A COMPARISON OF ABSOLUTE IDENTIFICATION AND FUNCTION LEARNING

MARK BROWN

A COMPARISON OF ABSOLUTE IDENTIFICATION AND FUNCTION LEARNING

by

Mark Brown

A thesis submitted to the

School of Graduate Studies

in partial fulfillment of the requirements

for the degree of Master of Science

Psychology Department

Memorial University of Newfoundland

June 19, 2012

Abstract

Absolute identification (AI) experiments are interested in how people remember the identity of simple perceptual stimuli. Function learning (FL) explores how people learn continuous relationships between stimulus (S) and response (R) dimensions. Although AI and FL are used to explore different cognitive processes, there are several important similarities between the two tasks, most importantly, the congruent S-R mapping used in AI creates a positive linear function. Three experiments begin to explore the commonalities between AI and FL. Experiments 1 and 2 use an AI methodology with 2 phases and increased the number of stimuli in phase 2 by adding either interpolation or extrapolation items. Classic AI and FL data patterns were both found depending on how the data were analyzed. Also, there was some evidence that people could respond accurately to novel stimulus values. Experiment 3 manipulated the instructions given to the participant (either AI or FL instructions) and the type of response labels (letters or numbers). Classic AI effects were observed for all groups; also, there was no difference in extrapolation/ interpolation performance. Overall, Experiment 3 revealed little evidence for differences between AI and FL, suggesting that both AI and FL involve the same cognitive processes.

Acknowledgements

I would like to thank my supervisor, Dr. Ian Neath for his help, guidance and encouragement. I would also like to thank the members of my supervisory committee, Dr. Aimée Surprenant and Dr. Rita Anderson for their advice and feedback. I would also like to acknowledge the members of the Cognitive Aging and Memory Lab, Annie Jalbert, Sophie Kenny, Roberta DiDonato, Marlena Hickey, Jamie March and Brittany Faux for their help and support. Finally, I would like to thank my wife, Safina Dewshi for her endless support and encouragement.

List of Figures

Table of Contents

Chapter 1 Introduction

1.1 Purpose of Research

In order to survive, an organism must not only identify individual objects, but must understand how individual objects relate to each other. Understanding that two unique objects are similar provides a method for grouping these objects within a single category, these categories can then provide a way to predict the behaviour of novel objects. For example, when we see an unfamiliar animal we can predict something about its behaviour by determining it is of the category *dog*, and objects belonging to the dog category are associated with barking and tail-wagging. Psychologists are often interested in exploring how people identify/categorize stimuli and how they use conceptual information to make predictions. Two methods for exploring these questions are *absolute identification* and *function learning*.

Absolute identification (AI) involves the mapping of unidimensional stimulus magnitudes onto discrete responses. For example, a participant might need to remember that the 600 Hz tone is response 1 the 800 Hz tone is response 2, the 1000 Hz tone is response 3, and so on. . Stimuli are presented one at a time and the participant tries to select the correct response. Feedback is provided after each trial so the participant can learn the correct response to each stimulus. Where AI uses discrete categories as response options, function learning (FL), on the other hand, involves the mapping of a continuous set of stimulus magnitudes onto a continuous

response scale. For example, a participant might need to learn how much fuel is required to drive a certain distance. Participants try to learn the functional relationship between the predictor (e.g., distance) and the criterion (e.g., amount of fuel needed). Participants learn the function relationship by estimating criterion values for a series of predictor values and receive accuracy feedback. At test, participants must respond to new predictor values and the accuracy of these predictions reflects how well the function concept was learned. Both AI and FL can be interpreted as conceptual tasks: categorization in the case of AI, and prediction in the case of FL. Research on AI and FL differ in their focus, but there are overlapping features between the two tasks. Because AI and FL try to answer different questions, some of the methodological details differ between the two paradigms. For example, AI and FL often use different dependent measures and different types of stimuli. In the research presented here, similar stimuli and dependent measures will be used in order to directly compare AI and FL. The goal of the research presented here is to examine the amount and type of overlap between AI and FL tasks.

The next sections will first describe the AI and FL tasks and compare the classic effects found in both paradigms. Next, an overview of AI and FL theories will be provided. Finally, the major methodological differences between FL and AI will be described and how these differences could affect performance will be addressed.

1.2 Absolute Identification

In a typical AI task a single unidimensional stimulus is presented and the participant responds by choosing a discrete response label. Feedback about the correct label for the presented stimulus is then provided. Several key phenomena are associated with the AI paradigm including: a performance limit that is resistant to practice, set-size effects, edge/bow effects, and sequential effects (for recent reviews see Petrov & Anderson, 2005; Stewart, Brown & Chater, 2005).

AI performance is notoriously resistant to improvement despite extensive practice (Miller, 1956; Shiffrin & Nosofsky, 1984). People are not able to perfectly identify more than the equivalent of about seven unidimensional stimuli; a surprisingly small limit when compared to the near infinite number of multi-dimensional stimuli that can be identified (Miller, 1956; Shiffrin & Nosofsky, 1984; Siegel & Siegel, 1972). Although the AI performance limit is one of the classic psychological effects, some recent research has called this limit into question. For example, Rouder, Morey, Cowan and Phaltz (2004) found that AI performance *did* improve with practice, with participants able to identify the equivalent of between 12 and 20 unique items (also see, Dodds, Donkin, Brown & Heathcote, 2011).

The size of the stimulus set affects how accurately people can discriminate between individual items; for example, two lines that are easily discriminated in the context of a two-item set become much more difficult to discriminate in the context of a ten-item set. Lacouture, Li and Marley (1998) provide data that clearly illustrate both the set-size effect and the bow-effect. The bow-effect (or edge-effect)

refers to the finding that responses to items from the ends of the stimulus range are more accurate than responses to middle items. Lacouture et al. (1998) found that as the number of items increased, performance became worse and the bow-effect became more pronounced. However, Lacouture et al. (1998) attributed the drop in performance to the number of response categories, not the number of stimuli.

Sequential effects in AI refer to how the immediate context (i.e., previous stimuli, responses, and feedback) affects responses to the current item (Lockhead, 1984). For example, responses to a current item are often pulled toward the immediately preceding item (i.e., assimilation) and pushed away from items further back in the series (i.e., contrast).

The identification of one-dimensional stimuli is superficially a simple task, however it can be approached from several inter-related perspectives; as a psychophysical task, as a memory task, or, as a categorization task. The psychophysical approach focuses on perception and attempts to describe how stimulus magnitudes are psychologically represented. As the goal of such research is to describe perception, researchers attempt to control factors such as memory or sequential effects (Lockhead, 2004).

In memory research, the AI paradigm is used to study how well simple unidimensional items are remembered and the patterns of errors that people make. AI as a tool for studying memory has two advantages: Because the stimuli typically vary along a single dimension the physical magnitude of the stimuli can be used to calculate how similar or different a particular item is from other items in the set

(e.g., Murdock, 1960; Neath, Brown, McCormack, Chater & Freeman, 2006). Secondly, the unidimensional nature of the stimuli reduces the possibility of some confounds that may occur with more complex stimuli.

AI can also be viewed as a special case of categorization where the number of categories equals the number of stimuli, and category membership is determined by the stimulus magnitude (Garner & Hake, 1951; Nosofsky, 1984). The focus of the categorization literature is to study concepts. In other words, categorization is used to gain insight into the rules, processes, and mental representations involved in determining if an exemplar is a member of a particular category. Interpreting AI as categorization provides a theoretical link between AI and other concept-learning tasks such as FL.

1.3 Function Learning

In order to explore how conceptual knowledge is psychologically represented researchers often employ categorization tasks. A categorization task usually involves presenting a stimulus to a participant, who then chooses the discrete category to which the stimulus belongs. However, many concepts are better described as continuous functional concepts, as opposed to categorical concepts. The FL paradigm is used to explore concepts where both the stimulus (X) and response (Y) are represented on continuous scales and the relationship between X and Y is determined by a mathematical function.

Functional relationships between variables are common in the environment and learning these relationships allows people to respond accurately to novel

stimulus values. Kalish, Lewandowsky and Kruschke (2004) give the example of a city worker who could determine the distance to a water main break (Y) based on the frequency of the sound (X). Other examples include being able to convert the price of an item from one currency to another (Juliusson, Gamble & Gräling, 2005), or estimating the amount of pollution in the environment at some future point in time (Wagenaar & Sagaria, 1975). To explore function concepts in the laboratory, participants learn the X-Y relationship from a series of exemplars. For example, the participant may be asked to predict "level of physiological arousal" for different quantities of a drug (e.g., Kwantes & Neal, 2006). Participants are trained on X-Y pairs and receive feedback about their accuracy. At test, the participant is shown new X values from within the training range (interpolation items) as well as outside the training range (extrapolation items). To illustrate, participants might learn the relationship between the speed of a car and stopping distance for speeds between 40km/hr and 65 km/hr. At test, participants apply their knowledge to make stopping distance estimates for speeds between 10km/hr to 39 km/hr (lower extrapolation) and between 66 km/hr to 100 km/hr (upper extrapolation). In addition, they will have to respond to speeds between 40 km/hr and 65 km/hr that were not used as training items (interpolation). Accuracy in responding to these novel stimulus values indicates how well the participant learned the relational concept.

Several typical findings within the function learning literature include: positive linear functions are easier to learn than negative linear functions, linear

functions are easier to learn than non-linear functions, and interpolation is more accurate than extrapolation (for a review see, Busemeyer, Byun, Delosh & McDaniel, 1997). Participants also tend to underestimate Y values in the extrapolation regions of a linear function (Delosh, Busemeyer & McDaniel, 1997), although this effect may be more reliable for the lower region than the upper region (Kwantes & Neal, 2006).

1.4 A Comparison of Absolute Identification and Function Learning

The similarity between AI and FL arises because of the relationship between the stimulus and the response scales. Usually, AI response keys are labeled and arranged so they correspond to the magnitude of the stimuli they represent (e.g., the smallest stimulus is labeled 1, the next smallest is labeled 2, etc.). The ordered mapping means that there are at least two ways a participant can solve the identification problem. The first option is that specific S-R pairs can be memorized. A second option is that the overall relationship between stimulus magnitude and response magnitude can be used to infer stimulus identity. In other words, a positive linear function based on ordinal values can be used to complete the identification task. FL tasks involve a regular and continuous S-R mapping which differentiates it from other categorization tasks (Busemeyer et al., 1997). The congruent S-R mapping in AI means that it meets the criterion needed to be considered a FL task, therefore, could potentially share some underlying processes with FL.

FL and AI both involve participants making a response from an ordered set when presented with a stimulus from a ordered set. However, participants appear

to be much more accurate when completing a FL task than when completing an AI task. A classic finding within the AI literature is the inability of participants to correctly identify more than the equivalent of seven different unidimensional stimuli regardless of amount of training. Miller's (1956) paper emphasizes the ubiquity of this performance limit, as it occurs across stimulus modalities (e.g., line length, frequency, saltiness) and experimental paradigms. This classic limit is not readily apparent in the function learning literature. For example, Delosh et al. (1997) used a FL task and trained participants on 8, 20 or 50 unique stimuli. Across training blocks, absolute deviations from the true function decreased to an average of 2.5 units on a 250 unit scale, and the number of unique training stimuli did not affect accuracy. In contrast, previous research on AI performance would predict accuracy to decrease as the number of training items increased and little improvement despite extensive training.

Participants appear to be very accurate by the end of FL training; however, the rate of learning is similar to what would be expected in an AI task. For linear functions, most of the improvement occurs within the first few blocks of trials after which there appears to be little improvement (Delosh et al., 1997; Kwantes & Neal, 2006; Lewandowsky, Kalish & Ngang, 2002). Similarly, AI accuracy does not continue to improve after the first few blocks of trials despite prolonged training. For example, even after experiencing 12000 AI trials, performance will remain poor (Garner, 1953; but see, Rouder et al., 2004). Generally, FL experiments use fewer training trials than AI experiments, however, the number of trials can vary

substantially within the AI paradigm. For example, Garner (1953) presented each stimulus up to 600 times, where as Murdock (1960) presented stimuli 10 times each. FL does not typically use an extremely large number of trials, for example, Delosh et al. (1997) presented training items either 4, 10, or 25 times each. However, in both the AI and FL paradigms, the data suggests that additional practice has minimal effect on accuracy after peak performance has been reached.

During the training phase of a FL task participants receive feedback. Because feedback is given, a FL training phase can be thought of as an identification task where the participant must remember the correct response value for each presented stimulus value. Most FL studies do not plot the training accuracy as a function of stimulus magnitude, therefore it is not possible to determine whether accuracy follows the bow-shaped pattern typical of AI. One exception is Kwantes and Neal (2006) and their data do not show the bow-effect for training items. Kwantes and Neal (2006) presented X values (i.e., stimulus values) as marked points along a scale as well as the numeric values. This additional information likely increased stimulus discriminability and may have eliminated any advantage for the edge items of the training set. Delosh (1997) found a bow-effect when the S-R mapping was random but not when the S-R mapping followed a negative linear function. However, Delosh (1997) looked for a bow-effect as a function of serial position (i.e., accuracy as a function of when an item was presented) not stimulus magnitude.

The ordered S-R mapping in AI means that AI meets the criteria to be considered a FL task. If the S-R function in AI provides participants with an additional source of information, a random S-R mapping would be expected to make performance worse. In general, S-R compatibility improves speed and accuracy when the experimental S-R mapping is congruent with an intuitive mapping (Fitts & Deininger, 1954), however, within the AI literature, the advantage of S-R compatibility is less clear. Lacouture and Lacerte (1997) showed that a congruent S-R mapping improved AI performance only marginally and that the effect was limited to the mid-range items. Eriksen and Hake (1957) also found that altering the S-R mapping did not affect accuracy, and the bow-effect remained as a function of stimulus magnitude. Additionally, Eriksen and Hake (1957) found that when AI stimuli varied on a dimension that had no natural end points, the bow-effect remained as a function of the response scale (but see Costall, Platt & Macrae, 1981).

Delosh (1997) studied FL and used a S-R mapping that was either random or followed a negative linear function. Participants were less accurate in the random mapping condition compared to the function mapping condition, suggesting that the congruency between the stimulus and response dimensions is an important source of information. Also, only when the mapping was random did increasing the number of items result in poorer performance, a pattern typical of AI. This finding is interesting because, as previously stated, the S-R mapping in an AI task is usually *not* random and is therefore more similar to the function mapping condition of Delosh (1997). An important difference between Delosh (1997) and the AI studies

conducted by Lacouture and Lacerte (1997) and Eriksen and Hake (1957) may be that Delosh (1997) used a random S-R mapping, whereas the S-R mapping in the AI studies maintained some structure.

The effect of a preceding stimulus on the response to a current stimulus is often explored in AI (for a review see Matthews & Stewart, 2009). Little work has been done on sequential effects in FL; however, McDaniel, Dimperio, Griego, and Busemeyer (2009) looked at ordered and random presentation in FL. If stimulus presentation is ordered during training, training performance is more accurate than if presentation is random. However, being trained on ordered items does not improve transfer performance (transfer items were presented randomly). Hu (1997) found similar results using an AI task. Hu manipulated the variability of the step size (either small or large) during training; participants were then tested without feedback (in random order). Performance in the training phase was better for the small-step group compared to the large-step group; however, during the test phase, the small-step group were less accurate than the large-step group. The results of both Hu (1997) and McDaniel et al. (2009) have parallels in the categorization literature if the range of stimuli is viewed as the category and the stimuli are viewed as category members. Receiving a highly variable set of category exemplars during training improves transfer performance compared to receiving a less variable training set (Posner & Keele, 1968).

1.5 Theories of Absolute Identification and Function Learning

Theories of AI and FL often overlap; for example, both FL and AI can be modeled using an exemplar framework. Exemplar models (e.g., Nosofsky, 1984; Nosofsky, Kruschke & McKinley,1992) propose that a stimulus is categorized based on how similar it is to the stored exemplars in memory. Exemplar models can model categorization in general, as well as AI in particular (Kent & Lamberts, 2005). Busemeyer et al. (1997) proposed a modified exemplar model (Extrapolation-Association Model [EXAM]; see also, Delosh et al., 1997) with the aim of explaining FL within a general categorization framework. In order to account for extrapolation, EXAM includes a linear rule component that allows it to respond to novel items outside the training range. Without the rule component, exemplar models of FL underestimate accuracy on extrapolation items because the model can only output the response associated with the nearest training item.

Alternative to exemplar theories in the AI paradigm are relative judgment theories (Laming, 1984; Stewart, et al., 2005). Relative judgment theories posit that AI performance is dependent on comparing the current stimulus to the immediate context (i.e., recently presented items). Because responses are made relative to recent items, there is no need to assume that representations of absolute magnitude play a significant role in AI performance (see Stewart, et al., 2005, for a review of absolute and relative models).

Alternative to exemplar models in FL are rule-based models (see McDaniel & Busemeyer, 2005 for a review). The rule-based approach proposes that during

training participants learn an abstract rule that represents the relationship between the predictor and the criterion (Koh & Meyer, 1991). The rule learning process is often conceptualized as learning the correct parameter values for a regression equation. When a novel predictor value is presented, participants use the rule to determine the correct criterion value. One problem with the rule-based models is that they overestimate extrapolation accuracy. If participants use a regression-like rule, performance should remain accurate regardless of how far an item is from the training range; however, participants do not extrapolate as well as rule models predict (Delosh et al., 1997). Recent rule models, such as the Population of Linear Experts (POLE; Kalish, et al., 2004) are more successful at predicting human performance by assuming that complex functions are approximated by selecting from a set of simple linear functions.

Relative models of AI could potentially be used to model FL performance. Because relative models do not respond based on stored absolute magnitudes, they may provide a parsimonious explanation of transfer performance, with responses being determined relative to recently presented items. For example, the Relative Judgment Model (RJM; Stewart et al., 2005) uses the *differences* between stimuli in order to model AI. In effect, participants learn the difference between stimuli that is equal to a unit change in the response category (Stewart & Matthews, 2009); this is, in some ways, very similar to learning the slope that relates the stimulus and response scales (see Kwantes, 2003 for a similar approach to modeling FL). However there may be important theoretical differences between the slope involved

in the RJM (or a similar approach) and the slope as conceptualized in rule-based FL models (e.g., Kalish et al., 2005, Koh & Meyer, 1991). One theoretical difference between the slope of an RJM-like approach and the slope of a rule-based approach is that, with a rule-based model, the slope is an abstraction representing the overall S-R relationship, whereas the RJM-like slope is derived from instances.

1.6 Differences Between Absolute Identification and Function Learning

The defining feature of a FL task is the continuous S-R mapping (Busemeyer et al., 1997); a characteristic shared by AI. Therefore, although AI can be thought of as a FL task, AI performance seems to be quite different than FL performance. There are several key differences between the tasks that need to be addressed. These differences include: how performance is measured, the cognitive strategy participants use, the experimental design, the presence or absence of feedback, and the surface features of the stimuli and responses.

1.6.1 Performance Measures

Perhaps the simplest explanation for the discrepancy between AI and FL performance is how performance is measured in the respective tasks. FL studies often measure deviations from the correct response (either absolute or signed), whereas AI experiments may use proportion correct, information transmitted (IT) or measures of discriminability. Therefore, participants in a FL experiment are given credit for being close to the correct answer. Averaging responses in FL may result in performance that appears very accurate, despite participants never being *exactly* correct. More stringent measures of performance used in AI (e.g., proportion

correct) will result in performance that appears inferior when compared with FL performance. Although it is not uncommon for both AI and FL experiments to use different measures of accuracy within the same study, to my knowledge there has been no cross-paradigm examination of how the performance measures affect data patterns. Therefore, it remains an open question as to whether AI performance and FL performance will mimic each other if performance is measured the same way.

The dependent measure can be critical, not only for assessing accuracy in general, but also for elucidating different qualities of the response pattern. For example, the mean response to a stimulus provides a measure of both the direction and degree of error in mapping stimulus magnitudes onto response magnitudes; however, the mean response might look quite accurate despite large response variability. Koh and Meyer (1991) addressed the averaging problem by measuring both *constant errors* (CE; the mean response for a particular stimulus) and *variable errors* (VE; the standard deviation of response values for a particular stimulus). CEs and VEs address two different aspects of performance; CEs are a measure of how well participants have learned the correct S-R mapping (i.e., the functional relationship), whereas VEs assess how consistently participants respond to a stimulus regardless of the experimental mapping.

Similar to VE, the information transmission (IT) measure used in AI is a measure of consistency, but unlike VE, IT is non-metric (Garner & Hake, 1951). For example, using the IT measure, a participant consistently calling stimulus 5 response 10 has the same effect as consistently calling stimulus 5 response 6. VEs,

on the other hand, are affected by the distance between a response and the mean response to a particular stimulus. Overall, because different measures of accuracy are used in AI and FL it may look as though people are much more accurate in FL simply because of how accuracy is assessed. For example, Petrov and Anderson (2005) show that when the range of errors is taken in to account, AI performance can look more accurate than when using the IT measure.

1.6.2 Strategy

Another possibility for the superior FL performance compared to AI is that participants may use different strategies for the two tasks. Lindahl (1964, 1968) emphasizes the distinction between general and non-general strategies. A general strategy is similar to adopting an abstract rule that can be applied to novel stimuli, whereas non-general strategies are based on specific stimulus/perceptual characteristics. A FL task may induce participants to adopt a general strategy involving learning the S-R relationship, similar to participants learning an *intervening concept* that relates stimulus magnitude to response magnitude (Busemeyer, McDaniel & Byun, 1997). In the case of AI, participants may use a non-general strategy focusing on specific stimulus characteristics (i.e., the stimulus magnitude). Although the AI problem is solvable by using relational information between the stimulus and response scales, participants may simply not recognize this strategy and rely on an item memorization strategy.

Differentiating between item information and relational information has been explored with verbal tasks, and there is evidence that relational and item

information are separate and additive (Hunt & Einstein, 1981; Hunt & Seta, 1984). However, the relational information in a verbal memory task involves the relationship between stimuli (e.g., words from the same category are highly related). In contrast, a FL task focuses on the relational information between the stimulus and response scales. Within the categorization literature, strategy has been shown to affect how a participant categorizes novel stimuli (Medin & Smith, 1981). For example, participants can be induced to categorize based on rules or on overall similarity depending on strategy instructions (Allen & Brooks, 1991; Smith, Patalano & Jonides, 1998). Also, and more generally, the literature on transfer-appropriate processing (Morris, Bransford & Franks, 1977), and encoding-specificity support the view that how an item is encoded (e.g., the strategies used, the task relevant features, etc.) has a strong effect on performance. It is therefore plausible that a task that focuses on the global relationship between stimuli and responses (i.e., FL) will result in a different level and pattern of performance compared to a task that focuses on item identity (i.e., AI).

1.6.3 Experimental Design

AI experiments involving the effect of set size often use a between-subjects design. In FL studies, the number and range of stimuli is a within-subjects factor, increasing from the training to the test phase. The AI studies that manipulate set size within-subjects show participants display both higher accuracy and improvement with training compared to between-subjects experiments (Dodds, Donkin, Brown, Heathcote & Marley, 2011; Dodds, et al., 2011; Kent & Lamberts,

2005; Rouder, et al., 2004). Also, within-subject designs can modulate the bow effects found in between-subjects designs (Dodds, Donkin, Brown, Heathcote & Marley, 2011). The within-subjects design of FL experiments may partially explain the discrepancy between AI and FL in the level and pattern of performance.

1.6.4 Feedback

Because experimenters conducting FL studies are interested in participants' ability to apply learned concepts to novel exemplars, feedback is not provided during the transfer test. The absence of feedback during transfer means that the participants must rely on their knowledge of the X-Y relationship rather than their memory for specific items (Delosh et al., 1997). In contrast, in AI experiments participants are typically give feedback throughout testing. When the effect of feedback is explored in AI, providing feedback tends to improve performance. However, feedback may act to influence the S-R mapping rather than improve stimulus discriminability (e.g., Eriksen, 1958). Mori and Ward (1995) found that feedback did not affect the discriminability of stimuli, but instead altered how the current response was affected by the preceding stimulus and response.

Brehmer and Svensson (1976) examined the effect of feedback in a FL experiment. Participants were informed of the shape of the function (either U or inverted U shaped) and had to predict a criterion for different levels of a predictor variable. Brehmer and Svensson found that providing feedback did not improve performance compared to a no-feedback condition.

1.6.5 Surface Characteristics and Response Scales

AI responses are usually made by selecting a discrete (but ordinal) response category, whereas FL responses are made by selecting a point along a continuous response scale. This difference means that FL gives participants access to many more unique response values compared to AI. Although this difference may seem important, previous research would suggest that making the response scale continuous would have little effect on performance, at least in terms of IT (Hake & Garner, 1951). However, others have found that AI performance gets worse as the number of response categories increases (Lacouture, et al., 1998). The method of responding may play an important role in how participants approach the task. Specifically, a continuous response scale may make the S-R *relationship* more salient compared to a discrete response scale.

Surface characteristics of the stimuli in a typical FL task may make these stimuli easier to remember and therefore result in accurate performance compared to AI. Within the FL paradigm, stimuli can take a variety of forms, for example, position of a marker on a scale (Delosh, et al., 1997), a line length (Kalish, et al., 2004), or numerals and letters (Sniezek & Naylor, 1978). In some FL procedures the stimulus characteristics may provide the participant with additional information that is typically not present in AI. For example, if the stimulus is presented as a marker along a scale, a participant may use the distance from the beginning of the scale, the distance from the end of the scale, and/or the distance from specific tick marks to aid in discriminating stimulus values. McDaniel et al. (2009) addressed the

role of tick marks on the stimulus scale in a FL experiment. Stimulus values were presented either as a filled bar on a marked scale or as segments of a circle with no scale markings. If tick marks on the stimulus scale provided additional information, participants should have been more accurate in the marked scale condition compared to the circle segment condition. However, McDaniel et al. found the opposite effect; participants performed better in the circle segment condition than in the scale condition. However, note that the segment stimuli also have additional information not usually present in an AI task because the size of the filled portion of the circle is perfectly correlated with the size of the unfilled portion. In general, the stimuli in FL may be easier to discriminate or remember because they have multiple correlated dimensions (see Garner, 1974).

When AI experiments use visual stimuli, additional cues, such as the distance from the end of a line to the edge of the screen, are often controlled. Although many FL experiments use multidimensional stimuli, some use stimuli that are very similar to those used in AI. For example, some FL experiments use the length of a line, or the distance between two markers to represent the level of a predictor variable (e.g., Brehmer, 1979, Koh & Meyer, 1991); a stimulus dimensions commonly used throughout the AI literature. Therefore, if a FL experiment uses line length as a predictor, the functional relationship is positive and linear, and participants receive feedback (e.g., the training phase), the FL task essentially becomes an AI task with the number of available responses exceeding the number of stimuli.

1.7 Summary

AI and FL paradigms both involve the study of concepts, but focus on different *kinds* of concepts. In FL, the participants' task is specifically to learn a relational concept; in AI the participants' task is to categorize stimulus magnitudes with discrete responses. In AI, the correspondence between the stimulus scale and the response scale makes the task solvable using a general function concept that relates stimulus magnitude to response magnitude. The results of AI and FL experiments differ substantially in terms of both level and pattern of performance. Different patterns of results in the two paradigms could be the result of: different strategies, different measures of accuracy, experimental design, and differences in the stimulus/response discriminability.

Therefore, in the present series of experiments, the goal was to begin to examine the factors that result in different data patterns in FL and AI. Experiments 1 and 2 follow a general AI procedure; participants were told to remember the correct numeric label for each stimulus and received feedback. Participants responded to a subset of items during the first phase of the experiment, then, the number of items was increased. Therefore, experiments 1 and 2 are identification experiments with a within-subjects set-size manipulation. The parallel to FL lies in *how* the set size was increased. The new items were either intermediate in size to the phase 1 items (i.e., interpolation items), or were larger and smaller than the phase 1 items (i.e., extrapolation items). Of particular interest in Experiments 1 and 2 is whether both classic AI and FL data patterns can be found in an identification

experiment simply by changing how the data are analyzed. Specifically, will the data look typical of AI performance when the proportion correct is the dependent measure, and will the data look typical of FL performance when the mean response is the dependent measure.

Experiment 3 also involved two phases, however, Experiment 3 did not provide feedback in the second phase. The absence of feedback means that the ability to transfer knowledge to novel stimuli can be assessed. Participants were instructed to either learn the *relationship* between stimulus magnitude and response magnitude, or, learn the *identity* of individual stimuli. Also, the response labels were manipulated so they represented either discrete categories or a continuous response scale. One of the critical questions for Experiment 3 was whether FL instructions result in better transfer performance than AI instructions.

Chapter 2 Experiments

2.1 Experiment 1

2.1.1 Purpose and Predictions

Experiment 1 followed an AI procedure and had two main purposes. The first purpose was to examine performance from both an AI and a FL perspective. In other words, if the measure of accuracy is changed, do the same data qualitatively mimic classic patterns in the two paradigms?

AI studies will often use the proportion of correct responses as a measure of performance. FL studies, on the other hand, often average the responses given to each stimulus. Experiment 1 will look at the data using both approaches. If AI and FL are highly similar tasks, the data should indicate low accuracy, and a bow-effect when proportion correct is examined. In contrast, when mean responses are examined they should appear very accurate and closely follow the S-R function.

The second purpose was to examine how increasing the number of items affects performance, and if receiving extra training on specific items improves performance on those items. Previous AI experiments have shown that receiving additional training on items can sometimes improve performance (Dodds et al., 2011; Cuddy, Pinn, Simmons, 1973; but see, Chase, Bugnacki, Braida & Durlach, 1982).

In AI terms, Experiment 1 involved a within-subjects set-size manipulation. Participants were trained on a subset of possible items during Phase 1. In Phase 2

the number of items was increased by adding stimulus magnitudes that were between the Phase 1 items. Increasing the number of items within-subjects is analogous to a FL experiment where the number of items is increased in the test phase. A second parallel between Experiment 1 and FL experiments involves how the set size was increased. Experiment 1 increased the number of items by adding items that are intermediate to the initial stimulus magnitudes (i.e., interpolation items).

It was expected that typical AI effects would be observed in Experiment 1, namely, a bow-effect (i.e., improved accuracy for the items at the edges of the stimulus set) and a set-size effect (decreased accuracy when the number of items is increased). However, when mean responses are used as the performance measure, the bow-effect should not be observed and the mean responses should follow a linear pattern consistent with the S-R relationship. Also, if giving participants extra practice on items improves accuracy for those items, it is predicted that when the set size is increased in Phase 2, Phase 1 items should have an advantage over new items.

2.1.2 Method

2.1.2.1 Design

The basic design has two within-subjects factors: stimulus magnitude (14 different stimuli) and Phase (Phase 1 and Phase 2). In order to separate the effect of additional training from the effect of stimulus magnitude, a between-subjects factor (Training Set; Odd/Even) was used. In Phase 1, the Odd group saw stimuli 1, 3, 5, 7,

9, 11, 13, and the Even group saw stimuli 2, 4, 6, 8, 10, 12, 14. In Phase 2, both groups saw all 14 stimuli.

The Odd/Even manipulation means that when the Even group sees Stimulus 1 in Phase 2, Stimulus 1 is technically an *extrapolation* item because it is outside the training range. The compliment occurs for the Odd group; with Stimulus 14 being outside the training range. Because there is only one extrapolation item per group and it is a direct neighbor of a training item, for the sake of convenience, I will refer to all new items as interpolation items.

2.1.2.2 Participants

Forty undergraduate students (6 males and 34 females) were recruited from Memorial University. All participants gave their informed consent before participating in the study. The mean age was 19.4 years (SD = 1.9). Participants were paid $10, and the experiment lasted approximately 30 minutes.

2.1.2.3 Stimuli

Stimuli were 14 red circles presented on a computer screen. Each stimulus had a unique numeric label (1 through 14) corresponding to its ordinal magnitude. The smallest stimuli (labeled 1) had a diameter of 10 pixels. The diameter of the circles increased by a constant 10 pixels (e.g., circle 14 had a diameter of 140 pixels).

2.1.2.4 Procedure

Participants were tested individually in a quiet testing booth. An iMac computer was used to present stimuli and collect responses. Participants sat a comfortable distance from the screen. The experimenter explained that the

participants' task was to remember the correct label for each circle. Participants were told that there were two phases and that in the second phase they would see new intermediate items as well as the old items.

2.1.2.4.1 Phase 1

Before testing began, participants were shown the seven Phase 1 stimuli with their correct label one at a time (once in ascending order once in descending order). The seven Phase 1 stimuli were then presented 10 times each, in random order (completely randomized without replacement). Response buttons for Phase l were seven virtual buttons (with numeric labels) in a single line along the bottom of the screen. The response buttons for all stimuli were in two rows at the top of the screen but only the bottom buttons were used for Phase 1.

The Phase 1 items were then presented individually and participants used the mouse to click on a response button. After making a response, the stimulus disappeared and participants were given feedback. If the response was correct the participant saw, "Correct! It was" with the correct label (printed in green), if the response was wrong the participant saw "Sorry! It was" with the correct label (printed in red). Feedback was presented visually in the center of the screen and remained on the screen until the participant clicked the "Next Trial" button.. Upon completing the 70 Phase 1 trials, participants were told that they would now see new items as well as the Phase 1 items. It was made clear that the Phase 1 items kept the same numeric labels in Phase 2.

2.1.2.4.2 Phase 2

Unlike in Phase 1, the entire Phase 2 stimulus set was not presented before testing began. The procedure for Phase 2 was the same as Phase 1 except all 14 stimuli were presented and responses were made using the 14 buttons at the top of the screen. Participants continued to receive feedback on all trials.

2.1.3 Results

The alpha level was set at .05 for all statistical tests. When the sphereicity assumption was violated, the Greenhouse-Geisser correction was used, and the adjusted degrees of freedom reported.

2.1.3.1 Absolute Identification Analysis

2.1.3.1.1 Phase 1

A 2 (Training set) x 7 (Relative Stimulus Magnitude) mixed-model ANOVA was conducted to determine if the relative magnitude of a stimulus affected accuracy, and whether the absolute magnitude of the stimuli affected accuracy. The dependent variable was the proportion of correct responses (e.g., the number of times response 7 was chosen when stimulus 7 was presented, divided by the number of times stimulus 7 was presented).

Figure 1 shows proportion correct plotted as a function of the relative stimulus magnitude. Both training groups show the typical bow-effect, with performance being better for the smallest and largest items compared to the middle items. As is evident in the figure, there was a significant effect of relative stimulus magnitude (F (6, 38)= 32.43, MSE = 0.023, $p < .01$).

Unexpectedly, there was also a main effect for training group (F (1, 38) = 4.23, MSE = 0.06, p = .047. Participants who were trained on the odd-item stimuli performed better than participants trained on the even-numbered items (Odd items, M = 0.85, SE = 0.021; Even Items M = 0.79, SE = 0.021).

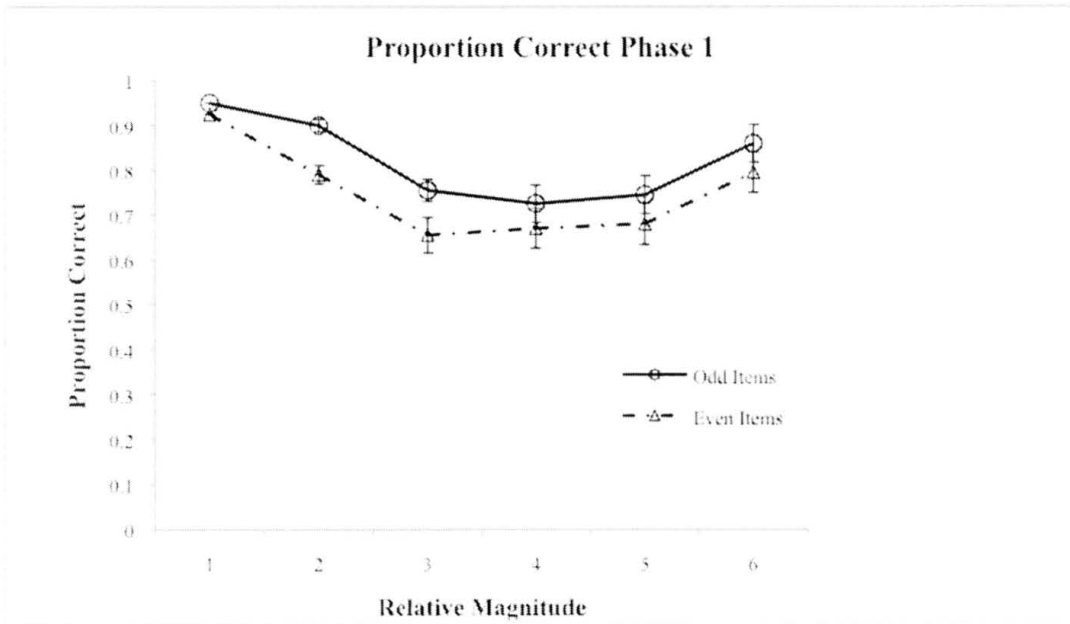

Figure 1: Proportion of correct responses in Phase 1 as a function of relative stimulus magnitude. Error bars show the standard error of the mean.

Evidently there was something that made the even stimuli more difficult to identify. However, there was no significant interaction between the relative stimulus magnitude and training set (F(6,38) = 0.842, MSE = 0.023 p = .514) indicating that the relative stimulus magnitude did not change the advantage held

by the odd-item group overall. Thus, this difference between the conditions did not affect the pattern of performance as a function of relative magnitude.

2.1.3.1.2 Phase 2

One purpose of Experiment 1 was to determine whether receiving extra training on an item improved accuracy for that item when the size of the stimulus set was increased. A 2 (Training Set) x 14 (Stimulus Magnitude) mixed-model ANOVA was conducted to determine if being trained on an item in Phase 1 improved performance on that item compared to novel items in Phase 2. The dependent variable was the proportion correct.

Figure 2 illustrates the proportion correct as a function of stimulus magnitude. As in Phase 1, a bow-effect is evident for both training groups. There was a significant main effect for stimulus magnitude ($F(13, 494) = 85.99$, $MSE = 0.036$, $p < .01$).

The training set had a significant effect on Phase 2 performance. Specifically, if participants were trained on the odd training items their performance in Phase 2 was superior to participants who were trained on the even items (Odd; $M = .497$, $SE = 0.02$; Even; $M = .426$, $SE = 0.02$; $F(1,38) = 6.08$, MSE = 0.115 $p = .018$).

The interaction between Training Set and Stimulus Magnitude did not reach significance ($F(13, 494) = 1.35$, MSE = 0.032, $p = .18$). Although the training set used in Phase 1 affected accuracy in Phase 2, there was no evidence for item specific effects.

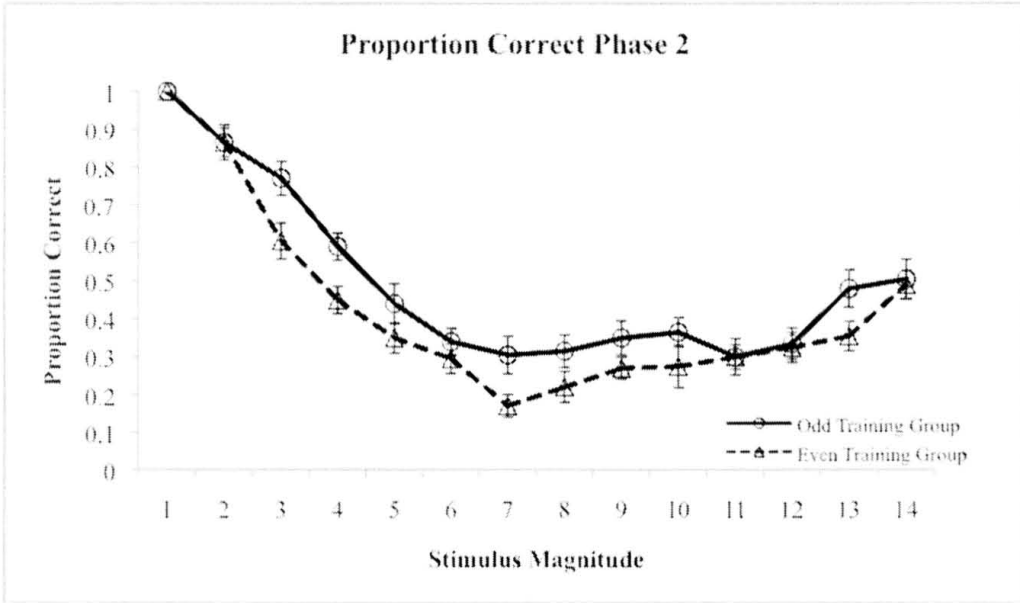

Figure 2: Proportion of correct responses in Phase 2 as a function of stimulus magnitude. Error bars show the standard error of the mean.

Visual inspection of performance in Phase 1 and Phase 2 shows that accuracy decreased when the set-size is increased. In order to examine the effect of set-size, performance on the seven Phase 1 items was examined in both Phase 1 (i.e., a small set context) and in Phase 2 (i.e., a large set context). The set-size effect was confirmed by conducting a 7(Stimulus Magnitude) x 2 (Phase) within subjects ANOVA for both the Odd and Even training groups. The proportion correct was the dependent measure.

Similar effects were found for both training groups. When the set-size increased in Phase 2, accuracy dropped for both the Even training group (Phase 1: M = 0.788, SE = 0.022; Phase 2: M = 0.417, SE= 0.018; $F(1,19)$ = 257.71, MSE = 0.037, p

$< .001$) and the Odd training group (Phase 1: $M = 0.848$, $SE = 0.019$; Phase 2: $M = 0.521$, $SE = 0.028$; $F(1,19) = 296.63$, $MSE = 0.025$, $p < .001$).

Increasing the set-size reduced participants' accuracy overall; however, not all items were equally affected. The interaction between Stimulus Magnitude and Phase was significant for both the Even ($F(6,114) = 5.51$, $MSE = 0.028$, $p < .001$) and the Odd training groups ($F(6,114) = 17.098$, $MSE = 0.018$, $p < .001$). Visual inspection of Figure 3 reveals that items from the ends of the range were less affected when the set size was increased.

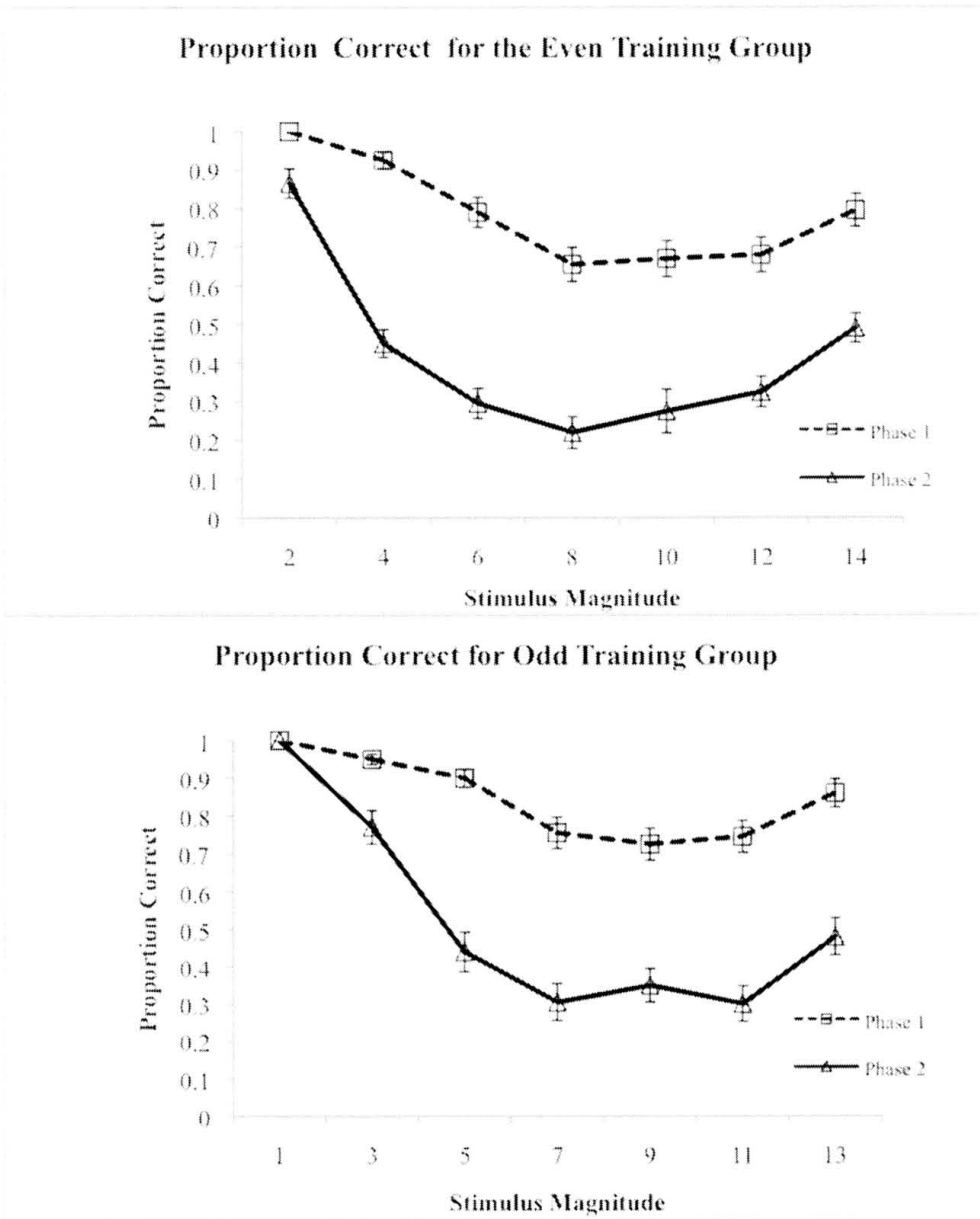

Figure 3: Training item accuracy in Phase 1 and Phase 2 for the Even Training Group (top) and the Odd Training Group (bottom). Error bars show the standard error of the mean.

There was a significant effect of Stimulus Magnitude for both the Even training group ($F(6, 114) = 53.42$, $MSE = 0.025$, $p < .001$) and the Odd training group ($F(6, 114) = 50.44$, $MSE = 0.023$, $p < .001$). The data show the bow-shaped pattern previously discussed.

2.1.3.2 Function Learning Analysis

2.1.3.2.1 Phase 1

Because the relative stimulus magnitude can perfectly predict the relative response magnitude, it is possible that participants are using conceptual information as described in the FL literature to make their responses in the context of an AI task. It is worth clarifying that in Experiment 1 feedback was provided during the Phase 2 trials, therefore a direct comparison between FL and AI is not possible in the current design. However, the data from the current experiment can be explored using the methods common to FL experiments.

To examine the effect of learning over trials, the mean absolute deviation of the participant's response from the correct response was calculated for each trial in Phase 1. The absolute deviations were then averaged into blocks of 10 trials.

A 2 (Training Set) x 7 (Block) mixed-model ANOVA was conducted to determine whether participants became more accurate with practice. Figure 4 shows the reduction in error as a function of trial block. It is apparent that participants were learning over trials, however, most of the improvement occurred during the first block of trials. There was a significant main effect of trial block ($F(6, 228) = 5.53$, $MSE = 0.016$, $p < .001$). The linear trend was significant ($F(1,38) = 8.23$,

MSE = 0.026, p = .007), however, higher order trends were also significant (quadratic $F(1,38)$ = 9.9, *MSE* = 0.016, p = .003; cubic $F(1,38)$ = 4.94, *MSE* = 0.018, p = .032; order 4 $F(1,38)$ = 6.64, *MSE* = 0.01, p = .014). These results replicate the finding that AI performance shows little overall improvement with practice. FL experiments also show a similar pattern learning over training blocks. For example, Delosh et al. (1997) found that absolute errors decreased quickly and asymptoted to a mean error of 2.4 or roughly 2.7% of the training range of the response scale. In the current experiment the mean error at the end of Phase 1 was 0.175 (*SE* = 0.018). Interestingly, if the mean absolute error in Experiment 1 is taken as a percentage of the number of responses, the resulting value is 2.5% similar to the error found by Delosh et al. (1997).

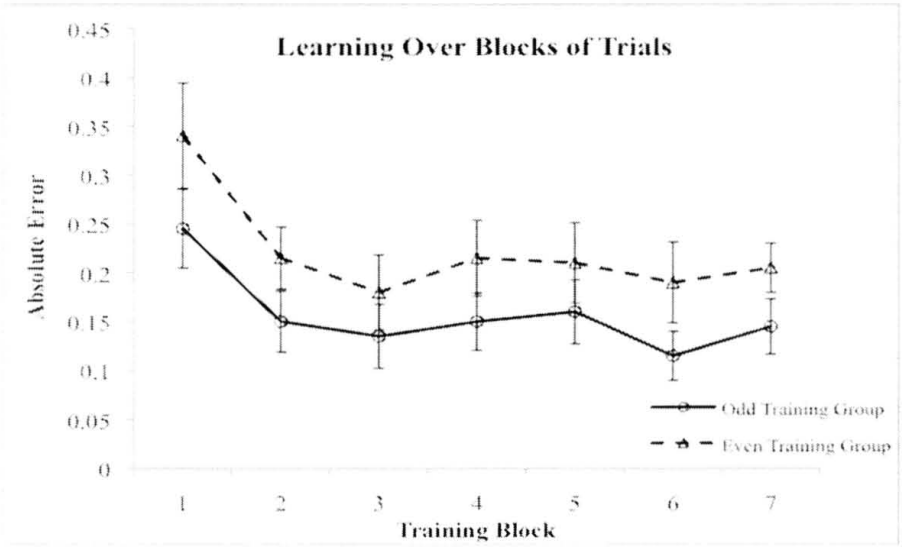

Figure 4: Mean Absolute Errors in Phase 1 as a function of trial block. Error bars show the standard error of the mean.

The effect of Training Set was marginally significant ($F(1,38) = 3.97$, $MSE = 0.074$, $p = .053$). The trend suggests that individuals in the Odd training group ($M = 0.157$, $SE = 0.023$) were more accurate than the Even training group ($M = 0.222$, $SE = 0.023$); a conclusion which is supported by the statistically significant difference found when proportion correct was used as the dependent variable. There was no significant interaction between Training Set and Block ($F(4.84, 183.86) = 0.171$, $MSE = 0.074$, $p = .971$).

2.1.3.2.2 Phase 2

Phase 2 performance was examined by calculating the mean response for each stimulus and plotting it as a function of stimulus magnitude. Figure 5 shows the effect of stimulus magnitude on the direction of errors; as stimulus magnitude increases participants tend to underestimate more. It is apparent that when the mean response is the dependent measure, participants appear to be much more accurate than when the proportion correct is used as the performance measure. In order to examine the pattern and direction of errors in more detail, the difference between each participant's mean response and the correct response was calculated for each stimulus. A 2(Training Set) x 14 (Stimulus Magnitude) mixed-model ANOVA was conducted to determine if the direction of errors differed as a function of stimulus magnitude, and, whether this effect depended on the training items. There was a significant main effect for stimulus magnitude ($F(4.63, 175.84) = 12.75$, $MSE = 0.735$, $p < .001$. The main effect of Training Set was not significant ($F(1,38) = 3.01$, $MSE = 0.866$, $p = .09$), nor was the interaction between Training Set and

Stimulus Magnitude ($F(4.63, 175.84) = 0.91$, $MSE = 0.735$, $p = .469$). The linear trend was significant ($F(1,38) = 45.35$, $MSE = 0.873$, $p < .001$).

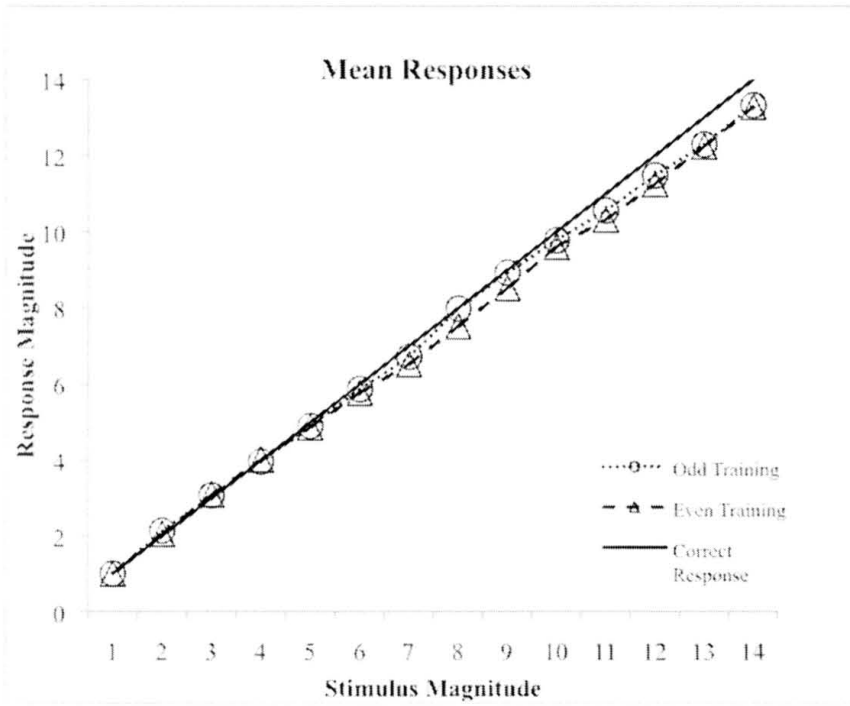

Figure 5: Mean responses as a function of stimulus magnitude.

Participants' ability to use conceptual information to infer the identity of novel items can be assessed, in part, by examining responses to the first presentation of an item in Phase 2. The number of participants who were correct on the first presentation of each stimulus in Phase 2 was summed. Figure 6 plots the percentage of participants who were correct on the first presentation of each stimulus (for both trained and untrained items) as a function of stimulus magnitude.

As a comparison, the mean number of participants who responded correctly over all stimulus presentations was calculated. Qualitatively, the shape of the function for first presentations is very similar to the mean of all presentations. There is some suggestion that seeing an item in Phase 1 increases the probability of a participant being correct on the first presentation of that item in Phase 2, particularly, for the largest items. However, the similarity between accuracy for first presentations and mean accuracy suggests that, to some extent, people are able to use what they learned in Phase 1 to respond to new items in Phase 2.

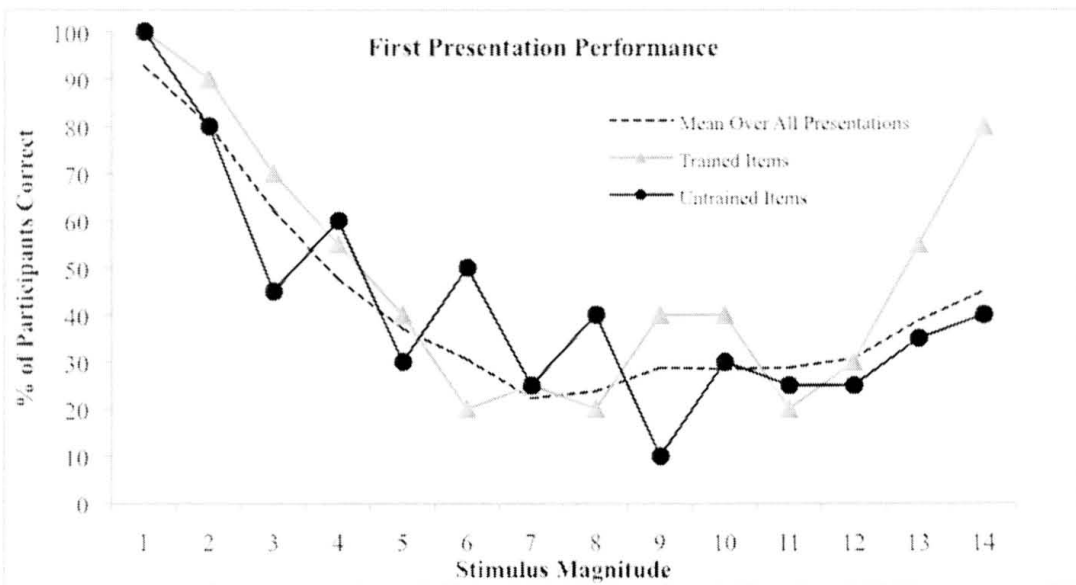

Figure 6: The percentage of participants who were correct the first time they responded to an item in Phase 2. Untrained items had never been seen before, whereas Trained items had been seen in Phase 1. The *Mean Over All Presentations* is the mean number of participants who were correct over all stimulus presentations.

2.1.4 Discussion

Experiment 1 addressed two main issues: first, whether receiving extra training on a subset of items improves accuracy for those items when the number of

stimuli is increased, and second, if the method of analysis is a major difference between AI and FL performance.

From an AI perspective, Experiment 1 replicated the bow and set-size effects typical of AI performance. However, Experiment 1 did not support the proposition that increasing the amount of practice on specific items improves performance on those items compared to less practiced items. AI usually does not improve beyond a low limit, however some recent studies have found that people can continue to improve if given enough practice (Dodds et al., 2011; Rouder et al., 2004). For example, Dodds, Donkin, Brown, Heathcote and Marley (2011) found improved performance on items that were presented more often. There are several important differences between their procedure and Experiment 1. Dodds, Donkin, Brown, Heathcote and Marley used stimuli that varied on a single dimension (the distance between two markers, or tones), whereas Experiment 1 used stimuli that were simple but not strictly unidimensional. Also, participants in Dodds, Donkin, Brown Heathcote and Marley experienced many more trials (1600 overall) than participants in Experiment 1 (210 trials overall). However, perhaps the critical difference is how the set-size was increased. They presented the two middle stimuli more often than the other stimuli, but in the current experiment, seven items received extra practice, and, these items were every-other item from the whole stimulus set. The distribution and number of stimuli that receive extra practice within the stimulus set may play a role in the efficacy of practice (also see Experiment 3 of Dodds, Donkin, Brown, Heathcote & Marley, 2011). If similar items

are seen more often, it may allow participants to organize the stimulus set into chunks, and therefore reduce memory load. In Experiment 1, the distribution of Phase 1 items (i.e., every other item from the whole set) may have made organizational strategies more difficult (see, Miller, 1956; Seigal & Seigal, 1972).

Surprisingly, Experiment 1 revealed an effect of training set on accuracy. Participants who were trained on the odd stimuli were more accurate than participants trained on the even stimuli. This effect in Phase 1 is likely due to the psychological spacing of the stimuli. If the psychological distance between stimuli is estimated by taking the log value of the stimulus diameter, the mean psychological distance between stimuli is greater for the odd set ($M = 0.427$, $SD = 0.351$) than the even set ($M = 0.324$, $SD = 0.202$). The increased stimulus spacing in the odd set might make these items less likely to be confused with each other, and therefore improve accuracy compared to the more closely spaced even set.

Psychological stimulus differences are a likely explanation of the advantage held by the odd-set in Phase 1, however, this does not explain why the odd-set advantage carries over into Phase 2. In Phase 2, both training groups saw the exact same stimuli; yet, the odd-set group was more accurate than the even-set group on both old items and new items. It is also worth noting that the variability of the stimulus differences is greater for the odd set than for the even set. Therefore, it is equally plausible that increased *variability* of the differences, not the *size* of the differences is the root of the odd group advantage. Within the categorization literature increasing the variability of the training items improves transfer

performance (Posner & Keele, 1968), yet, from an absolute judgment perspective, Lockhead (2004) would suggest that increased variability (on a trial by trial basis) would make performance worse. Resolving this issue is beyond the scope of the current paper but may be a interesting topic for future research.

The effect of the Phase 1 training set on performance in Phase 2 is interesting because it implies that people are learning something in Phase 1 that alters how they respond to Phase 2 items. The FL literature specifically focuses on this kind of knowledge transfer from training to test. Experiment 1 shows that a similar kind of transfer can occur even when feedback is provided on all trials.

When the results of Experiment 1 were approached from a FL perspective, two notable patterns emerged. First, when the mean response was plotted as a function of stimulus magnitude, participants appeared to be highly accurate and mean responses followed a linear pattern. Also, similar to Delosh et al. (1997), there was a tendency for larger stimuli to be underestimated. However, the pattern of underestimation of smaller items found by Delosh et al. (1997) and Kwantes and Neal (2006) was not apparent in the current study. The pattern of responses in Experiment 1 was qualitatively similar to FL results for positive linear functions despite Experiment 1 using very different stimuli and procedures. This similarity suggests that the accuracy measure in FL is why performance appears so accurate compared to AI performance.

In addition to the overall accuracy data mimicking both AI and FL patterns, the learning data in Experiment 1 also showed a pattern that is typical of both FL

and AI. Most learning occurs early in training and does not continue to improve across blocks of trials.

Responses to new items are the main focus of FL. Examining accuracy for the first time a stimulus was presented suggests that participants are relatively accurate in responding to novel items. This result implies that participants can infer something about novel items based on what they know about other items, and, that this learning can occur even when the task is to identify items, not learn a relational concept. Additionally, interpolation performance in Experiment 1 involved item specific interpolation rather than interpolation based on mean responses. Interpolation, as measured in a FL task uses the mean response given to a new item; therefore, it is possible that people are not inferring a specific response value, but rather, the distribution of errors centered on the correct response. The examination of first-presentation performance in Experiment 1 used a stricter criterion (right/wrong), therefore suggesting that item specific interpolation can occur within AI.

Four observations from Experiment 1 support the position that AI and FL are similar tasks. First, the shape of the learning curve was similar to previous work in both AI and FL; performance improved quickly, then leveled off at a suboptimal level. Second, by changing the dependent measure, the data mimic the patterns found in both AI and FL. Third, accurate responding to novel items suggests that participants can use knowledge about previous items to interpolate. Finally, exposure to a particular training set can affect how people respond to new items.

2.2 Experiment 2

2.2.1 Purpose and Predictions

In Experiment 1 the number of items was increased from Phase 1 to Phase 2 by including intermediate items. In Experiment 2 the number of items from Phase 1 to Phase 2 was again increased, however, the new items were stimuli that are smaller and larger than the Phase 1 items. Expanding the stimulus set in this way is analogous to a FL task where participants must extrapolate above and below the training range. In addition to changing how the set-size is increased, Experiment 2 altered the stimulus spacing. The diameter of the stimuli in Experiment 1 increased by a constant (10 pixels), in Experiment 2 the diameter of the stimuli increased exponentially (increasing by 30%). Exponential stimulus spacing may be interpreted as a logarithmic function between stimulus magnitude and response magnitude, assuming the psychological spacing of the responses is linear and there is a linear relationship between the stimulus physical magnitude and psychological magnitude.

Thirteen different stimuli were used in Experiment 2 compared to the 14 in Experiment 1. Although the number of stimuli was different in Experiment 2, using 13 items allowed the number of training items to be the same as Experiment 1 (7) and allowed an equal number of upper and lower extrapolation items.

The purpose of Experiment 2 was to examine an AI task where the range of stimuli is increased by including extrapolation items. Similar to Experiment 1, performance was analyzed from both an AI perspective and a FL perspective. It was

predicted that both a bow-effect and a set-size effect would be observed when proportion correct is examined. Mean responses to stimuli should follow a linear pattern.

The effect of practice on Phase 1 items cannot be answered definitively with the current design because the effect of practice is confounded with stimulus magnitude. Instead, the effect of practice will be addressed qualitatively by examining the overall pattern of accuracy in Phase 2.

2.2.2 Method

2.2.2.1 Design

Experiment 2 was a within-subjects design, the variables of interest were stimulus magnitude (13 different stimuli) and Phase (Phase 1 and Phase 2). Phase 1 used the seven stimuli from the middle of the stimulus set; Phase 2 used all 13 stimuli, therefore, each phase could be examined separately to determine the effect of stimulus magnitude on performance. Also, accuracy for Phase 1 items could be examined in a small set context (Phase 1) and in a large set context (Phase 2).

There are several confounds that occur with this design, for example, the order in which Phase 1 and Phase 2 was presented is not counter-balanced. Also, because only middle items are used as training items, stimulus magnitude is confounded with training. However, the purpose of Experiment 2 was to expand the stimulus set in a way that mimics FL and to approach the analysis from both an AI and a FL perspective. Therefore, the confounds that exist in Experiment 2 are the

same confounds that exist in a typical FL task. Because the present research is exploratory these confounds are not fatal to the current objectives.

2.2.2.2 Participants

Twenty undergraduate students (8 males and 12 females) were recruited from Memorial University. All participants gave their informed consent before participating in the experiment. The mean age was 21.1 years (*SD* = 3.0). Participants were paid $10 for their time. The experiment took approximately 30 minutes.

2.2.2.3 Stimuli

The entire stimulus set consisted of 13 circles. The diameter of the circles ranged from 30 pixels to 699 pixels, with the diameter of each circle increasing by 30%. Each circle was given a numeric label (1-13) corresponding to its ordinal magnitude.

2.2.2.4 Procedure

Experiment 2 followed the same general procedure as Experiment 1 with the following exceptions. In Experiment 2, the number of items was increased in Phase 2 by adding extrapolation items. In contrast, Experiment 1 increased the number of items by adding interpolation items. Thirteen stimuli were used in Experiment 2 instead of 14 in Experiment 1, and the Experiment 2 stimuli increased in diameter geometrically instead of linearly.

2.2.3 Results

The alpha level was set at .05 for all statistical tests. When the sphereicity assumption was violated, the Greenhouse-Geisser correction was used, and the adjusted degrees of freedom reported.

2.2.3.1 Absolute Identification Analysis

2.2.3.1.1 Phase 1

Accuracy for Phase 1 was assessed using a one-way, within-subjects ANOVA with seven levels representing the seven Phase 1 stimuli. The dependent variable was the proportion of correct responses. Participants were more accurate when responding to end items compared to middle items (see Figure 7). There was a significant effect of stimulus magnitude ($F(6,114) = 16.33$, $MSE = 0.027$, $p < .001$). The quadratic trend was significant ($F(1,19) = 83.69$, $MSE = 0.026$, $p < .001$).

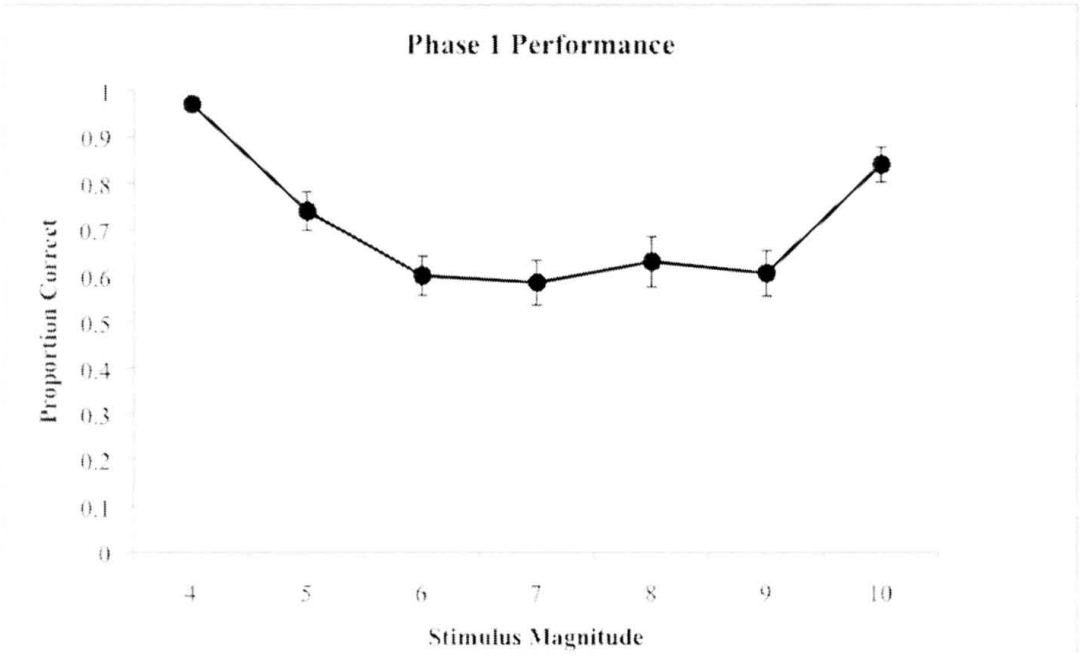

Figure 7: Proportion correct in Phase 1 as a function of stimulus magnitude. Error bars show the standard error of the mean.

2.2.3.1.2 Phase 2

Accuracy in Phase 2 was assessed using a one-way within-subjects ANOVA with 13 levels representing the 13 stimulus magnitudes. The dependent variable was the proportion correct. There was a significant effect of stimulus magnitude ($F(12, 228) = 14.71$, $MSE = 0.028$, $p < .001$. As in Phase 1, the quadratic trend was significant ($F(1, 19) = 47.77$, $MSE = 0.07$, $p < .001$). More interestingly, items that were edge items in Phase 1 were responded to more accurately than would be expected if accuracy was a simple U-shaped function of stimulus magnitude. Instead of a simple U-shaped function, the advantage held by the Phase 1 edge items resulted in a "double-bow" effect (see Figure 8). This pattern was significant, as evidenced by a sixth-order trend ($F(1,19) = 10.25$, $MSE = 0.033$, $p = .005$).

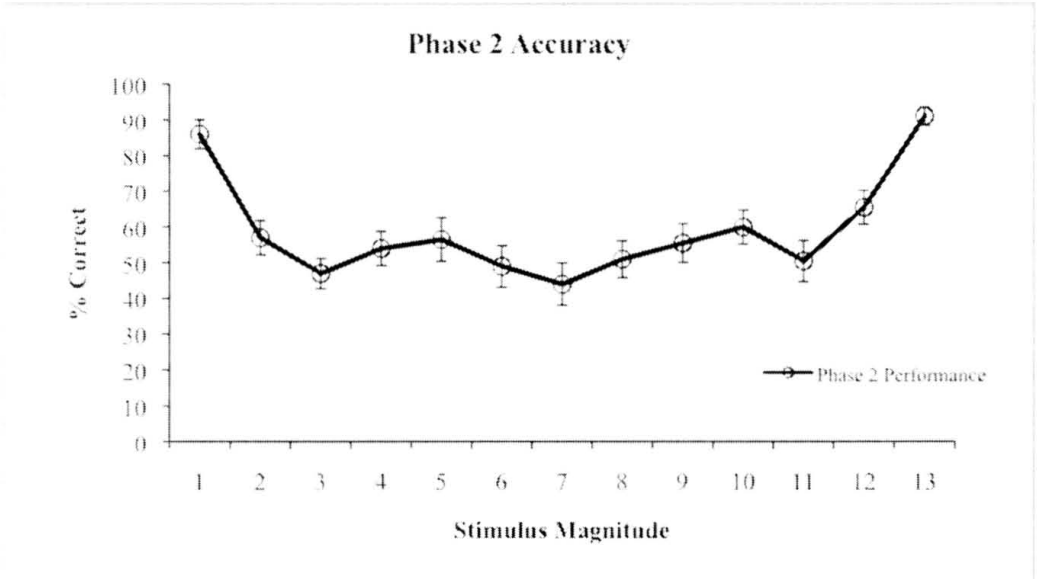

Figure 8: Percent correct as a function of stimulus magnitude in Phase 2. Error bars show the standard error of the mean.

In order to confirm that the data demonstrate a set-size effect, a 7 (Stimulus Magnitude) x 2 (Phase) within-subjects ANOVA was conducted. This analysis compared performance on the seven Phase 1 items to performance on the same items when they were seen in Phase 2. The dependent variable was the proportion correct.

The data demonstrate a set-size effect: when the items were presented in the context of a larger set, performance dropped from $M = .71$ ($SE = 0.025$) to $M = .529$ ($SE = 0.042$, $F(1,19) = 25.43$, $MSE = 0.091$, $p < .001$). As in Experiment 1, there was a significant effect of Stimulus Magnitude ($F(6,114) = 10.928$, $MSE = 0.031$, $p < .001$). Averaged over phases, performance still showed a bow effect (quadratic trend: $F(1,19) = 54.218$, $MSE = 0.034$, $p < .001$).

Figure 9 displays the proportion correct as a function of Stimulus Magnitude for both Phase 1 and Phase 2 and shows that increasing the set-size reduces accuracy more for the Phase 1 edge items than the middle items. Increasing the set-size did not hurt accuracy equally for all stimulus magnitudes; the Stimulus Magnitude x Phase interaction was significant ($F(6,114) = 7.61$, $MSE = 0.02$, $p < .001$).

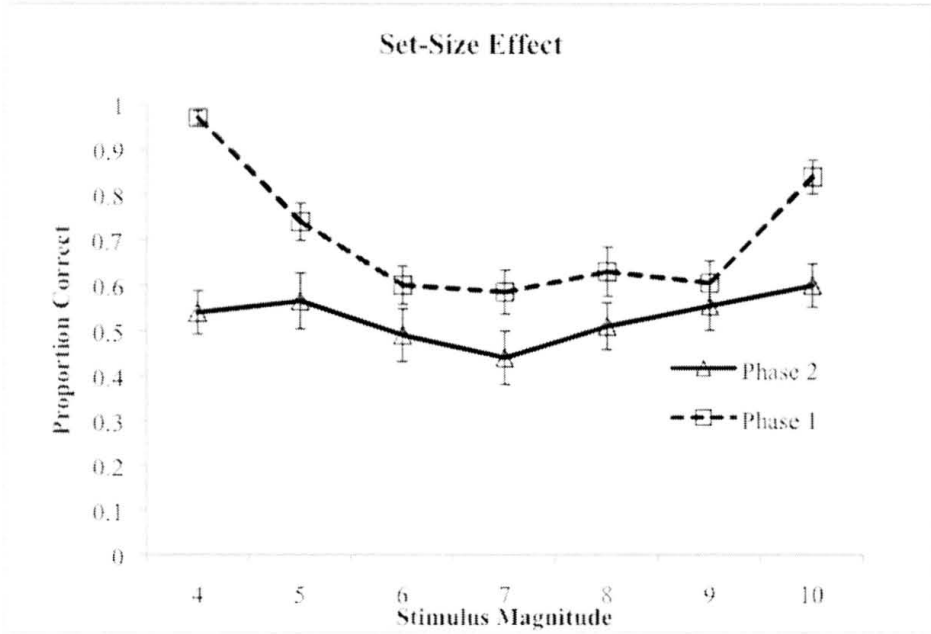

Figure 9: Proportion correct for training items in Phase 1 and in Phase 2, as a function of stimulus magnitude. Error bars show the standard error of the mean.

2.2.3.1 Function Learning Analysis

2.2.3.1.1 Phase 1

Learning in Phase 1 was assessed using a one-way, within-subjects ANOVA with 7 levels, representing seven 10-trial blocks. The dependent variable was the absolute difference between each participant's response and the correct response (averaged over blocks of 10 trials). Figure 10 plots mean absolute errors as a function of training block and shows a steady reduction in errors across blocks. In contrast to Experiment 1, learning in Experiment 2 appears to be a slower, more gradual process. There was a significant effect of training block ($F(6, 114) = 2.56$, $MSE = 0.026$, $p = .023$). Errors decreased from the first block of trials ($M = 0.405$,

SE= 0.047) to the last block of trials (*M* = 0.225, *SE*= 0.032). The linear trend was significant ($F(1,19) = 8.53$, $MSE = 0.04$, $p = .009$) but no higher order trends were significant (all Fs < 1).

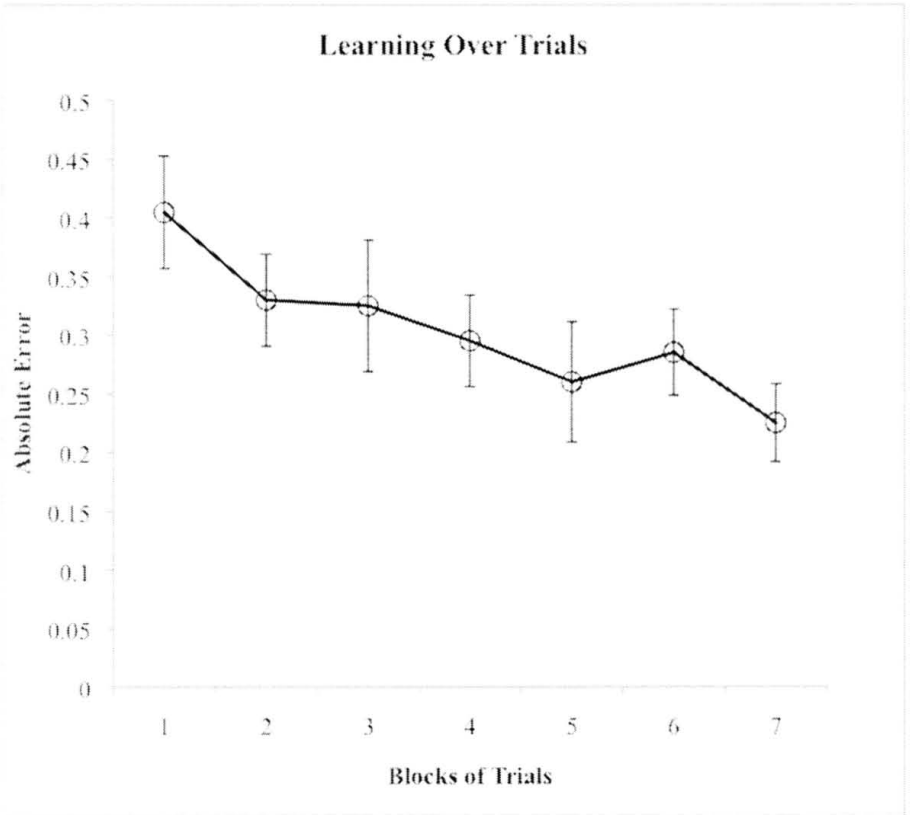

Figure 10: Mean absolute errors as a function of training blocks. Error bars show the standard error of the mean.

2.2.3.1.2 Phase 2

Performance in Phase 2 was analyzed as if it were a FL task by calculating the mean response for each stimulus magnitude. Figure 11 plots the mean response for

each item as a function of stimulus magnitude. When the identification task is plotted as if people were learning a conceptual S-R relationship, participants appear to be very accurate on average.

The direction and degree of error was analyzed for Phase 2. Each participant's mean signed response error for each stimulus was calculated. The mean signed error was used as the dependent measure in a one-way, within-subjects ANOVA with 13 levels for the 13 stimuli. Figure 12 shows the mean signed error plotted as a function of stimulus magnitude, all stimuli, with the exception of the smallest, tend to be underestimated. The U-Shape of Figure 12 illustrates better accuracy for the end items compared to the middle items. There was a significant effect of stimulus magnitude ($F(4.69, 89.04) = 3.902$, $MSE = 0.337$, $p = .004$).

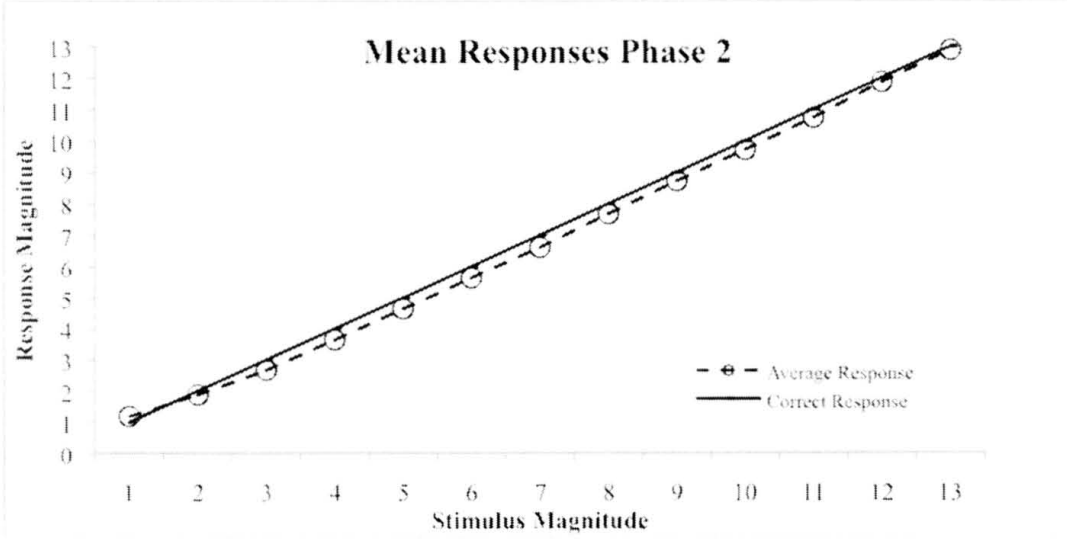

Figure 11: Mean responses as a function of stimulus magnitude.

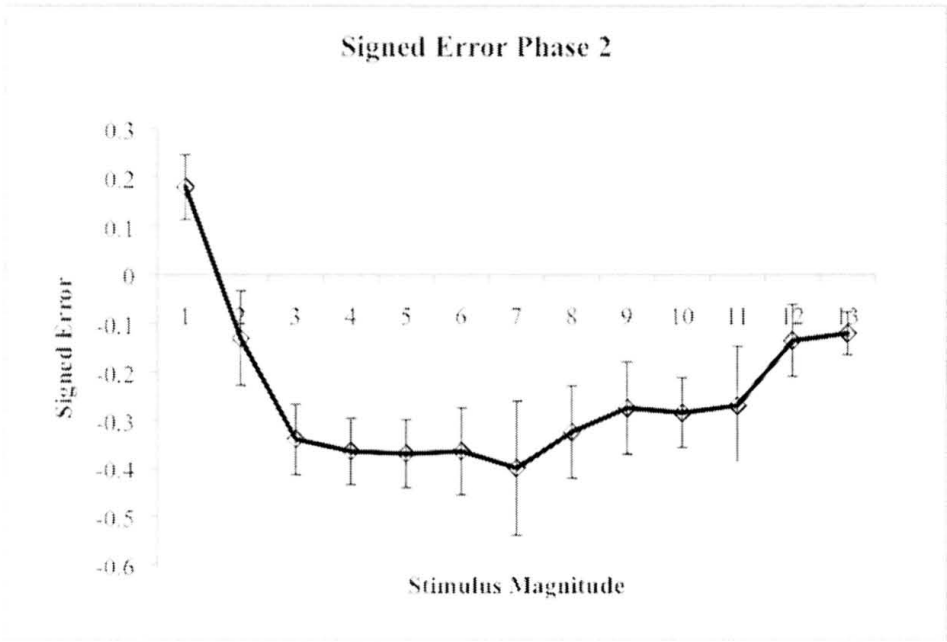

Figure 12: Mean signed error as a function of stimulus magnitude. Error bars show the standard error of the mean.

The ability of participants to infer the identity of novel items was explored by looking at the responses for the first presentation of an item in Phase 2. The percentage of participants who were correct on an item's first presentation is plotted as a function of stimulus magnitude, the number of correct participants averaged across all stimulus presentations is also plotted (see Figure 13). The main point of interest is that performance on the first presentation of an item is similar to mean performance. When presented with new items, there is some indication that participants are able to correctly infer the correct response for those items, especially the smallest and the largest items.

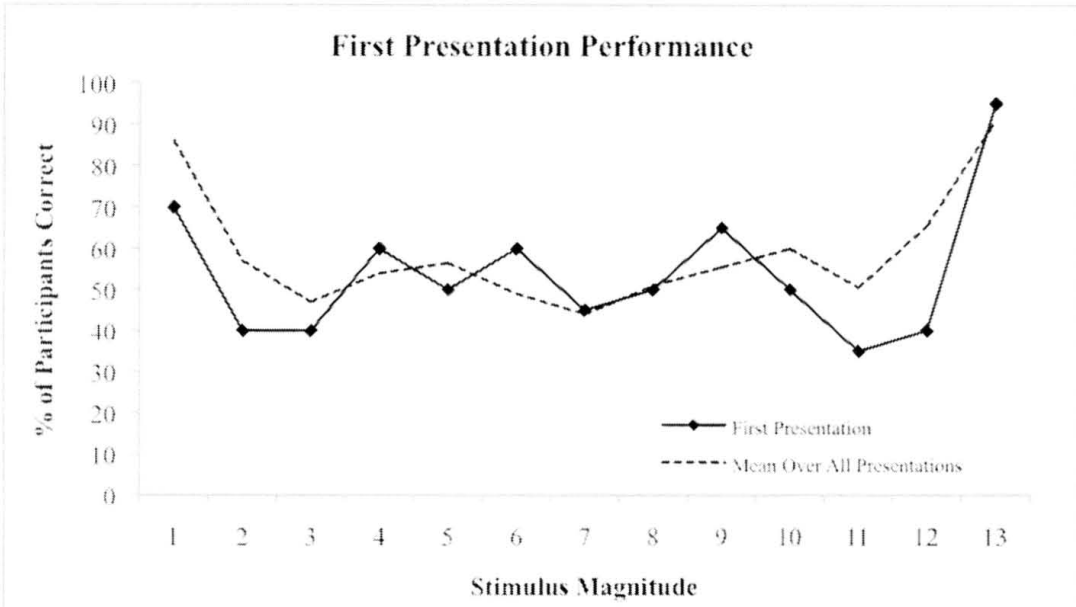

Figure 13: The percentage of participants who were correct the first time an item was presented in Phase 2. The *Mean Over All Presentations* is the mean number of participants who were correct over all stimulus presentations.

2.2.4 Discussion

Experiment 2 demonstrated both a set-size effect and a bow-effect typical of AI, however, the bow-effect in Phase 2 was not a simple U-shaped function. Specifically, edge items from Phase 1 maintained an advantage when new items were added to the ends of the stimulus range. Dodds et al. (2011) also found a modulated bow-effect using an AI task, and found that items presented more frequently were responded to more accurately. The results of Experiment 2 cannot

differentiate between the effect of additional stimulus presentations and the role of organization on performance. Receiving initial training on the middle items may allow participants to organize the stimuli into chunks (e.g., small, medium, large) and therefore facilitate performance (Miller, 1956; Seigal & Seigal, 1972). In Experiment 1 found there was no advantage for items that had been seen in Phase 1, perhaps because the structure of the Phase 1 items (every other item) did not allow efficient organization of the stimulus set.

When performance was examined using FL measures, several interesting patterns emerged. Similar to Experiment 1, the mean response to a stimulus appeared to be very accurate and followed a linear pattern. However, the direction of errors did not follow what is typical of FL. There was no indication that the largest or the smallest items were underestimated.

Comparing the learning rate in Experiment 1 and Experiment 2 (although qualitative) shows an interesting parallel between AI and FL. When the S-R relationship was linear (Experiment 1), participants quickly reached asymptotic performance. In contrast, when the S-R relation ship was non-linear (Experiment 2) performance improved linearly across training blocks. Similarly, within the FL literature, participants are able to learn linear functions more quickly than non-linear functions. However, it is not possible to say that the difference in learning between Experiment 1 and Experiment 2 is due to a different functional relationship between the stimulus and response items. It is just as likely that the difference is due to the stimulus spacing alone and not the S-R relationship.

Similar to Experiment 1, the number of participants who were correct when presented with a novel stimulus was similar to the mean number of participants who were correct across all stimulus presentations in Phase 2. This result suggests that participants can infer the identity of novel items, but the ability to infer a novel item's identity is most impressive for the Phase 2 edge items. The probability of a participant being correct on the first presentation of items between the Phase 1 edge and the Phase 2 edge is not much different from chance performance if it is assumed that participants know that the item is smaller (or larger) than the Phase 1 edge items (i.e., probability of guessing correctly is 1 out of 3). Although the results of Experiment 2 do not speak to what information participants are using when responding to novel items, the main point is that participants know something that allows them to be relatively accurate when responding to novel items.

Experiments 1 and 2 explored an AI task from the perspective of a FL task. The goal of the two experiments was not to provide definitive evidence that AI and FL involve similar processes, but rather, the intention was to approach the analysis of AI data from different perspectives, and determine whether the data matched classic patterns in the AI and FL paradigms.

Not surprisingly, how performance is measured plays a significant role in how accurate participants appear to be. The mean response can look very accurate and follow a linear trend (typical of FL) even when proportion correct displays relatively poor performance and follows a bow pattern (typical of AI). Therefore,

different accuracy measures in FL and AI probably account for the different levels of accuracy in the respective tasks.

In FL, responding accurately to novel stimuli is taken as evidence that the relational concept has been learned. In Experiments 1 and 2, novel stimuli were responded to relatively accurately indicating that participants can (at least to some degree) interpolate/extrapolate in an AI task.

FL studies show that non-linear functions are learned more slowly than linear functions (Busemeyer, et al., 1997). When the relationship between stimulus magnitude and response magnitude was linear (Experiment 1) accuracy improved quickly and leveled off. In contrast, when the S-R relationship was nonlinear (Experiment 2), accuracy improved gradually across training blocks.

2.3 Experiment 3

2.3.1 Purpose

Experiments 1 and 2 followed a general AI procedure. Experiment 3 used a procedure more similar to FL than AI. FL involves participants learning the correct S-R relationship during a training, during which feedback is given. At test, participants must respond to novel stimulus values, and are not given feedback. AI tasks typically provide feedback throughout the experimental session. If feedback is withheld, an AI task becomes absolute *judgment* rather than absolute *identification* (see Neath et al., 2006). For the sake of consistency, the term *AI* will be used to describe tasks that focus on item identity (even though feedback will not be

provided during testing). The term *FL* will be used to describe tasks that focus on learning the S-R relationship.

The goal of Experiment 3 was to directly compare AI performance with FL performance using a FL type procedure. The AI/FL comparison was made by manipulating aspects of the task participants performed. A significant difference between AI and FL is the strategy used when completing the tasks. Orienting participants toward either FL or AI strategy was done by providing participants with instructions highlighting either the S-R relationship (FL instructions) or highlighting item identity (AI instructions). In order to strengthen the relational/item processing distinction, FL participants responded by moving a slider underneath the response value they wanted, whereas AI participants clicked a response button. The type of instructions and the response method represent the general variable Task (FL or AI).

Another difference between AI and FL is the continuous response scale used in FL compared to the discrete/ordinal response scale of AI. The response scale in Experiment 3 used either letters or 3-digit numbers as response labels. Letter labels were meant to represent discrete response categories, whereas, numbers were intended to make the response scale appear more continuous. Experiment 3 manipulated these two variables in a 2 (Task; FL or AI) x 2 (Response Label; Letters/Numbers) between-subjects design. Therefore, the FL/Number cell is a good approximation of a typical FL task, while the AI/Letter cell approximates a typical AI task. The procedure followed a general FL methodology: participants

were trained on a subset of items from the middle of the range and given feedback, then, participants were tested (i.e., no feedback) on the training items, interpolation items, and upper and lower extrapolation items.

2.3.2 Predictions and Design

The experimental design was a 2(Task; FL/AI) x 2 (Label; Letter/Number) factorial. As previously noted, the FL/Number cell was the best approximation of a FL task, whereas, the AI/Letter was the best approximation of an AI task. Therefore, given the high performance levels found with FL and the poor performance associated with AI, participants in the FL/Number condition are predicted to be more accurate than participants in the AI/Letter condition.

If both FL instruction and a continuous response scale improve accuracy and instruction has a stronger effect, then, the FL/Number group should show the highest accuracy, followed by FL/Letter, followed by AI/Number, followed by AI/Letter. On the other hand, if the continuous response scale is a necessary condition for a FL instruction advantage, then the FL/Number group should show the highest accuracy and there should be no difference between the other groups. Predictions regarding accuracy can be examined for both the training phase and the test phase. If there is an advantage for the FL groups in the training phase (when feedback is provided) it would provide evidence that AI and FL strategies are inherently different because feedback should make the responses of FL and AI groups similar. Alternatively, the advantage of a relational (i.e., FL) strategy may only improve performance for new items.

One of the questions of interest is the degree to which participants can use previous experience to respond to new stimuli. The FL instructions should improve extrapolation performance compared to AI performance. If extrapolation performance depends on the response scale being perceived of as a continuous scale then the FL/Number group should extrapolate better than the FL/ Letter group. However, if extrapolation can occur with a discrete ordinal scale then extrapolation performance should be similar in the FL/Number and FL/Letter group.

As well as looking at how the Task and Response Label variables affect test phase accuracy, the data will be analyzed to look for classic AI effects, namely the bow-effect, the set-size effect, and asymptotic learning. If FL and AI represent two completely different kinds of tasks, the AI effects should appear only for the AI group, and, these effects should be most robust for the AI/Letter group. However, both AI and FL probably require some of the same processes and therefore an attenuation of the three AI effects in the FL groups is the most likely scenario.

2.3.3 Method

2.3.3.1 Participants

Fifty-two students (36 female and 15 male) from Memorial University of Newfoundland participated in the experiment. The mean age was 19 years old (*SD*= 1.59). Participants were paid $10 for participating, and the experiment took approximately 45 minutes. Participants were randomly assigned to groups and informed consent was obtained from all participants before the experiment began.

2.3.3.2 Stimuli

Stimuli were 25 vertical, blue lines measuring 9 pixels wide. The shortest line was 35 pixels long and the longest line was 765 pixels long, increasing by a constant 30 pixels (approximately 8mm). Lines were presented within a light grey rectangle (resembling an unmarked scale) 30 pixels wide and 800 pixels high, centered horizontally and positioned 319 pixels from the bottom of the screen. The distance between the top of the longest line and the top of the scale was 35 pixels, equal to the length of the shortest line. This control means the range of possible (but not presented) stimulus values was equal above and below the presented stimulus set. All lines were anchored at the bottom of the rectangle and extended upward.

Seven stimuli from the middle of the stimulus range were used as training items. The training range was from stimulus 7 (218 pixels long) to stimulus 19 (583 pixels long). Alternating stimuli were used from the training range providing seven unique training items. The remaining six items from the training range were used as interpolation test items; the six items below the training range and the six items above the training range were used as extrapolation items.

2.3.3.3 Response Scales

In order to strengthen the task manipulation, two different response procedures were used. When the instructions emphasized the S-R relationship (i.e., FL instructions), participants used the mouse to move a slider along a horizontal

track (from left to right) until it was positioned under the desired response label. Responses were registered after the participant released the mouse button.

When instructions emphasized item identity (i.e., AI instructions), participants made their response by clicking a response button. Response buttons were contiguous, light grey in colour, and arranged horizontally (in ascending order from left to right). The length of the response scale was the same for both response methods (approximately 47.5 cm), and the width of a button was equal to the width of slider range dedicated to each response label (approximately 19mm). Figure 14 illustrates how the stimuli were presented and the response method.

The labels used on the response scale were either the letters A through Y, or numbers corresponding to a linear function. Response labels were printed in a black 15pt. font. The use of letters should induce participants to view the responses as discrete categories, whereas numbers should make the response scale appear more continuous.

Because there are no numeric stimulus values, applying numeric response labels is arbitrary for a linear function. The numeric labels were based on the linear equation $y = 1.7x + 91$ with the 30 pixel difference between stimuli representing 10 theoretical units. The lowest response label was 108 and labels increased by 17, to a maximum of 516.

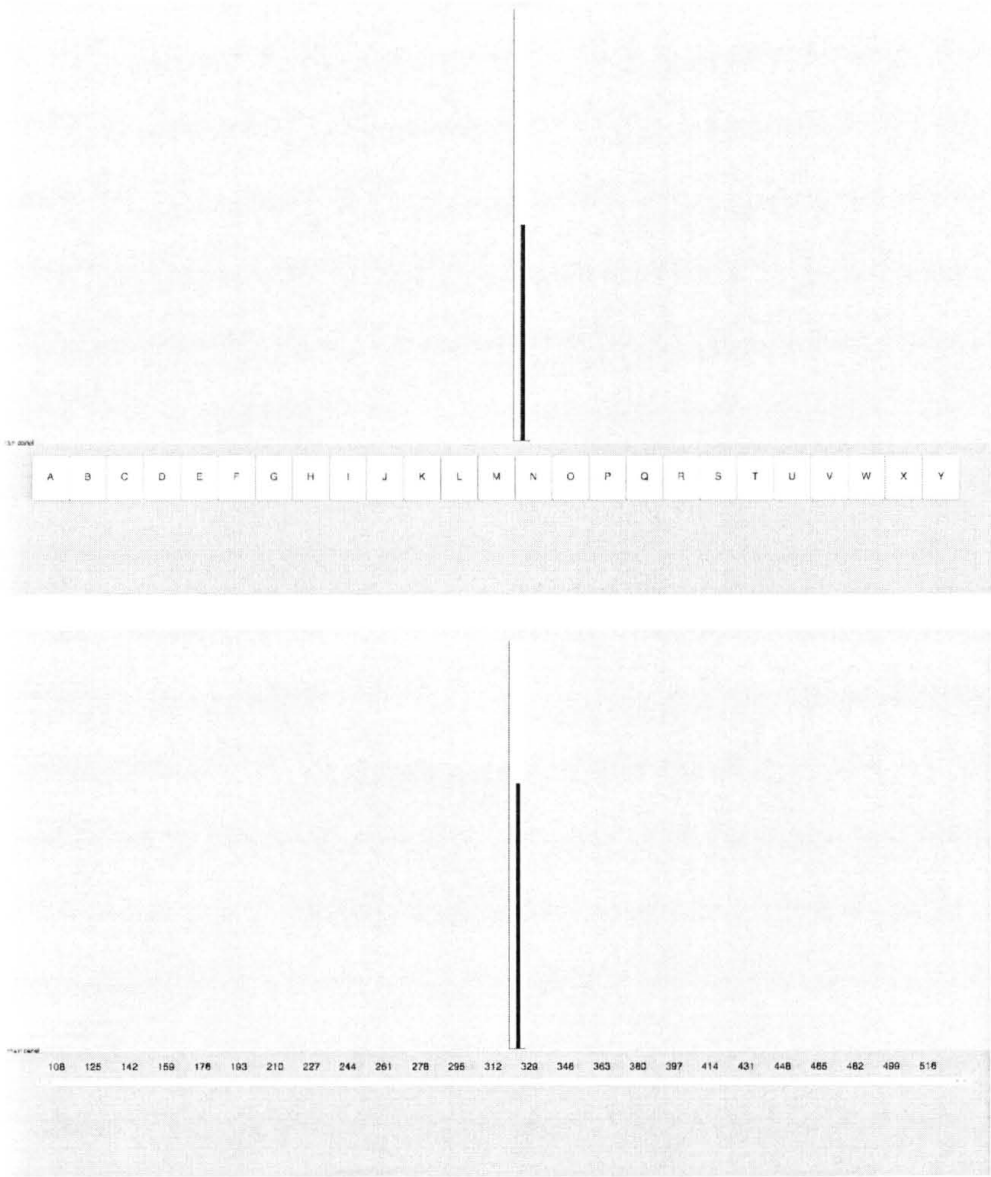

Figure 14: Screen shots for the AI/Letter condition (top) and the FL/Number condition (bottom).

2.3.3.4 Procedure

An iMac computer was used to present stimuli and collect responses.

Participants were arbitrarily assigned to one of four experimental cells from the 2 (Task: FL/AI) x 2 (Label: Letter/Number) design.

2.3.3.4.1 Instructions

Participants in the AI condition received instructions that emphasized the memorization of stimulus magnitudes. AI participants were told that the purpose of the experiment was to determine how well people could remember simple stimuli. Participants were told that they would see lines of different lengths and their task was to remember the correct label for each line length.

Participants in the FL condition received instructions that emphasized the relationship between line length and response magnitude. The cover story for the FL/Number condition was that a greenhouse owner had determined there was a relationship between the amount of fertilizer a plant receives and how tall the plant grows. The amount of fertilizer was represented by the length of the line, and plant height (in centimeters) was the numeric response label. Participants were told that their task was to learn the relationship between fertilizer and height.

Because there is not an intuitive relationship between amount of fertilizer and a letter, participants in the FL/Letter condition received slightly different instructions. FL/Letter participants were told that the greenhouse owner had developed a system for categorizing plants based on how much fertilizer they required and the categories were represented by the letters A through Y. Participants were told that their task was to learn the relationship between the amount of fertilizer and the category label.

All participants were told that there were two phases to the experiment and they would see a subset of stimuli during Phase 1 and be given feedback. They were also told that in Phase 2 they would see all of the items and they would not be given feedback.

The experimenter answered any questions and made sure participants understood how to make their responses.

2.3.3.4.2 Phase 1/Training

The seven training items were presented 15 times each in random order (without replacement).

After the participant selected his/her response, feedback was given. If the participant was correct, the words "Correct! The correct answer is" with the correct response label appeared, printed in green. If the participant was incorrect, the words "Incorrect. The correct answer is ..." with the correct response label, appeared printed in red. The feedback was presented in a grey box that appeared near the bottom of the screen.

For the conditions that used the slider response method, the slider remained in the response position the participant had chosen while feedback was presented. For conditions that used response buttons, the participant's response remained highlighted during feedback (a light blue highlight appear around the response button when that button was chosen). The letters (or numbers) of the correct response appeared in green on the response scale while feedback was presented. Participants clicked on the feedback box to proceed to the next trial. When the

feedback button was clicked, the stimulus line disappeared and a 750 ms delay preceded the next trial. The slider was reset to the far left, or, the button highlight was removed before each trial. At the end of training, a screen appeared providing instructions for the test phase. The instructions indicated that participants would now see all the stimuli and feedback would not be given.

2.3.3.4.2 Phase 2/ Test

All 25 items were presented 10 times each in random order during the test phase. When a response was made a grey box appeared at the bottom of the screen with " Click to Continue" printed in it. Participants clicked this box to proceed to the next trial. Upon completion of the test phase, participants were asked about any strategies they used while completing the task.

2.3.4 Results

Data from twelve participants were excluded from the analysis. One participant withdrew before completing the experiment. One participant responded in a highly idiosyncratic manner that appeared almost random. Nine participants were excluded because they reported explicitly limiting their responses to every other response option in the test phase. The training phase consisted of every other item from the middle of the set. It appears as if these nine participants extrapolated the same pattern throughout the test phase, despite being told that the test phase contained all of the items. Of the participants who explicitly constrained their responses, five were from the AI/Letter group, one was from the AI/Number group, one was from the FL/Letter group, and two were from the FL/Number group.

Although this kind of responding may be interesting in itself, I limited the analysis to participants who were at least open to using all the response options in the extrapolation phase. One additional randomly selected participant was removed in order to equate the number of participants in each condition (10 in each cell). The final sample was 28 female (Mean age = 18.9, *SD* = 1.8) and 11 male participants (Mean age = 19.4, *SD* = 1.6; one participant's demographic information was lost).

Responses that were more than six response categories away from the correct response were removed from the analysis. This criterion was set with the intention of including the full range of errors, while attempting to minimize noise from accidental responses. There were 84 responses (out of 15620) removed using this criterion. The alpha level was set at .05 for all statistical tests and the Greenhouse-Geisser correction was used when the sphericity assumption was violate

2.3.4.1 Phase 1/Training

It was expected that the edge items of the training range would be responded to more accurately than items from the middle of the training range. However, it was also expected that orienting participants toward a relational strategy would change the shape of the bow effect, namely, the bow effect was expected to be less pronounced for participants receiving FL instructions compared to participants receiving AI instructions, especially for numeric response labels.

Performance on the training phase was first assessed by calculating the mean absolute deviation (AD) of a response from the correct response for each stimulus

magnitude. As can be seen in Figure 15, all four groups show a bow-effect, with participants being more accurate in responding to the edge training items. A 7 (Stimulus Magnitude) x 2 (Task) x 2 (Response Label) mixed-model ANOVA was conducted to determine if either the response labels or how the task was framed affected accuracy.

There was a significant effect of Stimulus Magnitude on accuracy ($F(6, 216) = 8.24$, $MSE = 0.153$, $p < .001$). Participants were more accurate when responding to items from the edges of the training range compared to items from the middle, as evidenced by a significant quadratic trend ($F(1, 36) = 21.75$, $MSE = 0.275$, $p < .001$). Contrary to what was expected, Stimulus Magnitude did not interact with either Task or Response Label, nor was the 3-way interaction significant (all Fs< 1).

Neither the type of task ($F < 1$) nor the type of response labels ($F(1,36) = 2.93$, $MSE = 0.623$, $p = .096$) had an effect on accuracy. Additionally, the Task x Response label interaction was not significant ($F(1,36) = 1.05$, $MSE = 0.623$, $p = .312$).

The results from the training phase indicated that the type of instructions given to participants did not modulate the bow-effect during training. Therefore, when feedback is provided, focusing a participant on the relationship between stimulus magnitude and response magnitude does not affect accuracy.

The other classic finding in AI is the set-size effect; the finding that items are responded to more accurately in the context of a small set than in the context of a larger set. In order to determine if a set-size effect occurred in Experiment 3,

performance on the training items in the training phase was compared to performance on the same items in the test phase. A set-size effect would appear as a decrease in accuracy from the training phase to the test phase for the training items.

Figure 16 shows that when the Phase 1 items were seen in Phase 2, accuracy for the Phase 1 items decreased. However, the drop in accuracy in Phase 2 was not equal for all stimuli, specifically, the switch to Phase 2 was most detrimental for the Phase 1 edge items. A 2 (Phase) x 7 (Stimulus Magnitude) x 2 (Task) x 2 (Label) mixed-model ANOVA was conducted to determine if accuracy decreased from training to test, and, whether either instructions or responses labels moderated the drop in accuracy. The dependent variable was the AD scores.

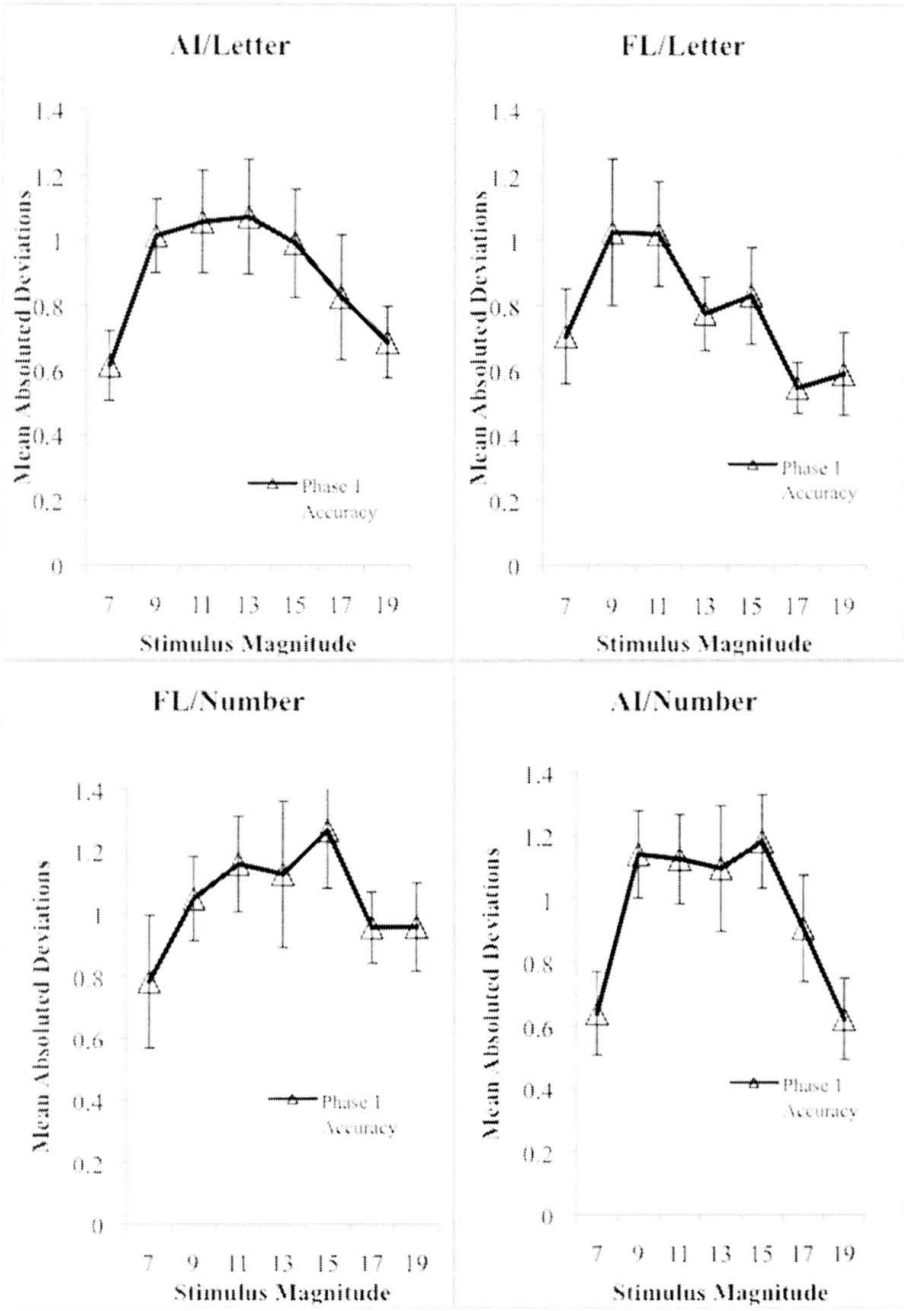

Figure 15: Mean absolute errors plotted as a function of training stimuli for each of the four groups. Error bars show the standard error of the mean.

There was an overall set-size effect. Errors increased from $M = 0.92$ ($SE = 0.047$) in the training phase to $M = 1.207$ ($SE = 0.083$) in the test phase ($F(1,36) = 17.29$, $MSE = 0.665$, $p < .001$). As is typical of the set-size effect, increasing the number of items did not affect all stimulus magnitudes equally, as evidenced by the significant Stimulus Magnitude x Phase interaction ($F(4.53, 162.94) = 12.25$, $MSE = 0.271$, $p < .001$). Figure 16 shows that increasing the number of stimuli increased error for the edge items, leaving the middle items relatively unaffected.

The three-way interaction between Phase, Stimulus Magnitude and Task was not significant ($F(4.53,162.94) = 1.89$, $MSE = 0.271$, $p = .106$), nor was the three-way interaction between Phase, Response Label and Task ($F(1,36) = 1.5$, $MSE = 0.271$, $p = .229$). Therefore, the overall set-size effect was not affected by how the participants were told to approach the task, or by the response labels used.

Neither the type of response label ($F(1,36) = 2.48$, $MSE = 1.897$, $p = .124$) nor the type of task ($F < 1$) had an overall effect on accuracy. Also, the main effect of Stimulus Magnitude was not significant (F (3.82, 137.59) = 1.89, $MSE = 0.555$, $p = .118$). No other effects were significant (all other Fs < 1).

Overall, the data show that when the number of items a participant must respond to was increased accuracy became worse. Finding a set-size effect in Experiment 3 is important because the procedure of Experiment 3 was more similar to a FL experiment than an AI experiment, yet, the data revealed a classic AI effect. It is worth pointing out that the set-size effect in Experiment 3 confounded set-size with feedback and therefore should be interpreted with caution. The point of the

set-size analysis was to provide preliminary evidence that a set-size effect is plausible with a FL task.

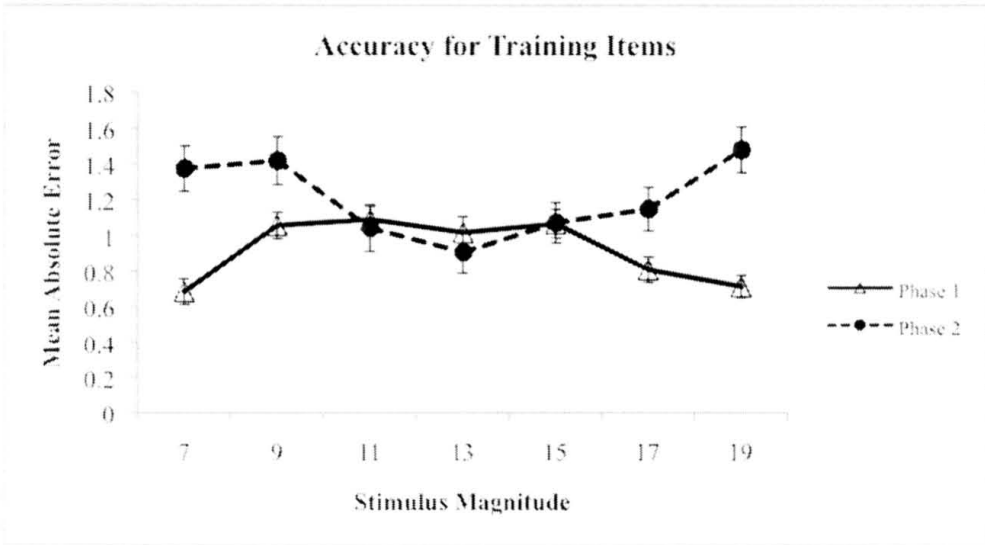

Figure 16: Mean absolute errors for training items in both the Training and Test phases averaged across groups. Error bars show the standard error of

So far, orienting participants toward either relational or item processing appears to have no effect on performance. Because the current research is exploratory, it is worthwhile to thoroughly examine the patterns of performance.

The mean absolute error is useful for measuring accuracy in general, however, absolute deviations may obscure directional trends in the data. In order to look at the direction of errors in Phase 1, the mean signed error was calculated for each stimulus. The signed error was used as the dependent measure in a 7 (Stimulus Magnitude) x 2 (Task) x 2 (Label) mixed model ANOVA.

There was a significant effect of Stimulus Magnitude ($F(3.53, 127.23) = 12.83$, $MSE = 0.433$, $p < .001$). Signed errors were more negative for the larger training items than for the smaller training items. The linear trend was significant ($F(1,36) = 37.13$, $MSE = 0.448$, $p < .001$). The quadratic and cubic trends were also significant (quadratic: $F(1,36) = 4.52$, $MSE = 0.28$, p = .04; cubic: $F(1,36) = 8.31$, $MSE = 0.17$, $p = .007$). Stimulus Magnitude did not interact with Response Label ($F(3.53, 127.23) = 1.23$, $MSE = 0.433$, $p = .302$), Task ($F < 1$), or the Response Label x Task interaction ($F < 1$).

There was a main effect for Task ($F(1,36) = 8.96$, $MSE = 0.877$, $p = .005$). Signed errors were more negative in the FL condition ($M = -0.255$, $SE = 0.079$) than in the AI condition ($M = 0.08$, $SE = 0.079$). The main effect for Response Label was also significant ($F(1,36) = 4.82$, $MSE = 0.877$, $p = .035$) with signed errors being more negative for the Number Label group ($M = -.021$, $SE = .079$) compared to the Letter Label group ($M = 0.35$, $SE = 0.079$). However both main effects were moderated by a significant Task x Label interaction ($F(1,36) = 5.72$, $MSE = 0.877$, $p = .022$).

In order to determine the nature of the Task x Label interaction, the difference between the Letter group and the Number group was examined separately for both task conditions. If participants performed a FL task with numeric labels, responses were more negative ($M = -0.512$, $SD = 0.32$) than if they performed a FL task with letter labels ($M = 0.002$, $SD = 0.439$; $t(18) = 2.99$, $p = .008$). However, if participants performed an AI task, the response labels did not make a

difference (Letter: $M = 0.069$, $SD = 0.375$; Number $M = 0.091$, $SD = 0.254$; $t\,(18) = -0.154$, $p = .879$; see Figure 17).

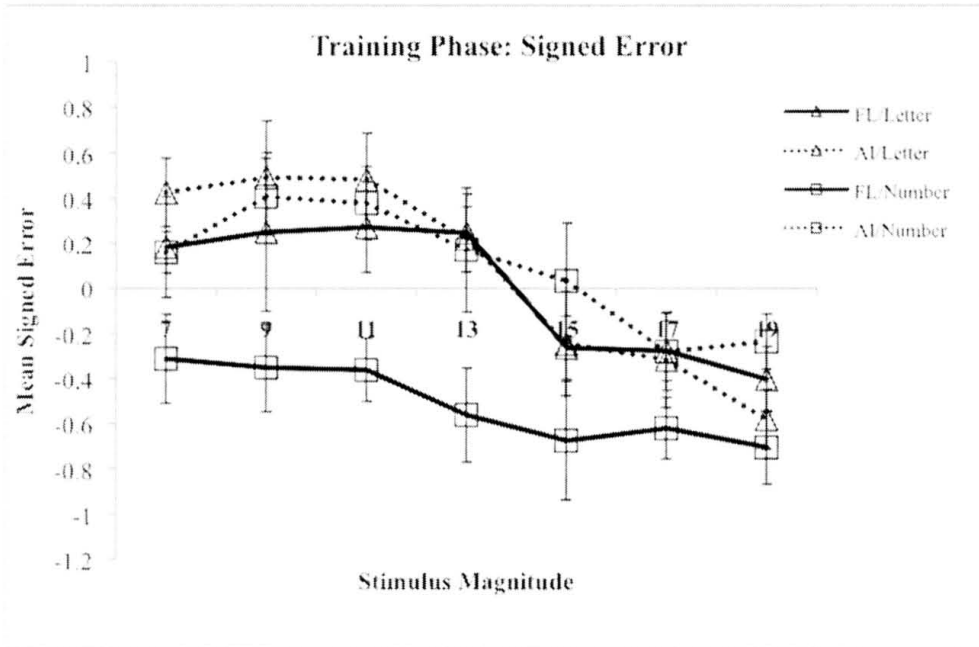

Figure 17: Mean signed errors in the Training Phase. Error bars show the standard error of the mean.

The third analysis for the training phase examined learning over trials. The 105 training trials were grouped into 7 blocks of 15 trials each. The mean absolute error was calculated for each block and used as the dependent measure. A 7(Block) x 2 (Task) x 2 (Response Label) mixed-model ANOVA was used to determine if either instructions or response label affected the rate of learning.

Figure 18 illustrates learning over blocks of trials. Generally, errors decrease across training blocks, with the most improvement early in training. The mean

error for the first block was 1.375 (*SE* = 0.078) and decreased to 0.737 (*SE* = 0.057) in the final block ($F(4.36, 156.85) = 19.5$, $MSE = 0.158$, $p < .01$; Linear Trend: $F(1, 36) = 44.05$, $MSE = 0.248$, $p < .001$). The quadratic and the 5th order trends were also significant (quadratic ($F(1, 36) = 26.4$, $MSE = 0.077$, $p < .001$; order 5 $F(1, 36) = 4.76$, $MSE = 0.069$, $p = .036$).

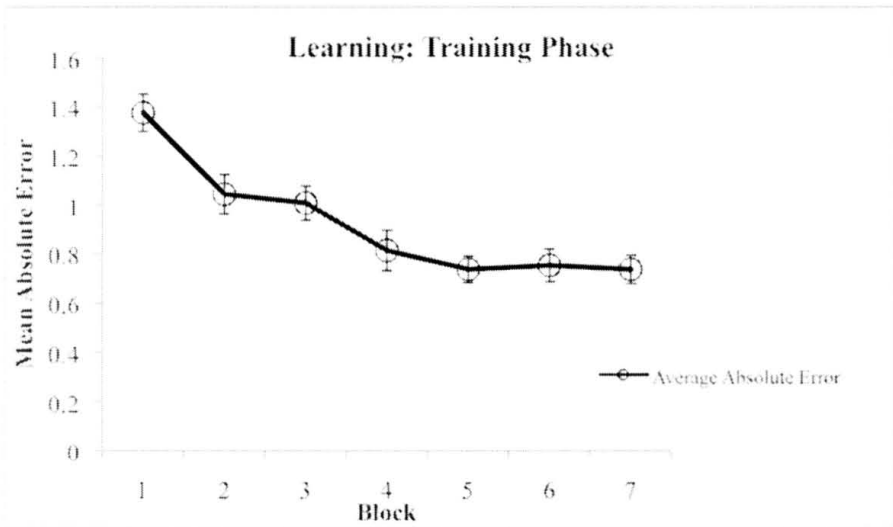

Figure 18: Mean absolute errors as a function of training block (averaged over all groups). Error bars show the standard error of the mean.

Because feedback was given throughout Phase 1, Phase 1 can be thought of as an identification experiment with 7 stimuli and 25 possible responses. Because the number of allowable responses is greater than the number of stimuli, Phase 1 was different from standard AI, yet, typical AI effects occurred.

Participants responded to items from the edges of the Phase 1 set more accurately than to items from the middle of the set, yielding the bow-shaped pattern

typical of AI. An important point is that, in Experiment 3, the bow-effect occurred even though participants had access to responses that are beyond the edges of the stimulus set. Because participants could make errors in both directions for the smallest and largest Phase 1 stimuli, the bow-effect cannot be due solely to the limited response options for edge items. However, even though smaller and larger responses were available, participants may have learned the set of valid responses and explicitly ignored the other response options. Therefore, although response options for the edge items were not objectively limited, they may be subjectively limited.

Participants seemed to know the set of possible responses, and restricted their responses accordingly. For example, incorrect responses to Stimulus 11 will usually be Response 9 or 13 (i.e., valid Phase 1 responses), rather than Response 10 or 12. In order to look at this pattern, I calculated the number of times each response was used incorrectly as a proportion of the total number of incorrect responses (calculated for each participant, then averaged). The data showed a saw-tooth pattern for responses across the training range (see Figure 19). Additionally, when the proportion of incorrect responses was calculated for only the first 50 trials a very similar pattern emerged. The similarity between the pattern of errors on the first 50 trials and pattern of errors on all trials suggests that participants quickly learned what response options were valid and limited their responses accordingly.

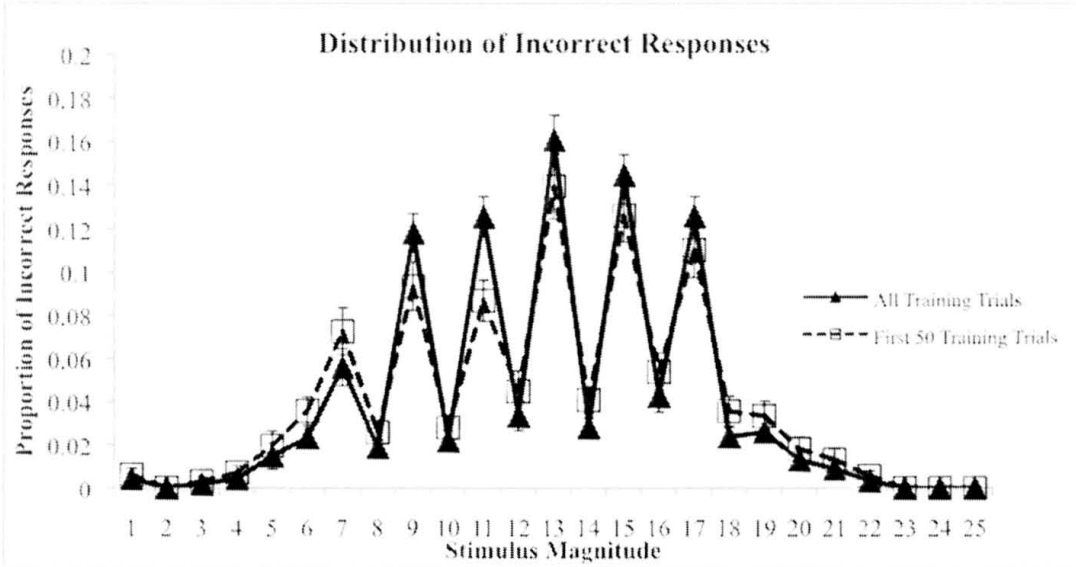

Figure 19: Proportion of incorrect responses in Phase 1 for each response category.

2.3.4.2 Phase 2/Test

The ability of participants to correctly infer the identity of novel stimuli was assessed by examining transfer performance. It was expected that if participants receive instructions that focus on the relationship between stimulus and response magnitudes, they would be able to use this information to accurately respond to novel items. On the other hand if the task is framed so that participants focus on the identity of individual items, transfer performance will be impaired.

The mean response was calculated for each stimulus magnitude. Figure 20 shows the mean response as a function of stimulus magnitude for the four experimental conditions. For all four conditions, mean responses appear to follow a

linear pattern with slight under-estimation occurring in the upper extrapolation region and slight over-estimation occurring in the lower extrapolation region.

In order to determine the exact pattern of errors, the mean signed error was calculated for each stimulus and used as the dependent measure. A 25 (Stimulus Magnitude) x 2 (Task) x 2 (Response Label) mixed-model ANOVA was conducted to determine if the pattern of errors differed among groups.

Stimulus magnitude had a significant effect on performance ($F(2.74, 98.58) = 50.29$, $MSE = 6.794$, $p < .001$). Figure 21 illustrates the pattern of errors; smaller stimuli tend to be overestimated whereas, larger stimuli tend to be underestimated.

Stimulus Magnitude did not interact with Response Label ($F < 1$) and the Stimulus Magnitude x Task x Response Label interaction was not significant ($F(2.74, 98.58) = 1.28$, $MSE = 6.794$, $p = .286$). There was no overall effect of Task or Response Label, and the Task x Response Label interaction was not significant (all Fs < 1).

The Task x Stimulus magnitude interaction was not significant ($F(2.74, 98.58) = 1.71$, $MSE = 6.794$, $p = .174$), however, visual inspection of Figure 21 suggests that the FL group may be more accurate than the AI group for a subset of stimuli.

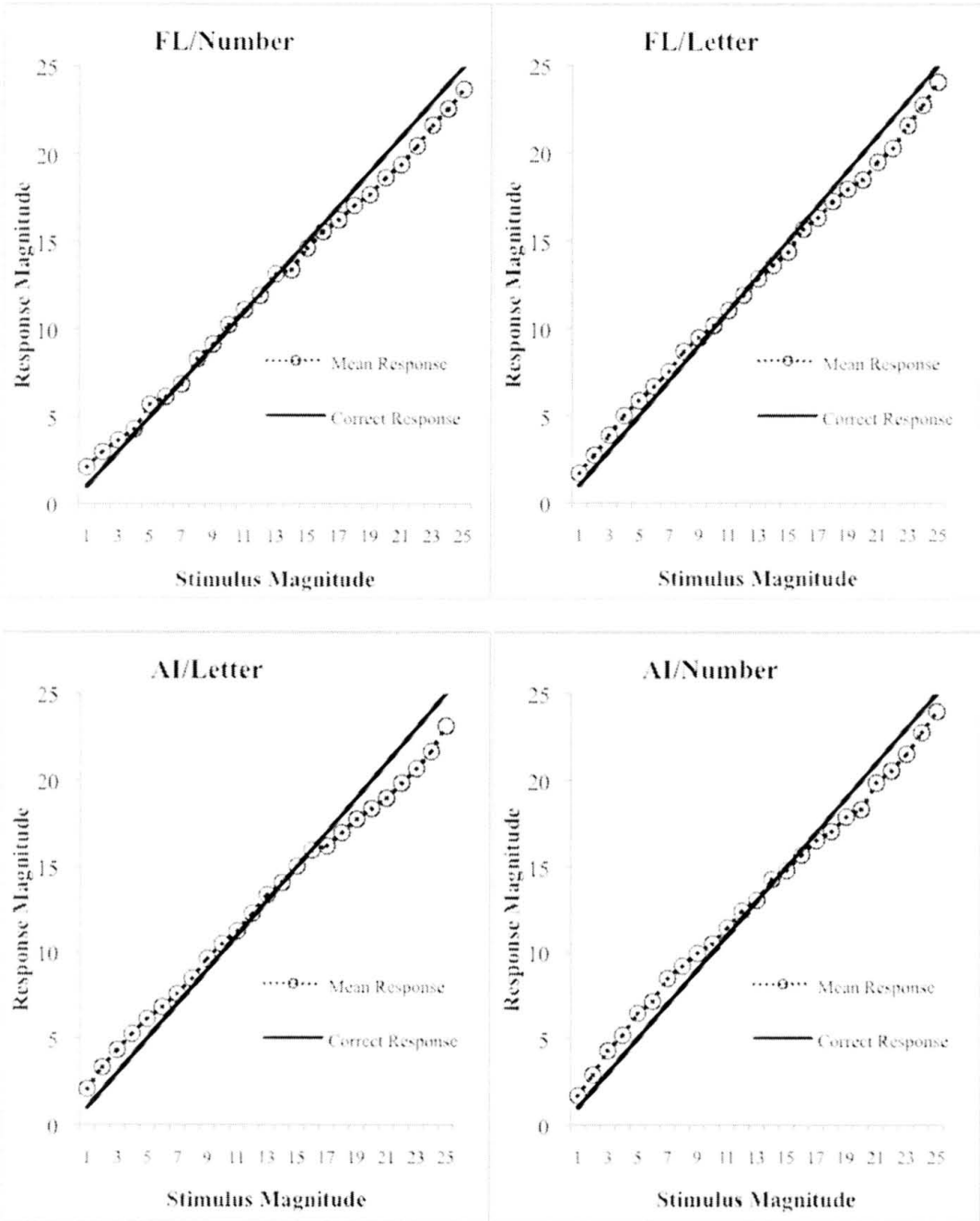

Figure 20: Mean response to each stimulus, plotted separately for each of the four conditions.

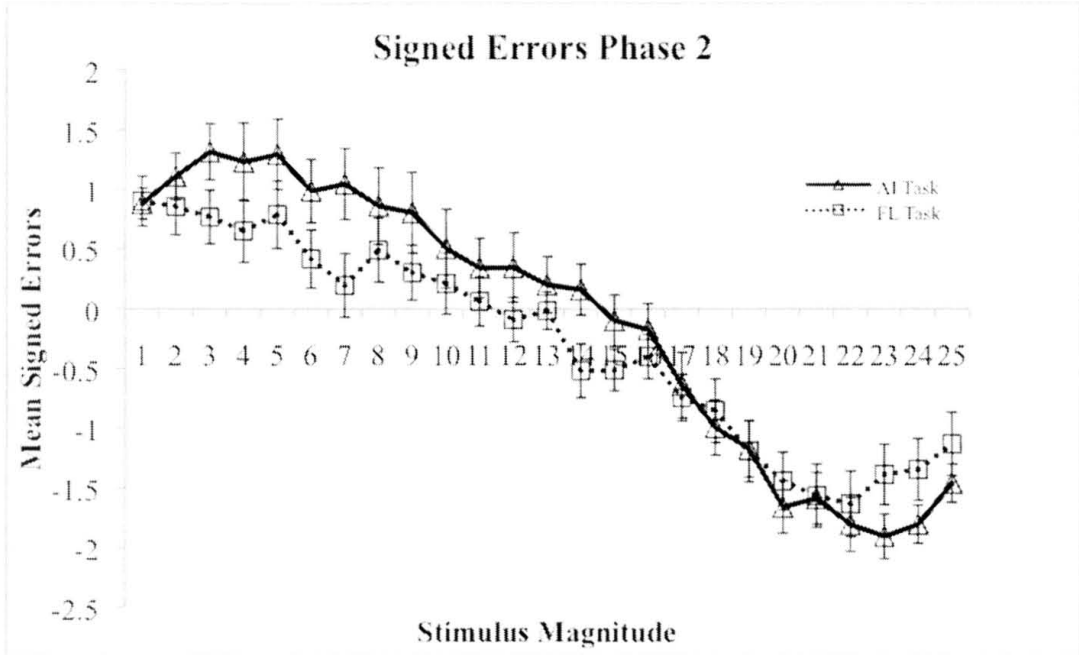

Figure 21: Mean signed error plotted as a function of stimulus magnitude for the FL and AI groups. Error bars show the standard error of the mean.

Before concluding that the type of task had no effect on participants' responses, a second analysis is warranted. The stimuli were grouped into the four important regions; lower extrapolation (stimuli 1-6), training items (stimuli 7,9,11,13,15,17,19), interpolation (stimuli 8,10,12,14,16,18), and upper extrapolation (stimuli 20-25). A 4 (Region) x 2 (Task) mixed-model ANOVA was used to determine if the type of task affected performance differently across stimulus regions. This analysis also allows for an examination of the Stimulus Magnitude main effect, with Stimulus Magnitude grouped by region. The mean signed error was the dependent measure.

There was a significant effect of Region ($F(1.47, 55.65) = 84.04$, $MSE = 1.029$, $p < .001$). Follow up paired t-tests confirmed the pattern implied by Figure 21. The items from the lower region ($M = 0.932$, $SD = 0.953$) were overestimated compared to the interpolation items ($M = -0.392$, $SD = 0.848$, $t(39) = -7.317$, $p < .001$) and the items from the upper region ($M = -1.56$, $SD = 0.883$) were underestimated compared to the interpolation items($t(39) = 8.944$, $p < .001$).

The Region x Task interaction was not significant ($F(1.47, 55.65) = 2.06$, $MSE = 1.029$, $p = .149$, observed power = .348). Therefore, if the marginal Task x Stimulus Magnitude interaction implied by Figure 21 is a real effect, the effect does not correspond to the important stimulus regions.

The pattern of signed errors suggests that people tend to underestimate items from the upper region and overestimate items from the lower region, regardless of either how the task is framed or the type of response labels. This pattern is not entirely consistent with previous FL studies that found underestimation in both the upper and the lower extrapolation regions.

The signed error (derived from the mean response) provides an estimate of the direction of errors, whereas the absolute error provides a more general estimate of accuracy. Delosh (1997) used absolute error as a dependent measure and found that when participants performed a FL task, there was no bow-effect. However, Delosh (1997) was looking for a serial position curve (accuracy plotted as a function of when the item was presented) rather than the bow-effect of AI experiments (accuracy plotted as a function of stimulus magnitude). If the flattening of the bow

effect is a result of how participants approach the task, the bow effect should only be present in the AI Instruction group.

A 25(Stimulus Magnitude) x 2(Task) x 2 (Response Label) mixed-model ANOVA was conducted with the mean absolute deviation (AD) from the correct response used as the dependent measure.

The magnitude of the stimulus had a significant effect on accuracy ($F(5.64, 202.85) = 6.15$, $MSE = 1.804$, $p < .001$). The pattern of errors did not constitute a typical bow-effect (quadratic trend, $F<1$). Figure 22 shows that accuracy took on a double-bow shape with accurate performance on the middle items as well as the typical advantage for the end items (order 4 trend, $F(1,36) = 24.59$, $MSE = 1.321$, $p < .001$).

Stimulus Magnitude did not interact with Task ($F < 1$) or Label ($F(5.64, 202.85) = 1.02$, $MSE = 1.804$, $p = .409$). None of the between-subjects effects were significant (all Fs< 1), nor was the Task x Label x Stimulus Magnitude interaction ($F(5.64, 202.85) = 1.61$, $MSE = 1.804$, $p = .15$).

The stimuli were grouped according to region (Lower, Training, Interpolation, Upper) and a 4(Region) x 2 (Task) x 2 (Label) mixed-model ANOVA was conducted. The Region x Task x Label interaction was not significant ($F(1.77, 63.59) = 1.56$, $MSE = 0.379$, $p = .219$) indicating that any potential differences among groups do not correspond to the important stimulus regions.

The effect of Region was significant ($F(1.77, 63.59) = 6.77$, $MSE = 0.379$, $p < .001$). One of the benchmark findings of the FL literature is that interpolation is

more accurate than extrapolation. Two paired t-tests compared accuracy for interpolation items to accuracy for lower extrapolation and upper extrapolation items. Interpolation was more accurate than extrapolation in the upper region (interpolation: $M = 1.39$, $SD = .448$; upper: $M = 1.66$, $SD = .79$, $t(39) = -1.964$, $p = .029$), however, there was no difference between interpolation accuracy and lower extrapolation accuracy (Interpolation: $M = 1.39$, $SD = .448$; lower: $M = 1.3$, $SD = .603$, $t(39) = 0.982$, $p = .166$). Therefore, Experiment 3 only partially supported the premise that interpolation is more accurate than extrapolation.

A critical factor may be that participants in Experiment 3 had to respond to both training items and interpolation items at test. Training items were responded to more accurately than interpolation items (Training: $M = 1.21$, $SD = 0.518$; Interpolation: $M = 1.39$, $SD = .448$, $t(39) = 6.109$, $p < .001$).

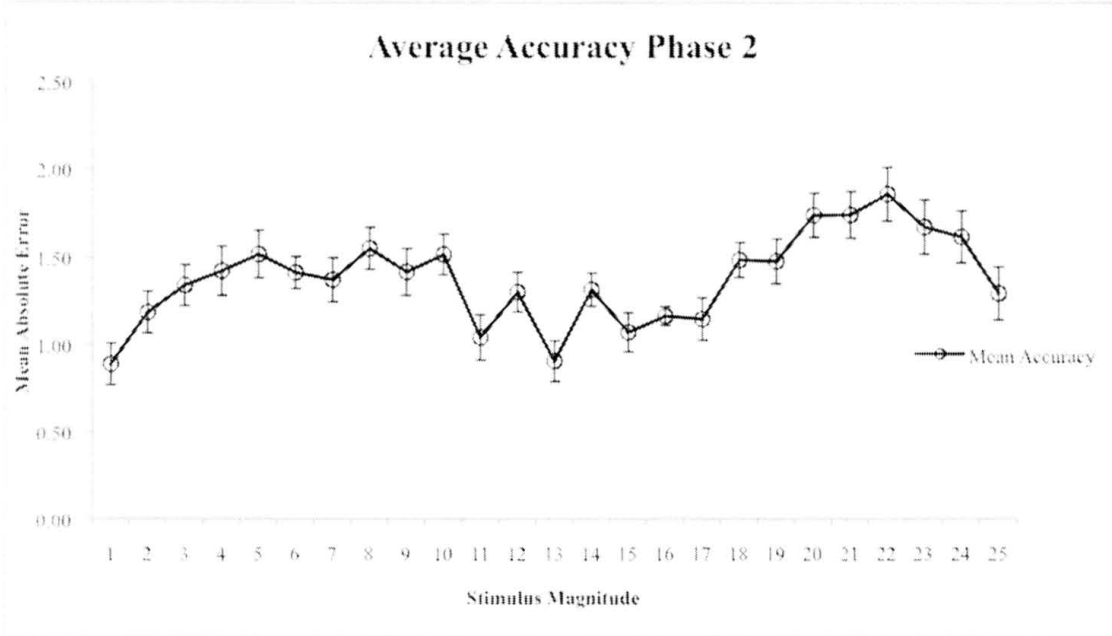

Figure 22: Mean absolute errors (averaged across conditions) plotted as a function of stimulus magnitude. Error bars show the standard error of the mean.

When accuracy was scored as the mean absolute error, there was no indication that changing how the task is framed affects the pattern of performance. This result disconfirms the prediction that participants given the FL task would show an attenuated bow-effect compared to participants given an AI task. For both groups, there was an accuracy advantage for middle items, resulting in a "M" shaped pattern, not the typical bow-shape.

Because feedback was not provided during the test phase, measuring performance relative to the "correct" response may not provide a complete picture of performance. In other words, how consistently a participant responds to a

particular item over multiple presentations provides a measure of performance relative to a participant's subjective S-R mapping.

In order to measure response consistency, the mean response for each stimulus was calculated for each individual participant. The mean absolute deviation for each stimulus was calculated relative to a participant's mean response to that stimulus (MDA) and submitted to a 25(Stimulus Magnitude) x 2(Instructions) x 2 (Response Label) mixed-model ANOVA to determine if either instructions or response labels affected consistency.

There was a significant effect of stimulus magnitude on consistency ($F(10.53, 379.11) = 10.7$, $MSE = .346$, $p < .001$). Participants were more consistent when responding to items from the ends of the stimulus range compared to the middle (quadratic trend; $F(1,36) = 86.05$, $MSE = 0.195$, $p < .001$). Several higher order trends were also significant, however the overall pattern in Figure 23 shows increased consistency for the edge items. Stimulus magnitude did not interact with Response Label ($F(10.67, 426.73) = 1.11$, $MSE = 0.346$, $p = .327$) or Task ($F < 1$). The 3-way interaction was also not significant ($F < 1$).

There was no main effect of Task ($F < 1$) and no interaction between Task and Response Label ($F < 1$). The effect of Response Label approached significance ($F(1,36) = 3.67$, $MSE = 1.476$, $p = .063$), suggesting that participants were somewhat more consistent when using letter response labels ($M = 0.704$, $SE = 0.054$) compared to numeric response labels ($M = 0.852$, $SE = 0.054$).

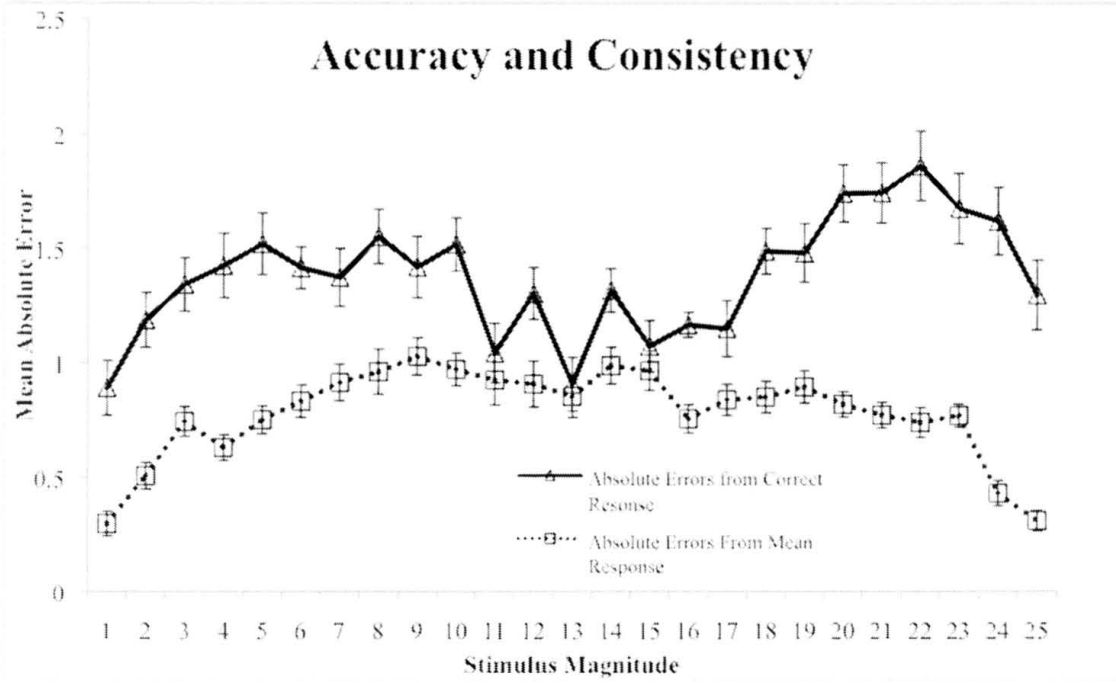

Figure 23: Mean absolute errors from the correct response, and mean absolute error from a participant's mean response as a function of stimulus magnitude. Error bars show the standard error of the mean.

Interestingly, when response consistency was measured the data take on a qualitatively different pattern compared to when accuracy was measured. Specifically, response consistency revealed a typical bow-effect, whereas, absolute deviations from the correct response showed an advantage for the middle items as well as an advantage for end items.

The results of Experiments 1 and 2 suggested that participants could infer the identity of novel stimuli within an AI task. Visual inspection of Figure 20 shows that mean responses to training range items were closer to the correct responses than items outside the training range. This result is consistent with the benchmark

FL result: interpolation is more accurate than extrapolation (Busemeyer et al., 1997). However, as pointed out previously, accurate interpolation in FL is often inferred from mean responses. The design of Experiment 3 allows for a stronger test of how well people are able to interpolate, specifically, whether item specific interpolation occurs or whether accurate interpolation is due to averaging. To clarify, if participants were presented with Stimuli 9 and 11 during training and then receive Stimulus 10 at test, the participant might not be able to differentiate Stimulus 10 from either 9 or 11 and might use Responses 9 and 11 when presented with Stimulus 10. Therefore, the mean response will be approximately 10 even though the participant never actually interpolated a response.

In order to see if item specific interpolation occurred in Experiment 3, the mean number of times each response was used was calculated. Figure 24 shows that participants rarely used interpolation responses and instead use the responses associated with the training items for interpolation items. Therefore, there seems to be little evidence for item specific interpolation in Experiment 3, rather, participants overwhelmingly used the Phase 1 responses when responding to interpolation stimuli.

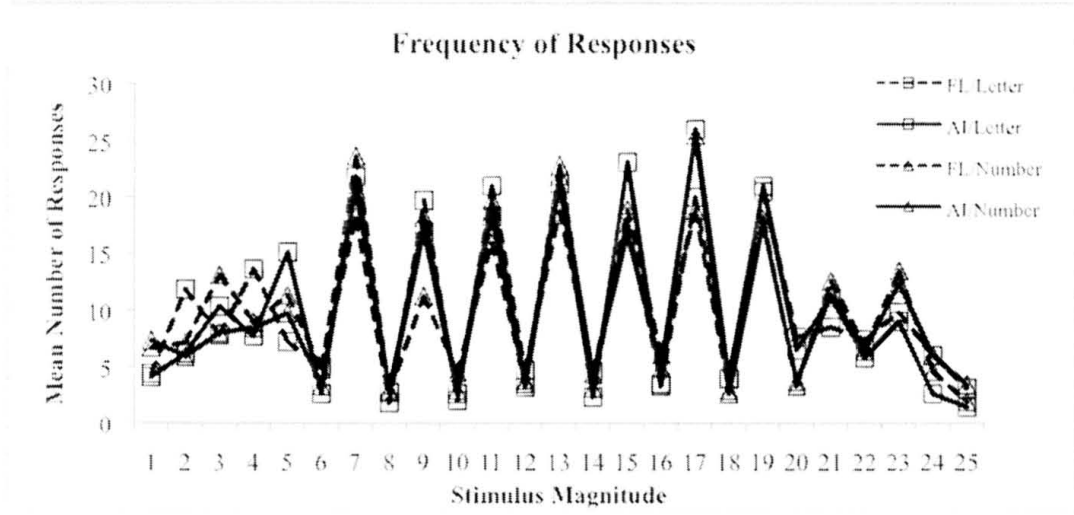

Figure 24: The mean number of times participants used each response category.

One possibility for participants' unwillingness to use interpolation responses is that the adjacent stimuli were not different enough to be perceptually discriminated and therefore interpolation could not occur because of a perceptual limit. In order to test this possibility, I examined trials in which the current stimulus was preceded by one of its immediate neighbours (e.g., Stimulus 5 followed by either Stimulus 4 or 6). Responses were then examined to determine if the direction of responding was the same as the direction of the stimulus change. If adjacent stimuli cannot be discriminated, the response should be the same on both trials (i.e., response repetition) and non-repetitions should be due to random error and therefore approximately evenly distributed on either side of the previous response. If adjacent stimuli can be discriminated, the response should change in the direction of the stimulus change.

Adjacent stimuli were presented on 733 trials. Of these trials, response repetitions (RR) occurred 346 times and response changes in the correct direction (CCD) occurred 370 times, leaving 17 responses that changed in the wrong direction. The CCDs were examined as a function of stimulus magnitude. The CCDs were calculated as a proportion of the number adjacent trials that occurred for that stimulus. Figure 25 plots CCD as a function of stimulus magnitude. The general pattern is that participants were more likely to shift their responses in the correct direction when the stimuli were from the ends of the stimulus range. Therefore, the items from the ends of the stimulus set appear to be easier to discriminate than items from the middle of the stimulus set.

If participants always made CCD responses and never repeated responses, it would provide strong evidence that the stimuli were different enough to be discriminated. The data indicate that response repetitions were very common; therefore, it is possible that neighbouring stimuli were too similar to allow interpolation to occur. However, participants rarely made responses in the wrong direction. If two neighbouring stimuli were perceptually indistinguishable, when a previous response is *not* repeated, responses should be equally likely to occur in the wrong direction as in the right direction. CCDs were much more frequent (n = 370) than response changes in the wrong direction (n = 17; Sign Test; $p < .001$). Because participants rarely made responses in the wrong direction there is some evidence that the stimuli were pair-wise discriminable. Also, the 30 pixel difference (equal to

approximately 8mm) between stimuli was similar to the stimulus differences used in previous FL studies (e.g., Brehmer, 1979; Kalish et al., 2004).

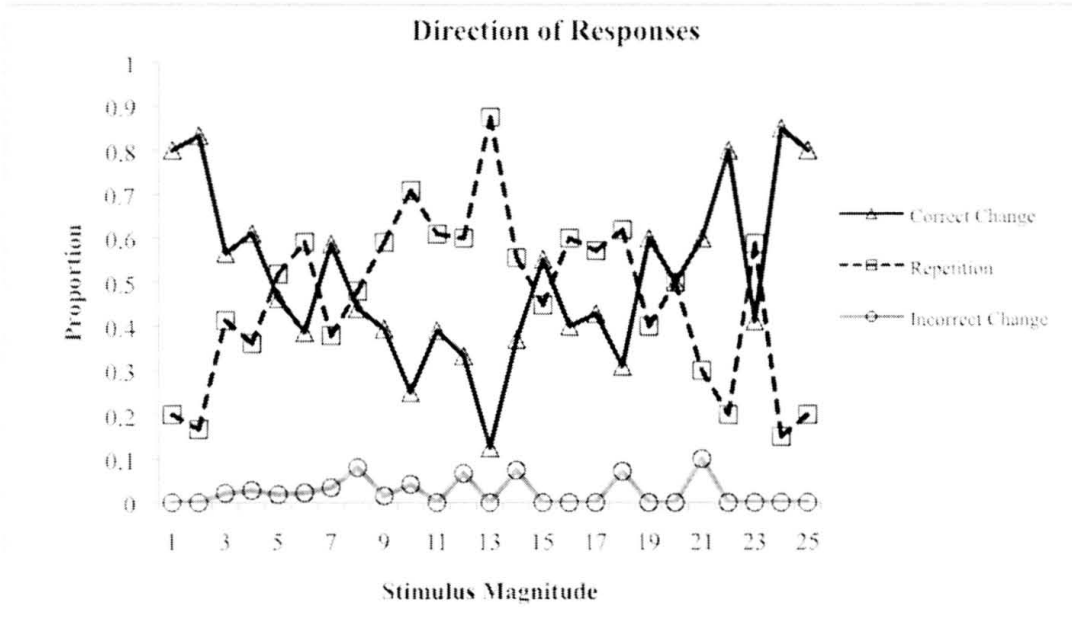

Figure 25: Plots the proportion of times responses to adjacent stimulus presentations were: repeated, changed in the correct direction, or, changed in the wrong direction.

2.3.5 Discussion

The goal of Experiment 3 was to compare FL and AI by manipulating how the task was framed, and the kind of response scale used. It was expected that drawing attention to the S-R relationship would result in a different pattern of performance than if attention was drawn to item identity. Specifically, transfer performance should be better if participants are given FL instructions than if they are given AI

instructions. The results of Experiment 3 did not support the conclusion that focusing on the functional relationship improves transfer performance.

When the results of Experiment 3 are considered overall, the data seem to support the conclusion that participants use similar processes in both AI and FL. However, one result points to a difference between AI and FL tasks: the interaction between Task and Label in the training.

The best evidence for differences between AI and FL comes from the Task x Label interaction during the training phase, when feedback should have made responses more similar among groups. The FL/Number group underestimated responses more than the FL/Letter group or the AI groups. This difference did not translate into a difference in accuracy, but rather, reflected a tendency for the FL/Number group to use lower response magnitudes. Research on numeric estimation suggests that people can have different representations of numeric magnitude (Seigler & Opfer, 2003); therefore, one possible explanation is that the numeric label determined the subjective response magnitude for the FL/Number group, while the ordinal response value determined the subjective response magnitudes for the FL/Letter and AI groups. In other words, the response label *210* may be subjectively larger when interpreted as a magnitude (i.e., FL instructions) compared to when it is interpreted as a label (i.e., AI instructions). Differences in the representation of the response magnitudes may account for the lower responses given by the FL/Number group.

Previous research in AI and related tasks, such as magnitude estimation, has shown that participants' responses can be shifted by giving them prior experience with a particular S-R mapping (Ward & Lockhead, 1970; West, Ward & Khosla, 2000). The current research provides some preliminary evidence that it is possible to shift the pattern of responding simply by changing how participants interpret the task. Because FL instructions with letter labels did not affect responses, the source of the effect may be due to the interpretation/mental representation of the response scale, not a distinction between relational and item-based strategies.

The results of Experiment 3 provides three strong lines of evidence that support the premise that AI and FL involve overlapping processes; the lack of significant differences between groups, the presence of classic AI patterns in the FL group, and the presence of FL patterns in the AI group.

Over multiple comparisons the AI and the FL groups were not significantly different from each other (with the exception of the Task x Label interaction during training). However, it is difficult to use null results to conclude that there is no difference between AI and FL tasks. Experiment 3 may not have had enough statistical power to detect differences between AI and FL. The lack of statistical power means a claim that AI and FL are essentially the same task is weakened. In future studies, the statistical power problem could be addressed by increasing the number of participants. Also, the prolonged testing period may have increased the amount of statistical noise because of participant fatigue, thus making the encoding strategy manipulation less influential as testing progressed.

An alternative to relying on null results is to determine how well performance on one task predicts performance on the other. For example, overall performance on a FL task can be viewed as output from an undefined cognitive model. If mean FL performance (the model's output) can accurately predict AI performance, it provides evidence that the same processes are involved in both tasks. In the case of Experiment 3, the mean responses of the FL/Number group almost perfectly predict the mean responses of the AI/Letter group (i.e., the two groups that should have been the most different; $R^2 = .997$, $F(1,24) = 8394.61$, $p < .001$; AI = 0.909 + 0.938(FL)) . This suggests that both tasks involve similar processes, and could potentially be explained using a common theory.

The pattern of responses also provides evidence that AI and FL involve similar processes. Participants were provided with feedback during Phase 1 of Experiment 3; therefore, the training phase was equivalent to an identification task with 7 stimuli and 25 allowable responses. The training phase of Experiment 3 showed that when performance was examined as a function of stimulus magnitude, participants responded more accurately to edge items than to middle items (i.e., bow-effect). In addition, when accuracy was measured across blocks of trials, accuracy did not continue to improve with more practice (i.e., asymptotic learning). Although the bow-effect and asymptotic learning are usually associated with AI, framing the task as FL did not change the pattern of performance.

When the training items were presented in the context of a larger set of stimuli (i.e., the test phase), accuracy for the training items decreased. This is

typical of the set-size effect found in AI, and, like the bow-effect, was not affected by the task. However, not all training items were affected to the same degree; specifically, most of the reduction in accuracy occurred for the edge training items. This pattern is the same as observed in Experiment 2 despite very different methods. Most notably, participants were provided with feedback throughout Experiment 2 but not during the test phase of Experiment 3.

Previous FL studies have shown that increasing the number of training items does not reduce accuracy during training (Delosh, 1997; Delosh et al., 1997). This lack a set-size effect is interesting because it stands in stark contrast to what would be expected given typical AI results. One possible explanation is that the lack of a set-size effect in previous FL experiments involves an interaction between three factors: the discriminability of the stimuli, the response spacing, and the measure of accuracy (i.e., absolute deviations). If a small training set and a large training set are taken from the same training range, the small set stimuli will be more widely spaced than the large set stimuli, making the small-set stimuli easier to discriminate. However, the small set also has a disadvantage because the valid responses are also widely spaced. The type of errors participants made in Experiment 3 suggested that participants quickly learned the valid responses and limited their responses accordingly. If participants only use the learned valid responses, absolute errors in the small set would be larger than absolute errors in the large set. For example, if the small set contains Stimuli 3, 5, and 7, an error on Stimulus 5 would probably be either Response 3 or 7 (i.e., absolute error of 2), even though participants have

access to Responses 4 and 6. However, if the large set contains stimuli 3, 4, 5, 6, and 7, an error on Stimulus 5 will probably be either Response 4 or 6 (i.e., absolute error of 1). Therefore, it may be easier to be exactly correct with a small stimulus set because of stimulus discriminability, but, when errors do occur, the errors will be relatively large. In contrast, it may be difficult to be exactly correct with a large number of stimuli (because the stimuli are more similar/confusable), but the magnitude of errors will be relatively small. These two effects may cancel each other out resulting in a null effect of set-size. If this explanation is correct, it implies that in tasks such as FL, participants may not treat continuous response scales as continuous, but rather constrain their responses to the set of learned valid response values.

In order to look for the bow-effect in the test phase of Experiment 3, the mean absolute deviations (AD) were used as the measure of accuracy. ADs followed a double-bow pattern in the test phase, not the typical single bow found in AI. When a measure of response consistency was used (the mean deviation from a participant's mean response; MDA scores), the data resembled a single bow pattern typical of AI.

Both AD and the MDA are measures of variability. The main difference between these measures is the reference point from which the variability is calculated. Because AD uses the correct response as its reference point, it can be considered a measure of how well participants have learned the correct S-R mapping. MDA, on the other hand, is a measure of performance that is independent

of the correct S-R mapping, because error is calculated relative to a participant's mean response to each stimulus.

Interestingly, these two measures show very different patterns. When the correct mapping is considered (ADs), participants are more accurate for the items in the middle of the stimulus set, as well as items at the ends of the stimulus set. This pattern is consistent with previous FL studies showing higher accuracy for training range items than for extrapolation items. In contrast, the MDA scores do not show the advantage for items from the middle of the set and are consistent with the single bow pattern typical of AI. In addition, MDA scores appear to be more accurate overall than AD scores. MDA scores may be more accurate because, essentially, any error that is due to incorrect S-R mapping is being ignored in the performance measure.

One way to interpret the pattern of MDA scores is to attribute them to the psychological discriminability of the stimuli. Items from the ends of the stimulus set may be easier to discriminate from their neighbours and therefore it is easier for participants to respond consistently to those items. If MDA scores represent effects attributable to stimulus characteristics and AD scores represent effects attributable to S-R mapping errors, the different pattern of results for the two measures suggests that these effects may be due to distinct processes. Theories of AI often distinguish between stimulus and response effects (see, Nosofsky, 1983), as well as effects due to S-R mapping (Lacouture & Marley, 1995). The AD and MDA scores may provide a intuitive method for measuring different components of AI and FL performance;

however, more research is needed to determine the validity of the measures and specify their underlying assumptions.

All groups in Experiment 3 demonstrated effects typically associated with AI (bow-effect, a set-size effect and asymptotic learning). If the bow-effect, the set-size effect and asymptotic learning are important phenomena in the AI paradigm, and these effects are found in a FL tasks it suggests that whatever processes cause these effects in AI are also affecting FL performance.

The AI/Letter group was the best approximation of an AI task because the instructions focused on item identity, and the response labels were discrete categories. Even though the experimental conditions for the AI/Letter group did not emphasize learning a functional S-R relationship, the AI/ Letter group's mean responses to novel items were still quite accurate. The mean responses followed the general pattern typical of FL experiments, with accurate performance on items from the training range, and worse performance on extrapolation items. The accurate transfer performance of the AI/Letter group suggests that even in a simple perceptual identification task, people are able to respond accurately to novel items.

One of the benchmark findings in the FL literature is that interpolation performance is more accurate than extrapolation performance (Busemeyer et al., 1997; Delosh et al., 1997). The results of Experiment 3 showed that mean responses in the training region were more accurate than responses outside the training region, hence replicating the advantage for interpolation over extrapolation found by Delosh et al. (1997) and Kwantes and Neal (2006). However, closer

examination of the data suggests some potential limitations. First, when the mean absolute deviations are measured, interpolation performance only holds an advantage over the upper extrapolation items, not the lower extrapolation items. The mean absolute deviation score may not have revealed an interpolation advantage because the AD scores are a stricter measure of accuracy compared to mean responses. That is, overestimation and underestimation will cancel each other out when the mean response is calculated, but not when the mean absolute deviation scores are calculated. Participants rarely used interpolation responses; therefore, accurate interpolation appears in Experiment 3 mainly because of averaging responses over stimulus presentations.

Experiment 3 revealed little evidence for item-specific interpolation; however, the FL/Number group may have perceived the response magnitudes as discrete categories rather than a continuous scale, and this may have hindered interpolation by facilitating bias toward specific training responses. If the response scale had been continuous (with no scale markings), it would be more difficult to remember the exact location of previous responses, and, therefore, participants would be less likely to be biased toward any specific response value. However, even if a continuous response scale is used, interpolation responses may come from two distinct response distributions associated with the nearest training items. In order to determine whether item-specific interpolation occurs in FL, future studies could use a continuous response scale without intermediate labels and examine the distribution of responses to interpolation items. Item-specific interpolation would

reveal itself as a uni-modal distribution centered on the correct response value, whereas, a bi-modal distribution with the peaks centered over the nearest training responses would be evidence against item-specific interpolation and would be more indicative of stimulus generalization.

Previous research has shown that people can interpolate even when the response categories are discrete (Levine, 1960). In addition, people are able to perform a wide variety of inference tasks (e.g., transitive inference, categorization of novel exemplars, etc.), therefore, it is likely that item-specific interpolation could occur in FL if the experimental procedure better supported interpolation. It could be argued that the stimuli in Experiment 3 were too similar to each other to allow participants to discriminate interpolation items from training items, and this is why people did not interpolate. However, when neighbouring stimuli were presented on consecutive trials, participants' responses rarely broke monotonicity. This suggests that the lack of interpolation was not due to a perceptual limit. Determining the factors that allow for item-specific interpolation with a continuous response scale would have both theoretical and practical implications. Practical application of this knowledge may include determining the best kind of scales or dials to use on equipment, as well as determining the most efficient training methods.

Intuitively, the distribution of interpolation responses in Experiment 3 is more consistent with exemplar-based theories than rule-based theories. An exemplar approach would predict that when presented with a novel stimulus, the responses associated similar training stimuli would be recalled. A rule-based

approach would probably predict a more continuous distribution of responses; however, both of these predictions are speculative, and a more formal test is necessary in order to differentiate the two theories. Also, although responses in the training range are more consistent with exemplar theories, a strict exemplar theory would have trouble accounting for the relatively accurate extrapolation performance.

Strong evidence that participants in the FL task learned a relational concept and participants in the AI group learned the identity of specific items would involve accurate extrapolation for the FL group and poor extrapolation for the AI group; this pattern was not found in Experiment 3. However, it is important to recognize that the task manipulation involved only changing how the task was framed (i.e., instructions) and the response method, and was therefore a relatively weak manipulation. Additionally, both groups were informed of the test phase at the beginning of the experiment. Informing the AI group of a test phase may have caused them to pay more attention to the S-R relationship during training in order to respond accurately during test. Therefore, both the AI and FL groups may have approached the task in similar ways, reducing the strength of the Task manipulation.

The effect of instructions on performance has sometimes been found to influence participants' responses in tasks such as probability learning; a task similar to FL (Brehmer & Kuylenstierna, 1980). One possibility is that FL instructions give meaning to the stimuli and responses, causing participants become more engaged in

the task. This engagement may result in better performance compared to when abstract stimuli and responses are used (as the case with AI). The results of Experiment 3 did not find improved accuracy for the more engaging FL instructions, therefore there seems to be no differences in participant motivation between the FL and AI groups. The possibility of different levels of engagement/motivation may be an important factor to consider in future research. For example, it may be necessary to provide a cover story for the identification group as well as the FL group in order to equate how interested participants are in the task.

Delosh et al. (1997) found that specifically telling participants to learn the functional S-R relationship did not change the pattern of responses compared to when the S-R relationship was not emphasized. However, even when the functional relationship was not emphasized, participants were still told that the stimulus and response magnitudes represented the values of variables (amount of growth hormone and plant height). The use of these labels may have induced participants to focus on a predictive relationship despite not being instructed to do so. The results of Delosh et al. in combination with the results of Experiment 3 suggest that specifically looking for a functional relationship is not necessary for accurate transfer.

Although there seems to be little evidence to suggest that participants used different strategies or processes for AI and FL in Experiment 3, a stronger manipulation may show different results. For example, not informing participants of the test phase may accentuate differences in encoding strategy. Also, positive

linear functions are known to be the easiest to learn and the most intuitive, therefore, using a less intuitive function (e.g., exponential, quadratic etc.) might yet highlight differences between FL and AI processes.

Overall, the results of Experiment 3 suggest a significant amount of overlap between AI and FL tasks. Three general findings support the idea that the same processes are involved in AI and FL. First, classic AI effects appeared for all groups. Second, the type of task did not change transfer performance. Finally, the strong correlation between the AI/Letter group's mean responses and the FL/Number group's mean responses suggests that both tasks could be explained with a common theory or model.

The best evidence for differences between AI and FL comes from the pattern of responses in the training phase. During training, the FL/Number group tended to use lower responses than the FL/Letter or the AI groups. Tentatively, this pattern may be better explained by differences in the mental representation of the response scale, not a distinction between relational and item-based strategies.

Chapter 3 General Discussion

There are several reasons why AI and FL may be similar tasks. First, both AI and FL can be viewed as conceptual tasks. Second, the stimuli in some FL experiments can be quite similar to the kind of stimuli used in AI experiments (e.g., line length). Third, and most importantly, the congruent S-R mapping used in AI means there is a continuous relationship between the stimulus and response values. This continuous mapping means that an AI task can be solved by learning the correct label for each stimulus, and/or, by learning the functional relationship between the stimulus and response scales.

The experiments presented in this paper demonstrate several interesting similarities and differences between AI and FL and the methods used in the respective paradigms. Although, a claim that AI and FL are essentially the same task is weakened by a lack of statistical power in Experiment 3, the overall pattern of results suggest a significant amount of overlap between the tasks

3.1 Performance Measures

Previous FL and AI research would suggest that FL performance is more accurate than AI performance. The results of Experiments 1 and 2 indicate that people may appear to be more accurate in FL tasks because of how accuracy is measured in the respective tasks. If the participants' mean response is used as the measure of accuracy in an AI task, participants appear to be very accurate, however, when the proportion correct is used as the dependent measure, accuracy appears

much worse. This is not surprising because proportion correct is a stricter measure of accuracy than mean response.

Experiment 1 also showed that the pattern of performance can be changed depending on how accuracy is assessed. For example, Experiment 1 showed a bow-effect when the proportion correct was used (typical of AI experiments), but when the mean response was used, participants tended to underestimate the larger stimuli, which is arguably similar to the underestimation that occurs in the upper extrapolation region of FL experiments. A bow-effect for mean responses would be demonstrated if the mean response to the larger items trended back toward the S-R function line (this pattern was seen in Experiment 2).

Delosh et al. (1997) found that people underestimated a positive linear function in the upper and lower extrapolation regions. Kwantes and Neal (2006) found that underestimation was more reliable in the lower extrapolation region than in the upper extrapolation region. Experiment 1 revealed that, when the stimulus magnitudes increased by a constant, underestimation occurred for the larger items, even though feedback was given on all trials. Experiment 2 used stimuli that were geometrically spaced and larger on average than Experiment 1 stimuli. Experiment 2 found that, with the exception of edge items, there was a general tendency to underestimate. Speculatively, underestimation of the upper extrapolation region, found in FL studies, may be (at least partially) a perceptual phenomenon rather than a conceptual one. In other words, underestimation may

not be due to extrapolation processes, but larger stimuli may be generally underestimated.

Although larger stimuli seem to be underestimated in general, the pattern of under/overestimation may be due to how people use the response continuum and not how they perceive the stimuli. Musielak, Chasseigne and Mullet (2006) compared FL with positive linear, negative linear, U-shaped, and inverted U-shaped functions. The patterns of responses found by Musielak et al. suggest that the *response* magnitude, not the *stimulus* magnitude, controls the pattern of over/underestimation. For example, when the function was positive linear, larger stimuli were underestimated and smaller stimuli were overestimated; in contrast, when the function was negative linear, the larger stimuli were overestimated and the smaller stimuli were underestimated. This pattern of results is consistent with results from magnitude estimation studies showing that people have a bias toward using responses from the middle of the response range (i.e., contraction bias; Poulton, 1979). It is possible that the pattern of extrapolation found in FL is due to both the psychological representation of the functional concept and a general response bias. Future studies may try to separate these two effects by manipulating training region and stimulus magnitude independently.

3.2 The Bow-Effect

All three experiments presented in this thesis revealed that the edge items of the stimulus set have an advantage over items from the middle of the set. However, some important qualifications need to be considered. When the number of items

was increased by adding new items to the ends of the initial training range (Experiment 2), items that were previously edge items held an advantage over the intermediate items. Speculatively, adding items to the ends of the stimulus set may have provided a way for participants to break the stimulus set down into distinct sections (small, middle, large) and this organization may have aided performance. For example, the old edge items may have been used as anchors or *subjective standards* from which intermediate items were judged (Eriksen & Hake, 1957; Petrov & Anderson, 2005). Other AI studies have been able to modify the bow-effect through different means, such as stimulus spacing (Lacouture, 1997; Neath et al., 2006) or by presenting some items more often (Dodds, Donkin, Brown, Heathcote & Marley, 2011). The present research provides an additional demonstration that changing the experimental procedure can modify the bow-effect in an AI task.

Experiment 3 also showed a bow-effect during training, and this effect was not modulated by how the task is framed. If FL and AI are fundamentally different tasks, a stronger bow-effect was expected for the AI group than for the FL group. Kwantes and Neal (2006) found that FL accuracy was relatively constant across all training stimuli (i.e., no bow-effect). However, there are some important differences between the methods used by Kwantes and Neal and the methods used in Experiment 3. For example, Kwantes and Neal presented stimulus and response values numerically as well as graphically, thus providing participants with more information about stimulus and response identity. Under the conditions of the

training phase used in Experiment 3, both FL and AI strategies resulted in a bow-shaped pattern of accuracy.

Experiment 3 showed modulation of the bow-effect in the test phase, with participants being more accurate on the middle items as well as edge items. However, this modulation of the bow-effect was only evident when participants' responses were scored in relation to the correct response. When response consistency was the dependent measure, the middle items no longer had an advantage, and a more typical bow-effect emerged. A plausible explanation is that edge stimuli are perceptually more discriminable from their neighbours than items from the middle of the stimulus set, and this allows participants to respond more consistently to edge items.

3.3 Accuracy and Response Patterns

In Experiment 3, despite participants being no more accurate in the FL condition than in the AI condition during training, how the task was framed did affect the direction of responses. During training, participants in the FL/Number group had lower mean responses than the FL/Letter group or the AI groups. The interaction suggests that the different response pattern is not solely due to the response method/instructions, or the numeric labels, but rather it is the combination of both factors that affects mean responses. Speculatively, FL instructions may have affected how participants represented the numeric response values. For example, Response *210* may have been interpreted as a magnitude in the context of FL instructions, but interpreted as a label in the context of AI

instructions. The lower response magnitudes given by the FL/Number group may have been the result of a bias against large response magnitudes, or, perhaps a non-linear representation of the response scale. Although the exact reason for the lower FL/Number responses is unclear, the result demonstrates the importance of context on performance.

3.4 Learning

Previous FL studies have plotted learning over trials; these graphs usually show that, when the function is linear, most of the learning occurs early in training and performance does not continue to improve over all training blocks. This learning pattern found in FL studies parallels the pattern of asymptotic learning that occurs in AI. When there was equal stimulus spacing (Experiments 1 and 3) performance improved quickly then leveled off, replicating previous FL and AI findings. Interestingly, when the stimulus spacing increased geometrically (Experiment 2), performance gradually improved over all training blocks.

If the magnitude of the stimuli increases by a constant, it could be argued that this represents a positive linear function between the stimulus and response scales. If, on the other hand, the same response scale is used but the stimuli increase geometrically, there is a non-linear function between stimulus magnitude and response scales. The gradual improvement across training blocks when the stimulus spacing was geometric, and the quick, asymptotic learning when the stimulus spacing was constant is similar to FL studies that show that linear functions are learned more quickly than non-linear functions. However, a

comparison of learning rates between Experiments 1 and 2 is qualitative and therefore future research could examine the effect of stimulus spacing on learning rates under more controlled conditions.

Slower learning of geometrically spaced stimuli in an AI task raises some interesting questions about the difference between learning linear and non-linear functions. When linear and non-linear functions are compared in FL (e.g., Delosh et al., 1997), the same stimulus values are used for both linear and non-linear groups; this means that the difference between the functions is in the spacing of the responses, not the spacing of the stimuli. If the S-R function is what determines task difficulty, then adjusting the spacing of the stimuli and responses independently of the function could help clarify the issue. For example, if exponentially-increasing stimuli were mapped on to exponentially-increasing response values, the S-R relationship would be linear and should be easy to learn. If the response spacing increased exponentially while the stimulus spacing increased by a constant (or vice versa), the S-R relationship would be non-linear and should be more difficult to learn. By manipulating the stimulus and response spacing independently of the mathematical function, it may be possible to determine whether the formal function is the important variable in FL.

Different effects of stimulus spacing and response spacing have been explored in AI tasks. For example, Bahrick and Nobel (1961) found that when responses were widely spaced, accuracy was better when the stimuli were also widely spaced, compared to when the stimuli were narrowly spaced. However,

when the response spacing was narrow, the stimulus spacing did not make a difference. Exploring the effects of stimulus and response spacing may be a productive avenue within the FL paradigm in order to determine if the functional S-R relationship is the factor that influences different performance levels and learning rates.

3.5 Interpolation

Experiments 1 and 2 provided some evidence that participants could infer the correct responses for specific items; however, the response distributions in Experiment 3 provided evidence against item-specific interpolation. It is likely that item-specific interpolation can occur under appropriate circumstances. However, two questions need to be addressed in future research. The first question is whether item-specific interpolation occurs with the kind of stimuli and response methods used in FL experiments. Ferrando (2003) explored the difference between continuous and discrete response scales when participants respond to questionnaire items (e.g., a personality instrument). When continuous response scales were used, people tended to limit their responses to a few points on the continuous scale. In the case of a FL experiment, participants may learn a set of discrete response values during training and continue to use these responses for interpolation items, resulting in accurate mean performance but no item-specific interpolation.

A second question regarding interpolation is: What information does the participant use? A rule-based FL approach may assume that participants are

learning the formal function and are able to use this abstract information to respond to new items. However, item-specific interpolation could also occur by a deductive process based on exemplar knowledge. For example, when presented with a new intermediate stimulus (e.g., Stimulus 5), the participant might be able to respond correctly by recognizing that the new stimulus is too large to be 4 and too small to be 6. Assessing the merits of rule-based and exemplar-based theories is one of the main theoretical issues within the FL literature (Kalish et al., 2004; Koh & Meyer, 1991; McDaniel & Busemeyer, 2005). The apparent lack of interpolation in Experiment 3 suggests that examining the response distributions may provide a way of determining which theories are more correct.

3.6 Summary

The goal of the present thesis was to compare AI performance to FL performance. The congruent S-R mapping that is usually present in AI means that the AI task is solvable by either remembering item-specific information, learning the functional relationship between the stimulus and response scales, or, a combination of both processes.

Three main findings speak to the similarities between FL and AI. Comparing previous FL and AI studies leaves the impression that FL performance is much more accurate than AI performance. Experiments 1 and 2 demonstrated that the appearance of highly accurate FL performance is probably due the measures of accuracy used in the respective tasks. Second, classic AI effects occurred in a FL task. Finally, participants were equally adept at extrapolation/interpolation,

regardless of whether they were instructed to learn the functional relationship or were told to memorize the identity of stimuli. These three general findings suggest a significant amount of overlap between AI and FL processes.

A fourth finding is perhaps the most interesting, as it speaks to a potential difference between FL and AI. Manipulating how participants interpret the task can affect the responses they make. Specifically, if participants were given FL instructions with numeric labels they tended to use lower response magnitudes than if they were given FL instructions with letter labels, or, if they are given AI instructions. Although interesting, more research is needed to determine the exact nature of the effect. Tentatively, the effect may be due to differences in how the response values are psychologically represented, not a difference between relational and item based strategies.

References

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General, 120*(1), 3-19. doi:10.1037/0096-3445.120.1.3

Bahrick, H. P., & Noble, M. (1961). On stimulus and response discriminability. *Journal of Experimental Psychology, 61*(6), 449-454. doi:10.1037/h0045585

Brehmer, B. (1979). Effect of practice on utilization of nonlinear rules in inference tasks. *Scandinavian Journal of Psychology, 20*(3), 141-149. doi:10.1111/j.1467-9450.1979.tb00694.x

Brehmer, B., & Kuylenstierna, J. (1980). Content and consistency in probabilistic inference tasks. *Organizational Behaviour and Human Performance, 26*(1), 54-64. doi:10.1016/0030-5073(80)90046-X

Brehmer, B., & Svensson, C. (1976). Learning to use functional rules in inference tasks. *Scandinavian Journal of Psychology, 17*(4), 313-319. doi:10.1111/j.1467-9450.1976.tb00246.x

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. R. Shanks (Eds.), *Knowledge, concepts and categories* (pp. 408-437). Cambridge, MA US: The MIT Press.

Busemeyer, J., McDaniel, M. A., & Byun, E. (1997). The abstraction of intervening concepts from experience with multiple input–multiple output causal environments. *Cognitive Psychology, 32*(1), 1-48. doi:10.1006/cogp.1997.0644

Chase, S., Bugnacki, P., Braida, L. D., & Durlach, N. I. (1983). Intensity perception. XII. Effect of presentation probability on absolute identification. *Journal of the Acoustical Society of America, 73*(1), 279-284. doi:10.1121/1.388700

Costall, A., Platt, S., & MacRae, A. (1981). Memory strategies in absolute identification of 'circular' pitch. *Perception & Psychophysics, 29*(6), 589-593. Retrieved from EBSCOhost.

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. doi:10.1037/h0035439

Delosh, E. (1997). Effect of mnemonic variables on function and category learning (Doctoral dissertation, Purdue University, 1997). *Dissertation Abstracts International, 58*, Retrieved from EBSCOhost..

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 968-986. doi:10.1037/0278-7393.23.4.968

Dodds, P., Donkin, C., Brown, S. D., & Heathcote, A. (2011). Increasing capacity: Practice effects in absolute identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 477-492. doi:10.1037/a0022215

Dodds, P., Donkin, C., Brown, S. D., Heathcote, A., & Marley, A. (2011). Stimulus-specific learning: disrupting the bow-effect in absolute identification. *Attention, Perception & Psychophysics*, *73*(6), 1977-1986. doi:10.3758/s13414-011-0156-0

Eriksen, C. W. (1958). Effects of practice with or without correction on discrimination learning. *The American Journal of Psychology, 71*, 350-358. doi:10.2307/1420079

Eriksen, C. W., & Hake, H. W. (1957). Anchor effects in absolute judgments. *Journal of Experimental Psychology, 53*(2), 132-138. doi:10.1037/h0047421

Ferrando, P. J. (2003). A kernel density analysis of continuous typical-response scales. Educational and Psychological Measurement, 63(5), 809-824. doi:10.1177/0013164403251323

Fitts, P. M., & Deininger, R. L. (1954). S-R compatibility: Correspondence among paired elements within stimulus and response codes. *Journal of Experimental Psychology, 48*(6), 483-492. doi:10.1037/h0054967

Garner, W. R., & Hake, H. W. (1951). The amount of information in absolute judgments. *Psychological Review, 58*(6), 446-459. doi:10.1037/h0054482

Garner, W. R. (1953). An informational analysis of absolute judgments of loudness. *Journal Of Experimental Psychology, 46*(5), 373-380. doi:10.1037/h0063212

Garner, W. R. (1974). *The processing of information and structure*. Oxford England: Lawrence Erlbaum.

Hake, H. W., & Garner, W. R. (1951). The effect of presenting various numbers of discrete steps on scale reading accuracy. *Journal Of Experimental Psychology, 42*(5), 358-366. doi:10.1037/h0055485

Hu, G. (1997). Why is it difficult to learn absolute judgment tasks?. *Perceptual and Motor Skills, 84*(1), 323-335. doi:10.2466/PMS.84.1.323-335

Hunt, R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning & Verbal Behavior, 20*(5), 497-514. doi:10.1016/S0022-5371(81)90138-9

Hunt, R., & Seta, C. E. (1984). Category size effects in recall: The roles of relational and individual item information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(3), 454-464. doi:10.1037/0278-7393.10.3.454

Juliusson, A., Gamble, A., & Gärling, T. (2006). Learning unit prices in a new currency. *International Journal Of Consumer Studies*, 30(6), 591-597. doi:10.1111/j.1470-6431.2006.00498.x

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychological Review, 111*(4), 1072-1099. doi:10.1037/0033-295X.111.4.1072

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*(2), 289-305. doi:10.1037/0278-7393.31.2.289

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(5), 811-836. doi:10.1037/0278-7393.17.5.811

Kwantes, P. J., & Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(5), 1019-1030. doi:10.1037/0278-7393.32.5.1019

Kwantes, P. J., & Neal, A. (2003). Function learning: An exemplar account of extrapolation performance. Defense R & D Canada – Toronto. Retrieved October 7, 2011, from http://pubs.drdc.gc.ca/PDFS/unc47/p521106.pdf

Lacouture, Y. (1997). Bow, range, and sequential effects on absolute identification: A response-time analysis. *Psychological Research/Psychologische Forschung, 60*(3), 121-133. doi:10.1007/BF00419760

Lacouture, Y., & Lacerte, D. (1997). Stimulus modality and stimulus-response compatibility in absolute identification. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 51*(2), 165-170. doi:10.1037/1196-1961.51.2.165

Lacouture, Y., Li, S., & Marley, 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. doi:10.1080/00049539808258793

Lacouture, Y., & Marley, A. J. (1995). A mapping model of bow effects in absolute identification. *Journal of Mathematical Psychology, 39*(4), 383-395. doi:10.1006/jmps.1995.1036

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

Levine, G. (1960). Stimulus-response generalization with discrete response choices. *Journal of Experimental Psychology*, 60(1), 23-29. doi:10.1037/h0041055

Lewandowsky, S., Kalish, M., & Ngang, S. K. (2002). Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General, 131*(2), 163-193. doi:10.1037/0096-3445.131.2.163

Lindahl, M. B. (1964). The importance of strategy in a complex learning task. *Scandinavian Journal of Psychology, 5*(3), 171-180. doi:10.1111/j.1467-9450.1964.tb01424.x

Lindahl, M. B. (1968). On transitions from perceptual to conceptual learning. *Scandinavian Journal of Psychology, 9*(3), 206-214. doi:10.1111/j.1467-9450.1968.tb00535.x

Lockhead, G. R. (1984). Sequential predictors of choice in psychophysical tasks. In S. Kornblum & J. Requin (Eds.), *Preparatory states & processes* (pp. 27-47). Hillsdale, NJ: Lawrence Erlbaum Inc.

Lockhead, G. R. (2004). Absolute Judgments Are Relative: A Reinterpretation of Some Psychophysical Ideas. *Review of General Psychology*, 8(4), 265-272. doi:10.1037/1089-2680.8.4.265

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. doi:10.1080/17470210802649285

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review, 12*(1), 24-42. Retrieved from EBSCOhost.

McDaniel, M. A., Dimperio, E., Griego, J. A., & Busemeyer, J. R. (2009). Predicting transfer performance: A comparison of competing function learning models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(1), 173-195. doi:10.1037/a0013982

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7*(4), 241-253. doi:10.1037/0278-7393.7.4.241

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97. doi:10.1037/h0043158

Mori, S., & Ward, L. M. (1995). Pure feedback effects in absolute identification. *Perception & Psychophysics, 57*(7), 1065-1079. Retrieved from EBSCOhost.

Morris, C., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior, 16*(5), 519-533. doi:10.1016/S0022-5371(77)80016-9

Murdock, B. R. (1960). The distinctiveness of stimuli. *Psychological Review, 67*(1), 16-31. doi:10.1037/h0042382

Musiélak, C., Chasseigne, G., & Mullet, E. (2006). The learning of linear and nonlinear functions in younger and older adults. *Experimental Aging Research, 32*(3), 317-339. doi:10.1080/03610730600699126

Neath, I., Brown, G. A., McCormack, T., Chater, N., & Freeman, R. (2006). Distinctiveness models of memory and absolute identification: Evidence for local,

not global, effects. *The Quarterly Journal of Experimental Psychology, 59*(1), 121-135. doi:10.1080/17470210500162086

Nosofsky, R. M. (1983). Information integration and the identification of stimulus noise and criterial noise in absolute judgment. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(2), 299-309. doi:10.1037/0096-1523.9.2.299

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104-114. doi:10.1037/0278-7393.10.1.104

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 211-233. doi:10.1037/0278-7393.18.2.211

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. doi:10.1037/0033-295X.112.2.383

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3, Pt.1), 353-363. doi:10.1037/h0025953

Poulton, E. (1979). Models for biases in judging sensory magnitude. *Psychological Bulletin, 86*(4), 777-803. doi:10.1037/0033-2909.86.4.777

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. Retrieved from EBSCOhost.

Siegler, R. S., & Opfer, J. E., (2003). The development of numerical estimation: evidence for multiple representations of numerical quantity. *Psychological Science, 14*, 237-243.

Shiffrin, R. M., & Nosofsky, R. M. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review, 101*(2), 357-361. doi:10.1037/0033-295X.101.2.357

Siegel, J. A., & Siegel, W. (1972). Absolute judgment and paired-associate learning: Kissing cousins or identical twins?. *Psychological Review, 79*(4), 300-316. doi:10.1037/h0032945

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition, 65*(2-3), 167-196. doi:10.1016/S0010-0277(97)00043-7

Sniezek, J. A., & Naylor, J. C. (1978). Cue measurement scale and functional hypothesis testing in cue probability learning. *Organizational Behavior & Human Performance, 22*(3), 366-374. doi:10.1016/0030-5073(78)90022-3

Stewart, N., Brown, G. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112*(4), 881-911. doi:10.1037/0033-295X.112.4.881

Stewart, N., & Matthews, W. J. (2009). Relative judgment and knowledge of the category structure. *Psychonomic Bulletin & Review, 16*(3), 594-599. doi:10.3758/PBR.16.3.594

Wagenaar, W. A., & Sagaria, S. D. (1975). Misperception of exponential growth. *Perception & Psychophysics*, 18(6), 416-422.

Ward, L. M., & Lockhead, G. R. (1970). Sequential effects and memory in category judgments. *Journal of Experimental Psychology, 84*(1), 27-34. doi:10.1037/h0028949

West, R. L., Ward, L. M., & Khosla, R. (2000). Constrained scaling: The effect of learned psychophysical scales on idiosyncratic response bias. *Perception & Psychophysics, 62*(1), 137-151. Retrieved from EBSCOhost.