.24	-22.265		
Tone Frequency (36)	1	0.4	0.29	-4.273		
	2	0.59	0.31	-0.394		
	3	0.2	0.13	27.8		
	4	0.21	0.09	24.934		
	5	0.19	0.08	22.549		
	6	0.22	0.14	0.634		

* Note that difference values are calculated by subtracting the likelihood values for ring structures from the likelihood values for chain structures.

Effect of Practice

Dodds et al. (2011) noted that participants improved their performance markedly given practice at AI for all stimulus sets except for tones varying in intensity. In order to examine whether the improvement in performance was associated with a change in psychological structure, we also examined the confusion matrices for each participant in the small set size experiments separately for early (1:5) and late (6:10) practice sessions (See Table 2). We did not examine this split in the data from the large set size experiments because the sample size was too small – an average of fewer than two observations per entry.

For those who practiced tone loudness (Table 2) there was no difference in the estimated structure between early and late sessions: the data from every participant, for both early and late sessions, were always better described by chain structures than rings. For those who practiced line length or dot separation (Table 2), the chain structure was also dominant for early sessions (10 out of 11 participants). For six participants however, the most likely structure changed from a chain to a ring from early to late sessions. Three participants demonstrated a chain structure both in the early sessions and in the late sessions, and one other demonstrated a ring structure in both early and late sessions. No participant demonstrated the reverse switch – from ring to chain structure. Consistent with the hypothesis that high performance in AI is only possible through more complex psychological representations, the single participant who demonstrated a ring structure during early practice also had very high performance in early practice, and the three participants who demonstrated a chain structure even late in practice were amongst the poorest performers late in practice.

A repeated theme in the above findings is that more complex (ring) structures are associated with better identification and with more improvement with practice. To

investigate this more formally, we calculated the correlation between both improvement in performance and initial accuracy, and log-likelihood. Both improvement in accuracy and initial accuracy demonstrated a strong negative relationship with log-likelihood difference values, where smaller log likelihood differences (representing a preference for a ring-structure) was associated with greater overall improvement in accuracy (r = -.70, p < .001) and greater initial performance (r = .65, p < .001; see Figure 2).

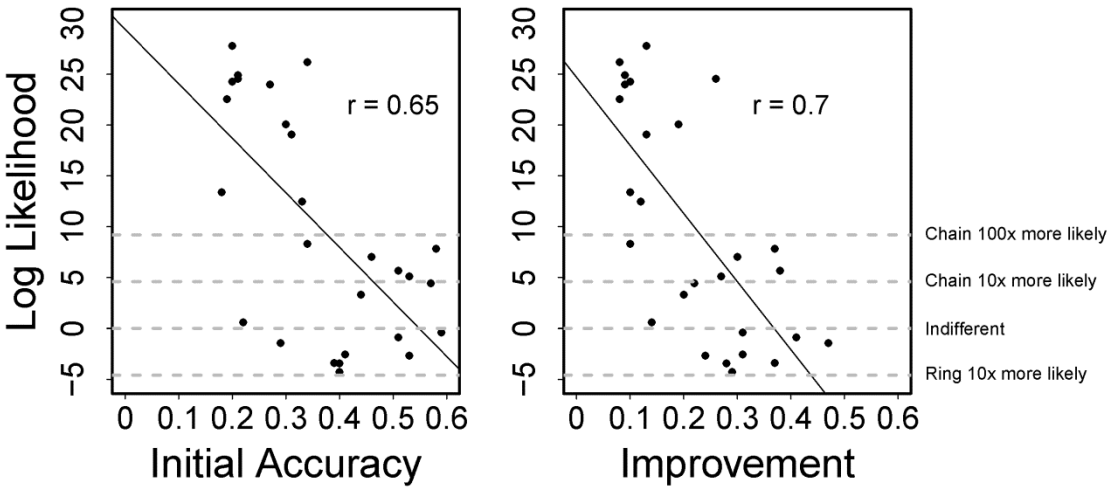


Figure 2. Accuracy and Improvement in accuracy as a function of difference in Log-Likelihood values for ring and chain structures (where a negative Log-Likelihood value indicates a preference for a ring structure). Note: two outliers were removed from this analysis (where log-likelihood difference was < -10 and >30).

Discussion

For more than fifty years, AI with unidimensional stimulus sets has been assumed to be subject to a strict performance limit, Miller's (1956) magical number 7+-2. More recently, Rouder et al. (2004) and Dodds et al. (2011) have shown that some stimulus sets support much greater performance than this limit (including line length or angle, and tone frequency) while at least one does not (tone intensity). One way to reconcile these new findings with previous literature is to hypothesize that some stimulus sets, while physically varying on only one dimension, give rise to a more complex psychological representation. The data from Dodds et al., and the new structural forms algorithm developed by Kemp and Tenenbaum (2008) provide a method for investigating this hypothesis in a way that was not previously possible because of limitations in analytic tools such as multidimensional scaling.

Our results provide consistent support for the previously untestable hypothesis that improved identification performance is only possible with more complex psychological representations. When we examined data from the identification of tones varying in intensity (for which identification performance was severely limited) we uniformly found strong support for the simplest unidimensional psychological representation – a chain, as assumed in all theoretical accounts of identification. This result was observed for all participants, and was also confirmed as the most likely structure in both the early and late practice data. Data from those stimulus sets for which Dodds et al. (2011) found significantly improved performance with practice yielded different results. The psychological representations of the stimuli for the more than one third of these participants (9 of 23) were better described by the ring structure than the chain structure. This figure rose to 8 of 11 participants when only data from the second half of practice were considered, in stark contrast to the 1 of 11 participants identified as

using a ring structure in the first half of practice. The hypothesized relationship between identification performance and structure was further supported by strong correlations between performance in practice and the goodness-of-fit of the ring and chain structures.

To check our results with data from another laboratory, we also analysed data from two of Rouder et al.'s (2004) participants. Those participants practiced line length stimuli in a similar procedure to that described above, using set sizes of 13 line lengths in one experiment and then 20 line lengths in another[10]. Both participants in the first experiment, and one out of two of the participants in the second experiment were better described by the ring structure. The participants that demonstrated a more complex structure were also those that demonstrated higher initial accuracy (M_{ring} = .85; M_{chain} = .68).

A natural question arising from our analyses is why we did not observe uniform results. That is, if improved performance in the identification task really is supported by more complex psychological representations of the stimuli, why did we not observe such representations for *every* participant? Two explanations seem plausible. First of all, in all experiments there was considerable variability amongst the participants in identification performance. About half of the participants did not learn to improve their performance beyond Miller's (1956) limit of 7±2 stimuli, and so it is consistent with the hypothesis that those participants should maintain the simplest (chain) psychological representations. Secondly, there is an inherent bias favouring the chain structure over the ring structure in noisy data. This bias arises because general noise (such as non-task-related responses, and random error) bias the confusion matrices towards uniformity, and uniform confusion matrices are – according to the structural forms algorithm –

[10] We did not analyse data from Rouder et al.'s 30-length experiment, as the sample sizes became prohibitively small.

better described by chain than ring structures due to the higher complexity penalty attracted by ring structures.

Our results indicate that better performance through practice in identification is associated with more complex psychological representations of stimuli. However, the results do not provide insight into exactly how those representations arise, nor what extra stimulus information is being represented. For example, it is easy to speculate that participants might learn to judge line lengths using information from several sources – perhaps the extent of the retinal image, or the magnitude of the saccade needed to traverse the line, or even cues gained by comparing the line to external objects such as the display monitor. Magnitude estimates obtained from these sources would presumably be highly, but not perfectly, correlated, which could result in psychological representations more complex than chains. Further studies might examine such hypotheses by attempting to limit the information available from such cues, for example by presenting visual stimuli using virtual reality goggles.

In summary, it seems that tone loudness was the only stimulus modality that showed consistent evidence for only a single underlying psychological dimension. Line length, dot separation and tone frequencies showed evidence for more complex psychological representations than simple chain structures - particularly for highly-performing participants and post-practice data. The implications of these results are remarkable for the study of memory in terms of AI – if these stimuli are truly represented on multiple dimensions, unidimensional AI does not apply to these stimuli. In the extreme, it might be that the long history of study of unidimensional AI should have been limited to the study of tones varying in loudness. Or in the very least, that the identification of other stimulus types only qualifies as unidimensional as long as participants are not well practiced.

Acknowledgements

We are grateful to Jeff Rouder and Richard Morey for sharing their data for this analysis, and to A.A.J. Marley and Chris Donkin for comments on an earlier version.

References

Bachem, A. (1950). Tone height and tone chroma as two different pitch qualities. *Acta Psychologica, 7*, 80-88.

Braida, L. D., & Durlach, N. I. (1972). Intensity perception: II. Resolution in one-interval paradigms. *Journal of the Acoustical Society of America,51,* 483–502.

Cox, T. F. & Cox, M. A. A. (1994). *Multidimensional Scaling*. London: Chapman and Hall.

Cox, T. F. & Cox, M. A. A. (2001). *Multidimensional Scaling*. London: Chapman and Hall.

Dodds, P., Donkin, D., Brown, S.D., Heathcote, A. (2010). Multidimensional scaling methods for absolute identification data *In S. Ohlsson & R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society. Portland, OR: Cognitive Science Society.*

Dodds, P., Donkin, C., Brown, S. D. & Heathcote, A. (2011) Increasing Capacity: Practice Effects in Absolute Identification *Journal of Experimental Psychology: Learning, Memory & Cognition, 37(2), 477-492.*

Eriksen, C. W., & Hake, H. W. (1955). Multidimensional stimulus differences and accuracy of discrimination. *Journal of Experimental Psychology, 50*(3), 153-160.

Garner, W. R. (1953). An information analysis of absolute judgements of loudness. *Journal of Experimental Psychology, 46*, 373-380.

Hartman, E. B. (1954). The influence of practice and pitch-distance between tones on the absolute identification of pitch. *The American Journal of Psychology, 67,* 1-14.

Kemp, C. & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105*, 10,687-10,692.

Lacouture, Y. (1997). Bow, range, and sequential effects in absolute identification: A response-time analysis. *Psychological Research*, 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.

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modelling. *Journal of Mathematical Psychology, 45*, 149-166.

MacLeod, D. I. A. (2003). New dimensions in color perception. *Trends in Cognitive Sciences*, *7*(3), 97-99.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63 (2),* 81-97.

Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. Psychonomic Bulletin & Review, 3(2), 222-226.

Pollack, I. (1952). The information of elementary auditory displays. *Journal of Acoustic Society of America, 24*, 745-749.

Rouder, J. N. (2001). Absolute identification with simple and complex stimuli. *Psychological Science, 12,* 318–322.

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.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. *II Psychometrika, 27 (3),* 219-246.

Weber, D. L., Green, D. M., & Luce, R. D. (1977). Effects of practice and distribution of auditory signals on absolute identification. *Perception and Psychophysics, 22*(3), 223-231.