The cost of errors: confusion analysis and the mental representation of familiar and unfamiliar digits

Paul M. Garrett[a], Murray Bennett[a], Zachary L. Howard[a], Cheng-Ta Yang[b], Daniel R. Little[c], and Ami Eidels[a]

[a] School of Psychology, University of Newcastle, Australia

[b] Department of Psychology, National Cheng Kung University, Taiwan

[c] School of Psychology, Melbourne University, Australia

Author Note

This work was supported by an ARC Discovery Project Grant DP160102360 to Ami Eidels and Daniel Little, and the John and Daphne Keats endowment research fund G1700468 to Zachary Howard, Paul Garrett, Cheng-Ta Yang, Daniel Little and Ami Eidels. Correspondence concerning this article should be addressed to Paul M. Garrett, School of Psychology, The University of Newcastle, Newcastle, New South Wales 2308, Australia. E-mail: paul.garrett@uon.edu.au.

Abstract

People express quantities using a remarkably small set of units – digits. Confusing digits could be costly, and not all confusions are equal; confusing a price tag of 2 dollars with 9 dollars is naturally more costly than confusing 2 with 3. Confusion patterns are intimately related to the distances between mental representations, which are hypothetical internal symbols said to stand for, or represent, 'real' external stimuli. The distance between the mental representations of two digits could be determined by their numerical distance. Alternatively, it could be driven by visual similarity. In an English speaking cohort, we investigated the mental representations of familiar and unfamiliar numbers (4 sets: Arabic, Chinese, Thai, and non-symbolic dots) through a set of identification experiments, using multi-dimensional scaling and cluster analysis. We controlled for undesired effects of response bias using Luce's choice model. Our findings show Arabic, Chinese and Thai numerals were represented in the mental space by perceptual similarities. We also find non-symbolic dots were represented by perceptual and numerical similarities. This work is a novel contribution to the literature and lays the foundation for further investigations into the mental representation of numerals across cultures.

The cost of errors: confusion analysis and the mental representation of familiar and unfamiliar digits

People express quantity through a remarkably small set of digits. These digits, 0–9, and the quantities they represent are fundamental to our understanding of finance, mathematics, programming, and time. The cost of confusing one digit for another may be minor, for example confusing $2 as $3, or major, for example confusing $2M as $3M. Many digits share similar visual features increasing the likelihood of a confusion. Although the visual properties of symbolic digits change between languages, for example, ‘2’, ‘⁚’ and ‘二’, their numerical value does not.

Digits maintain the numerical properties of cardinality, (i.e., unit-value), and ordinality, (i.e., unique sequential ordering). With use, these properties become embedded into our internal representation of number; our so called *mental number space*. But which plays a larger role in the confusion of digits: our visual perception or our internal representation of quantity?

The current study investigates the effect of perceptual and numerical similarities on the confusion of digits within an English speaking cohort. We analyzed confusion patterns in a digit-identification task (via confusion matrices) and assessed how perceptual and numerical properties influence the mental representation of familiar and unfamiliar digits. We also consider the mental representation of symbolic quantities, such as dice patterns. To foreshadow, we find evidence that confusions between digits depend primarily on perceptual similarities, and confusions between quantities depends on both perceptual and numerical similarities.

The value in digits

The approximate number system (ANS; Dehaene, 2011; Gallistel & Gelman, 1992) is the predominant account for how numbers are represented by humans (Dehaene, 2011), as well as many other species (Woodruff & Premack, 1981; Pepperberg & Gordon, 2005; Agrillo, Dadda, Serena, & Bisazza, 2008). The ANS detects differences in quantity through changes in relative magnitude (Gallistel & Gelman, 1992). Over time, human cultures have mapped these discrete differences onto non-symbolic representations of quantity, such as tallies and dots.

For small quantities, non-symbolic representations are useful (e.g., a tally of five apples), however, these representations become error prone at larger counts (e.g., a tally of 5,000 apples). Absolute symbolic representations, such as Arabic numerals, remove this problem (Menninger, 2013) by mapping differences in quantity to unique symbols — digits. This unique mapping between quantities and digits enforces the sequential ordering and cardinality of each unit. With use, digits become inherent of the quantities they represent, eventually assuming a role in how we represent quantities within our mental number-space.

The mental number-space

Mental representations, such as our mental number-space, are theoretical cognitive states thought to reflect the external world (Mueller & Weidemann, 2012; Eidels & Cassey, 2016). Digits are representations of quantity and inhabit the same mental number-space as the approximate number system. Numerical confusions between two digits, for example confusing 3 and 4, may index numerical proximity within this mental space.

Digits are prone to effects of numerical magnitude, specifically the *size-*, *distance-* and *ratio-effects* (Dehaene, 2011). The size-effect describes how, given the same difference, larger numbers such as 8 and 9 are harder to compare (less accurate and slower) than smaller numbers, such as 3 vs 4. The distance-effect describes how closer numbers (e.g., 4 vs 5) are harder to compare than numbers further apart (e.g., 4 vs 9). Finally, the ratio-effect describes how smaller ratios (e.g., 4 vs 6) are harder to compare than larger ratios (e.g., 2 vs 4). These effects index proximity within the mental number-space and reflect an ordering to our mental representations of number.

The size-, distance- and ratio-effects show that digits are i) represented within the mental number-space, and ii) subject to numerical ordering and proximity. As such, confusions between two-digits may be caused by their numerical proximity and numerical ordering within the mental number-space. However, digits do not always represent their numerical value. For example, passwords that contain digits, such as 'PA55WORD', may be read without numerical influence. As '5' could be read as either a five or 'S', the visual properties of the digit must be processed before its semantic meaning. As such, digit confusions may not be due to numerical proximity, but rather perceptual similarities.

The perception of digits

When we view a series of digits, for example 1, 2, 3..., we apply visual attention to guide and focus our search. By doing so, we enhance the 'signal' of a digit to more precisely determine its visual features (Wolfe & Horowitz, 2004). However, the context in which we view a digit does not always afford correct feature identification. Constraints to time, noise around or over the digit, brightness and contrast, all impact our decision and may result in an incorrect identification or confusion. With digits being integral to our daily lives, (e.g., maths, money, time, cooking), understanding what makes two digits (in)distinguishable has been a topic of deliberate research.

The visual features we use to represent digits were recently explored by Godwin, Hout, and Menneer (2014) using the spatial arrangement method (see Goldstone, 1994). Godwin et al. asked participants to spatially arrange the Arabic numerals 0–9 based on feature-similarity. Through multidimensional scaling (MDS), a method used to visually represent the dimensional properties of confusion patterns, they found digits were identified along two dimensions: i) 'roundness' and 'straightness', for example, 6 vs 9 are more similar than 6 vs 2, and ii) 'openness', for example, 2 vs 7 are more similar than 2 vs 1.

In the same experiment, Godwin et al. (2014) asked participants to complete a search task looking for a target digit among distractor digits. Eye-tracking analysis found an effect of perceptual similarity and numerical proximity. Digits that were perceptually similar to the target were fixated on for longer. Likewise, digits that were numerically close to the target were fixated on for longer than digits further apart.

However, this numerical effect was an order of magnitude less than that found for visual similarities. This finding echos similar results previously established in a related literature — the comparison of letters.

In his investigation of letter similarity, Townsend (1971) collected confusion data on all 26 upper-case letters of the English alphabet (see also Eidels & Cassey, 2016). Sixty-five trials were collected for each letter over 13 sessions, and modelled at the group and individual level. Townsend (1971) found 50% of letter confusions could be attributed to perceptual similarity, with remaining confusions accounted for by noise and alphabetic proximity. This study highlights an instance where perceptual similarity dominated semantic proximity within the mental space. Yet, this study did not determine which visual features are used in letter identification.

Fiset et al. (2008) applied the 'bubbles technique' to partially obscure letters and determine the key visual features used in letter identification. Fiset et al. found that line terminations were the most important feature for letter identification. As an example, 'J' and 'L' or '1' and '7' have similar line terminations, whereas 'U' and 'K' or '1' and '8' do not. This technique, while powerful, required participants to complete 26,000 trials and did not address confusions based on semantic proximity.

These investigations into the perception of digits and letters highlight three important visual properties of an item: i) straightness or roundness, ii) openness, and iii) line terminations. These studies also provide instances where perceptual similarity was more important than semantic similarity. Finally, these studies showcase multidimensional scaling as a method for simultaneously assessing the influence of perceptual and semantic similarity within a character-set, (e.g., an alphabet or a set of digits). In the next section, we discuss the advantages and limitations of MDS, and some advancements in the application of this technique.

Multidimensional scaling

Scaling methods have a long history in the social sciences (see Hefner, 1959; Gower, 1966) with Shepard, (1962a, 1962b) and Kruskal (1964b, 1964a) pioneering modern multidimensional scaling methods (see Groenen & Borg, 2014, for a review). A benefit of MDS is that it takes proximity data, such as similarity ratings or identity confusions, and plots these as distances among two-or-three dimensions of space. The relative distance between points in the MDS space is assumed to reflect the psychological distance or proximity between the stimuli.

As an example, if a participant perceived '1' and '7' as psychologically similar, these items would be clustered within the MDS space. Accordingly, individuals with similar psychological spaces would display similar clustering and MDS spaces, while dissimilar mental spaces would display unique MDS spaces. Traditionally, differences in the MDS space for alphabet and letter studies have been attributed to three causes: i) visibility of the stimuli, ii) similarity of the stimuli, and iii) response bias (Mueller & Weidemann, 2012). As such, MDS studies typically manipulate the visibility of the target stimulus in order to affect the rate of stimulus confusions.

In a classical MDS experiment investigating number or letter confusions, a target stimulus is presented, followed by a set of response-options — a correct item among distractor foils — from which the target must be identified. To increase the rate of confusions, the visibility of the initial target stimulus is degraded through backwards-masking, stimulus noise or item-feature obstruction (e.g., Fiset et al., 2008).

In an alternative to the MDS design, Goldstone (1994) asked participants to spatially arrange stimuli by similarity. This method requires fewer trials than classical MDS and purportedly assesses the underlying psychological map between distance and similarity. However, the design also encourages the visual comparison of all response-options, emphasizing visual similarity and possibly confounding latent cognitive dimensions, such as numerical proximity. Goldstone's method ensured all items were arranged and responded-to simultaneously, bypassing another key issue in classical MDS — response-bias.

Response bias vs feature similarity

Visual similarities between items and shared visual features, such as a straight horizontal line on top, common to both '5' and '7' , increase the likelihood of inter-item confusions. Separating perceptual item-confusion, (e.g., '5' and '7'), from a participant's bias in responding, (e.g., always responding with '5'), has been a central concern of the MDS literature.

In their investigation of letter confusions, Gilmore, Hersh, Caramazza, and Griffin (1979) noted a participant who favored responding with letters 'I', 'J', and 'L'. This resulted in high accuracy for these items, but also higher rates of confusions *with* these items. Factoring out the effect of response-bias on accuracy is a difficult task, yet, necessary to truly assess which visual features are used in symbolic identification. Fortunately, there are modeling techniques designed to tease these elements apart.

Luce's (1963) similarity choice model predicts how response bias, inter-item similarity, and the number of response-options, impact the choices we make. This machinery factors out the effect of individual response bias from the confusion data. Used in conjunction with MDS, this method provides a bias-free representation of the underlying psychological space (the details of this model are covered in Appendix A).

Combining Luce's (1963) choice model with MDS has proven effective in previous similarity studies. Townsend (1971) combined these techniques in his assessment of alphabetic similarities, with the bias-free MDS results providing the best explanation of the data — a finding recently replicated by Pleskac (2015). We extended this methodological approach to the study of digit confusions. Rather than be limited to a single set of digits, the current study examined four unique digit sets and explored the effect of familiarity on the resulting MDS space.

The language of numbers

Arabic digits, 0–9, are abundant in predominantly English-speaking countries. Familiar number sets may be internally represented by numerical proximity. As a comparison set, we also presented stimuli in the form of non-symbolic dots.

Non-symbolic dots, as found on domino tiles, playing cards, and dice, are a direct representation of numerical magnitude (hence, non-symbolic). Each increment in the quantity of dots presented coincides with an exact increase in numerical magnitude. Symbolic numerals present in consistent shapes and orientations. Accordingly, we presented non-symbolic numerals in consistent and familiar patterns. These dot patterns provide a comparison set of familiar, yet non-symbolic stimuli.

Deciding on the specific spatial arrangement of the dots in the stimulus display is complicated by two issues. First, increasing the number of dots in a display gives rise to new emergent features (Pomerantz & Portillo, 2011; Hawkins, Houpt, Eidels, & Townsend, 2016). For example, moving from a one dot-display to two dots adds not only an additional dot, but also new information concerning the relative position and distance between the two dots. Moving from two dots to three again results in a new emergent feature, co-linearity (whether all three dots are on the same line or not), and so on. Thus, unless carefully controlled, emergent features and numerosity are easily confounded. Second, numerosity and visual properties such as contrast energy or total area are also confounded; unless carefully controlled, displays with more dots would have larger stimulus area, or higher density. A complete investigation would therefore require many conditions, each designed to control for specific factors. Since this was not the goal of the present study, we selected a single set of dot displays, guided by familiar dot patterns based on playing cards (and slightly modified for the 8 and 9 dot displays to fit within a 3 x 3 grid). In addition to the familiar Arabic and dot sets, we presented two sets of unfamiliar numeric symbols, Chinese and Thai.

In their investigation of symbolic similarity, Yeh, Li, Takeuchi, Sun, and Liu (2003) presented participants with Chinese characters and asked the participants to arrange the characters by similarity. The spatial arrangement of these characters was thought to represent similarity in the mental space. Taiwanese and Japanese students who were familiar with Chinese characters, arranged items by configurable structures and treated characters as whole objects. By contrast, English and illiterate-Taiwanese students arranged items by feature components, focusing upon individual lines strokes and orientations within each character. The reported difference between these cohorts and the way they perceived similarity, was their level of expertise with the Chinese character set.

In the current study, we followed up on these findings and presented two unfamiliar numeric item-sets, Chinese and Thai. Chinese numerals are logographic (each character represents a word or phrase, Hung, Tzeng, & Tzeng, 1992; Shakkour, 2014) and the visual properties of Chinese numerals represent their numerical value. For example, the Chinese characters 一, 二, 三, 四 are the numerals 1–4 as indicated by the sum of their outer lines. Thai numerals, like Arabic digits, are non-logographic and impart no numeric value in their physical characteristics. Chinese and Thai numeric sets will provide insight into whether numerical proximity may be imparted by an unfamiliar logographic numeric set, as compared to an unfamiliar, non-logographic numeric set.

In an English speaking cohort, we combined a method for removing response bias using Luce's choice model with multidimensional scaling to assess the mental representation of digits from four different numeric-types: Arabic, Chinese, Thai and non-symbolic dots. We hypothesized familiar Arabic digits would be confused by dimensions of perceptual similarity and numerical proximity. We hypothesized non-symbolic dots would be confused by dimensions of numerical proximity. Finally, we hypothesized unfamiliar Chinese and Thai digits would be confused by dimensions of perceptual similarity.

Method

Participants

Participants were 11 student volunteers (4 females) from the University of Newcastle, Australia, who completed four 90 minute experimental sessions (one per numeric-type) and were reimbursed \$25 per half-hour. The average age was 24.45 years (SD = 1.53 years). All participants reported as having normal or corrected to normal vision, were proficient in English, and could not speak or read Chinese nor Thai.

Stimuli and apparatus

Stimuli were presented on a 23inch Dell s2240L (60Hz) monitor with a 16:9 aspect ratio at a display resolution set to 1920 x 1080. Arabic digits (from here on, numerals) were generated using calibri-body 80pt font[1], Chinese numerals were generated using DFKai-SB 80pt Font, Thai numerals were generated using unicode characters generated with Calibri 80pt font and non-symbolic dot patterns were generated as canonical representations of quantity within a 3 x 3 grid (see Figure 1). The canonical dots patterns were based on playing cards (8 and 9 dot patterns were slightly altered to fit within the 3 x 3 grid). The number '0' was avoided due to similarities across numeric-types. The experiment was coded and presented using Python 2.7.14 and the Pygame 1.9.2b package. Responses were recorded using a standard dell 9RRC7 optical mouse on a Windows 7 operating system with mouse sensitivity settings set to a default value of 10.

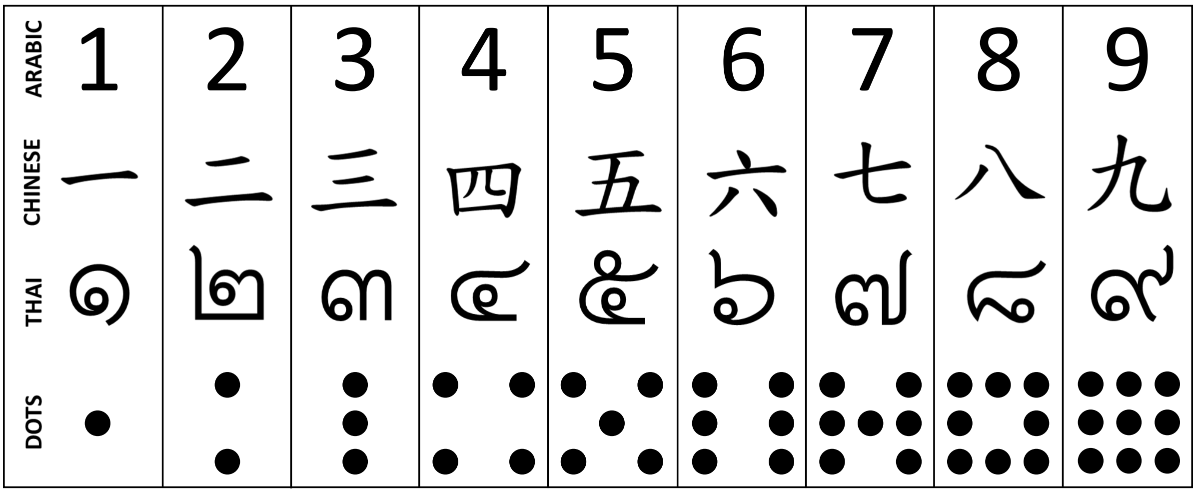

Figure 1. Arabic, Chinese, Thai and dot numerals for the range of one to nine (left to right).

Target stimuli were displayed in the center of the screen within a noisy field ($\mu = 0$, $\sigma = 0.125$; à la Eidels & Gold, 2014) and were followed by a central backwards-mask. At a viewing distance of 60cm, each noisy target stimuli subtended a visual angle of 5.53 degrees (5.8cm^2) and the mask subtended a visual angle of 11.61 degrees

[1] All stimuli were generated in Microsoft Powerpoint 2017 and saved as images that were displayed during the experiment.

($12.2cm^2$). Responses were made by moving the mouse to the matching numeral presented within a response-wheel (see Figure 2). The response-wheel was evenly divided into nine sectors, each containing one of nine numeric-symbols. The symbols were sampled from one of the four numeric-conditions (Arabic, Chinese, Thai and Dots; see Figure 1 again). Each numeral was randomly allocated to a wheel sector at the start of each session and displayed equidistant from the starting mouse location. This design ensured no one numeral was spatially biased towards the target.

At the start of each response-window, a mouse-cursor appeared in the center of the number-wheel. Participants responded by moving the mouse-cursor towards the segment that contained the best match to the previously presented stimulus. A response was taken once the cursor passed over the inner-circle of the response-wheel. At a viewing distance of 60cm, the inner-circle of the response-wheel subtended 20.04 degrees visual angle (diameter 21.2cm) and the outer-circle subtended 25.91 degrees visual angle (diameter 27.6cm). All experimental displays were presented on a gray background with RGB values (240, 240, 240).

Procedure

Participants completed four 90min sessions — Arabic, Chinese, Thai and non-symbolic dots. Session order was randomized using a Latin-square design. At the start of the first session participants were presented an information statement and provided signed consent before answering demographic questions regarding age, gender, handedness and vision. Participants reported whether they identified as being proficient in English, Chinese and Thai. At the start of each session participants were instructed to briefly view a noisy symbol, and using the mouse, identify the best matching symbol on the response-wheel.

Each trial began with a blank screen presented for 250ms, followed by a central fixation-cross for 500ms, followed by a 250ms blank screen. The stimulus display was then presented for 500ms, followed by a mask for 200ms. The response-window began at the presentation of the response-wheel and lasted 8000ms (see Figure 2). A trial ended when a response was made or when the trial timed-out.

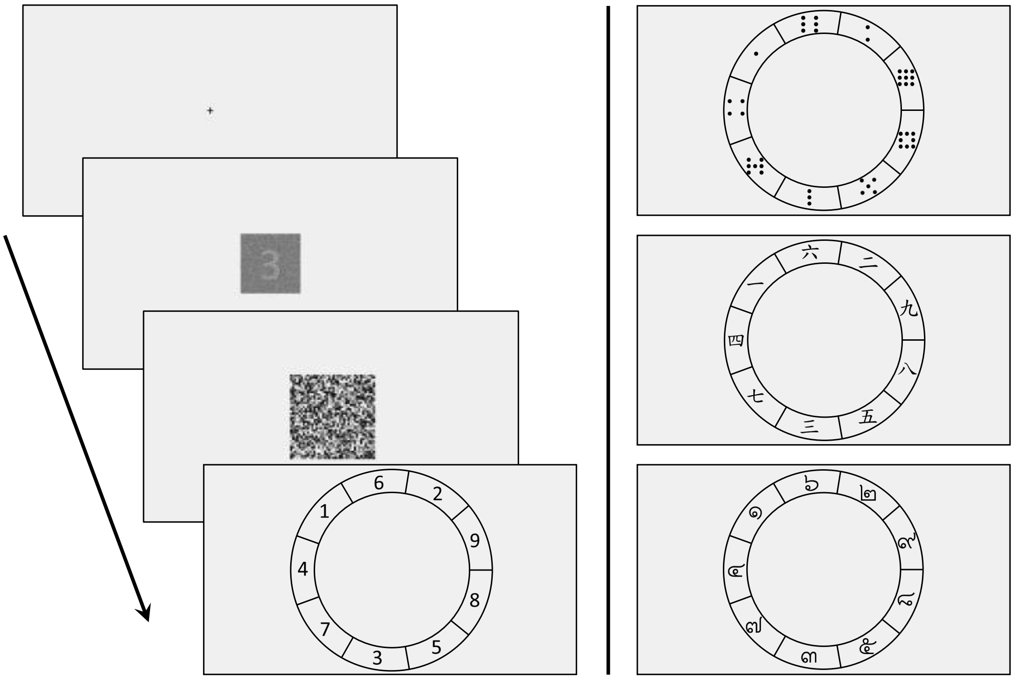

Figure 2. Illustration of a trial with Arabic numerals (left). Alternative response-wheels are displayed (right) for the non-symbolic dots (top), Chinese (middle) and Thai (bottom) numerals. For illustrative purposes, the position of each numeral is held constant between numeric-types.

At the start of each session, participants completed a practice-calibration block. By means of a single staircase procedure, we manipulated the contrast of the signal across trials according to the participant's responses. A modified 2-up 1-down rule was applied[2]. Signal contrast began at a fixed RGB value (153, 153, 153), with two contiguous correct responses decreasing the RGB signal-values by 1 (becoming harder), and a single incorrect response increasing the RGB signal-values by 1 (becoming easier). This staircase design allowed participants to quickly plateau at their perceptual threshold. Earlier piloting with this procedure resulted in approximately 60% accuracy in the main task.

Our proffered analysis technique, multidimensional scaling, requires a combination of correct and error responses. To ensure this, experimental stimuli were presented at

[2] A classical 2-up 1-down staircase requires two correct responses on each step to step-up. Our modified staircase required a correct response on two contiguous trials, (e.g., trial n-1 and trial-n), to step-up. This produced a faster and more responsive staircase procedure.

five signal levels of varying difficulty (following Eidels & Gold, 2014). A critical contrast level was determined by the mean RGB values of the final 30 calibration trials. The highest signal level (easiest) was presented three RGB steps above the critical contrast. The lowest signal level (hardest) was presented one RGB step below the critical contrast. Together, these formed the five stimulus-signal contrast-levels.

During each session, participants completed one practice block of 135 trials, and 13 experimental blocks each containing 90 trials. During an experimental block, each numeral was presented 10 times, twice at each of the five signal levels. Trial-by-trial accuracy feedback was provided during the practice-calibration block and trial order was randomized within each block. Block accuracy was displayed as a graph at the end of each experimental block to encourage participant engagement[3]. In total, each participant completed 130 experimental trials per numeric-symbol, and 1170 experimental trials per-session.

Data Analysis

Trials with no response were removed from the analysis. Calibration (practice) and experimental trials were assessed for accuracy to ensure an appropriately difficult stimulus-signal contrast was achieved. Repeated-measures ANOVAs and paired-sample *t*-tests were used to statistically compare differences between numeric sets. Where accuracy was matched by signal-contrast level, between-subject ANOVAs and independent-samples *t*-tests were used. Multiple comparisons were corrected for family-wise error using the bonferroni method.

For each participant, a 9 x 9 confusion matrix was generated for each numeric-type. To remove the effect of response-bias before MDS analysis, Luce's (1963) similarity choice model was applied to each confusion matrix. This model describes identification responses as probabilistic outcomes driven by the similarity of a stimulus to other in the choice set, as well as a response-bias parameter — one for each stimulus.

[3] At no point during the experiment were any numbers displayed except for those contained by the response-wheel and target-stimulus. Accuracy was presented as a line graph with no numbers, and countdown timer was displayed as a ticking sundial.

By estimating the parameters of the model, researchers can examine the theoretically meaningful similarity scores free from the effect of response-bias that can contaminate the observed data. In Appendix A we provide a formal description of Luce's choice model and describe the application to the current data.

After application of Luce's choice model, non-metric multidimensional scaling was conducted on the bias-free similarity matrices. For each participant and each numeric-condition, a scree analysis was conducted to determine the appropriate number of MDS dimensions. Group MDS plots were generated for each numeric-type using the individual differences scaling (indscal) MDS technique (Carroll & Chang, 1970). Indscal provides a group MDS fit by deferentially weighting the contribution of each individual to the overall MDS fit. K-means cluster analysis was applied to determine cluster patterns at both the group level and across individuals. The frequency at which items clustered were turned into proportions and displayed as a heatmap, separately for each numeric-type. Finally, we compared MDS and K-mean cluster results to an ideal observer analysis, used to simulate pure perceptual confusions of the numeric sets.

Results

Calibration Block

Figure 3 (top) depicts the calibration block for participant S1, responding to Arabic numerals. This staircase procedure was typical of all participants. The mean contrast level of the final 30 assessment trials (highlighted yellow) determined the critical contrast value — the value from which stimulus-signal levels were determined in the experimental session. Violin plots (bottom) depict the mean and standard-error of contrast values for assessment trials, for each participant and numeric-type. Critical contrast levels were relatively stable across numeric-type and participants.

Colored ticks on the violin-plot (Figure 3, y-axis) show, on average, critical contrast levels were lowest for Arabic numerals (RGB $\mu = 134.14$, $\sigma = 1.07$), then non-symbolic dots (RGB $\mu = 134.88$, $\sigma = 1.04$), then Chinese numerals (RGB $\mu = 134.91$, $\sigma = .86$) and finally, Thai numerals (RGB $\mu = 135.5$, $\sigma = .97$). A lower

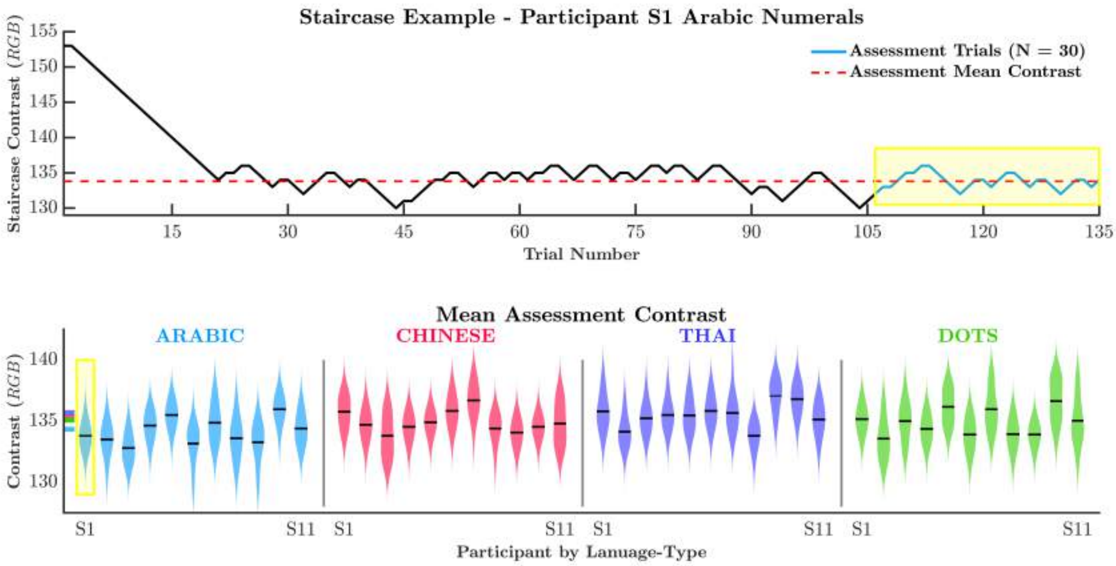

Figure 3. Plot of participant S1's Arabic numeral staircase procedure (top) and violin plots of individual participant's staircase assessment trials (bottom). Assessment trials (highlighted yellow for participant S1) determined the critical contrast value for the main experiment. For all numeric-types, participants displayed relatively stable contrast levels during the assessment window. Black lines on each violin plot represents the critical contrast value (mean RGB value over the assessment window). Colored ticks on the y-axis are the mean critical contrast values for each numeric-type.

signal-level suggests familiar numeric sets (Arabic and Dots) were easier to recognize than unfamiliar numeric sets (Chinese and Thai). However, the greatest difference between contrast levels, Arabic vs Thai, was only equivalent to a single RGB step.

Experimental accuracy

The staircase procedure was effective at reducing identification accuracy during experimental trials. On average, accuracy was highest for Chinese numerals (μ = .60, σ = .21), then Arabic (μ = .59, σ = .19) and non-symbolic dots (μ = .59, σ = .19), and finally, Thai numerals (μ = .54, σ = .19). Our manipulation of contrast accuracy was similarly effective. During experimental trials, stimuli were presented at five signal-levels, one step below the critical contrast value (level 1: hardest), at the critical value (level 2) and three steps above (levels 3, 4 and 5; easiest). As intended, mean accuracy increased linearly with the visibility of the contrast levels, being lowest at level 1 (μ = .32, σ = .02) and highest at level 5 (μ = .80, σ = .03). A full analysis of accuracy over contrast levels is provided in Appendix B. For now, we summarise by stating our manipulation of contrast worked as intended and produced error rates sufficient for our subsequent analyses.

Response Bias

Figure 4.a. shows the positive relationship (rank-order correlation $\rho = .83^{***}$) between response-frequency (blue) and response-accuracy (orange) for Arabic numerals in participant S1. Here, as the frequency of responding with a specific numeral, for example '4', increases, so too does identification accuracy. Similarly, as response-frequency decreases, such as with item '3', accuracy also decreases. This figure clearly displays the relationship between response-frequency ('strength' in Luce's choice model) and response-accuracy.

The dotted blue line in Figure 4.a represents a response-frequency matching the number of stimulus presentations. For example, a '5' response was made as often as '5' was presented, however, these responses were correct only half of the time. By contrast, a '4' response was made nearly twice as often as it was presented, showing an effect of response-bias. The positive relationship between response-frequency and response-accuracy is evident when we examine the scatter plot in Figure 4.b. This scatter plot depicts accuracy against response-frequency (bias) for each participant, for each stimulus and numeric-type. The positive relationship ($r = .58^{***}$) indicates that in a standard analysis, overall accuracy and response bias could be mistakenly conflated. This highlights the need for a bias free measure of similarity, as offered by Luce's choice model.

Figure 4.c shows the mean response-frequency and mean response-accuracy of each stimulus, separately for each numeric-type. Averaging response-frequency and accuracy diminishes their correlation, however, clearly illustrates response patterns and accuracy for each stimulus. Together, these results show how an increase in response-frequency (or strength in Luce's choice model) can artificially improve identification accuracy for any given stimulus. Similarly, these results show how a

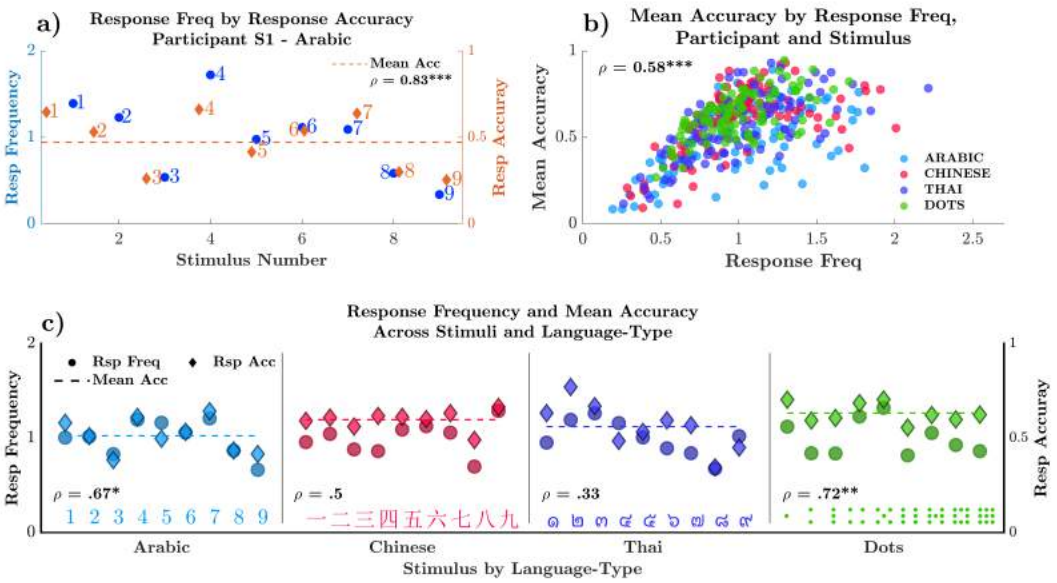

Figure 4. a) Response frequency by response accuracy for participant S1, Arabic numerals. b) Scatter plot depicting a positive correlation between mean stimulus accuracy and response frequency, across numeric-types. c) Response frequency by response accuracy for stimuli for the Arabic (left), Chinese (mid-left), Thai (mid-right) and Dot (right) numeric-types.

decrease in response-frequency can lead to poorer stimulus identification accuracy. To understand the impact response-bias has on confusion data and our analysis of the mental space, we now consider our MDS results.

Multidimensional Scaling

Response bias. Figure 5 shows MDS results from a representative data set (participant S4) where response-bias was unaccounted for (biased plots) and corresponding MDS results where response-bias was removed from the data using Luce's choice model (bias-free plots). Changes between item-proximity within each numeric-type (blue arrows) illustrates how response-bias alters the MDS solution. All future references to MDS within the results section will pertain to the bias-free MDS solutions.

Scree analysis identified two-dimensions as an appropriate MDS representation for most participants in each of the four numeric-types (for more details see supplementary materials, Figure S2.1 and Figure S2.2). Scree analysis identified three-dimensions as the appropriate MDS representation for three participants in the Arabic and Thai

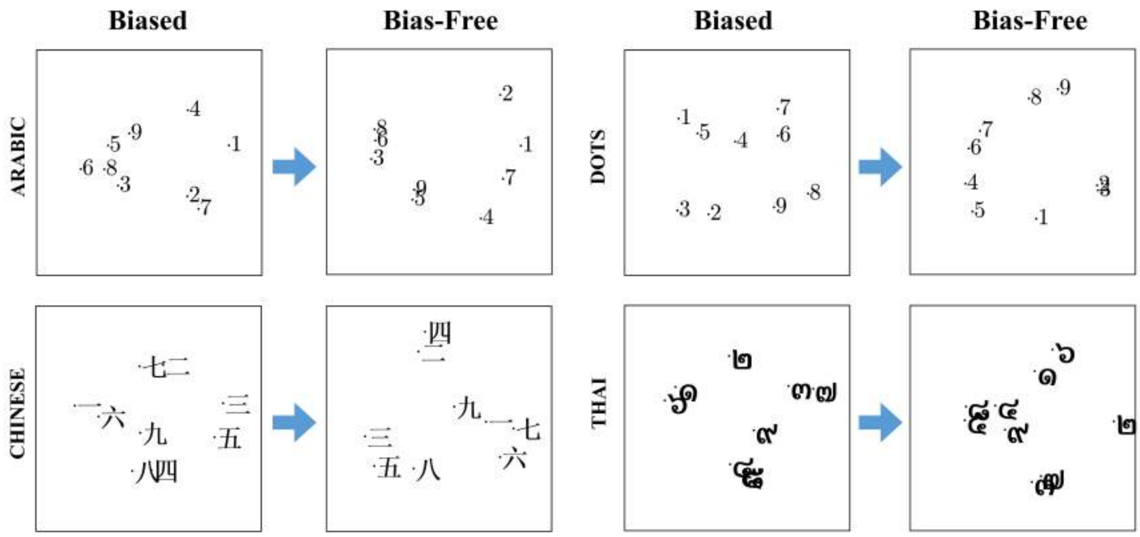

Figure 5. Biased (uncorrected) and bias-free (Luce's choice model corrected) MDS solutions for participant S4, displayed separately for each numeric-type. Changes in item-proximity between biased and bias-free MDS plots displays the influence of response-bias on the MDS solution.
Note. Numerals displayed within the MDS plots are for illustrative purposes and are not identical to the experimental stimuli; see Figure 1. Dots are presented as Arabic numerals to avoid misinterpretation where numerals spatially overlap.

numeric-types, and the appropriate representation for four participants in the Chinese numeric-type. Individual bias-free MDS plots are displayed in the supplementary materials, Figure S2.3 – Figure S2.6; biased MDS plots are displayed in the supplementary materials Figure S2.7 – Figure S2.10). To describe trends observed across participants, we conducted an Individual Differences Scaling (indscal) MDS analysis. Results of the three-dimensional MDS indscal solutions (see supplementary materials, Figure S4.1) closely resemble the results of the two-dimensional MDS solutions. For efficiency of exposition, the following will focus on the majority of participants and the two-dimensional indscal results.

MDS indscal interpretation. Figure 6 displays the group MDS indscal solutions for data collapsed across those participants best represented by two-dimensions, separately for each numeric-type. Interpretation of MDS plots is at times arbitrary and relies on visual inspection of the plots. Similarly, interpretation of the MDS solution's axes are also arbitrary with reference to the x- or y-axes (see e.g., Nosofsky, 1986; Nosofsky, Sanders, Meagher, & Douglas, 2018). Likewise, the notion of similarity is very broad and has been the centre of many disputes in the literature (e.g., Tversky, 1977; Medin, Goldstone, & Gentner, 1993). We begin with a visually-guided assessment of the MDS plots, and then move to more formal analyses of clustering (using K-means cluster analysis) and similarity (using a rudimentary yet useful ideal-observer analysis).

Arabic numerals appear to be arranged along dimensions of roundness (x-axis) and openness (y-axis; similar results were observed by Godwin et al., 2014). Arabic numerals formed four groups in the MDS space: [2,7], [1,4], [5,6] and [3,8,9]. The dimension of openness best described the diagonal of the y-axis[4] for all numerals except the closed shape of item '8'. Under noisy stimulus conditions, the concave exterior of '8' might be perceived as more 'open' than it would under ideal viewing conditions. These results show an apparent effect of perceptual similarity on the mental representations of Arabic numerals.

The non-symbolic dots indscal MDS solution (Figure 6) appears to be displayed across dimensions of alignment (x-axis; whether items are presented internally or externally in the nine-dot array) and quantity (y-axis). Non-symbolic dots show five distinct groupings: [1], [2,3], [4,5], [6,7], [8,9]. These groupings suggest items cluster by numerical proximity. Furthermore, if we ignore the unique case of one-dot, numerical magnitude increases in a clock-wise direction, possibly reflecting the mental number-line. These results show an apparent effect of numerical magnitude and perceptual similarities on the mental representations of non-symbolic dots.

The Chinese indscal MDS solution (Figure 6) appears to be arranged across dimensions of alignment (x-axis) and line terminations (y-axis). Chinese numerals are logographic, a trait captured by the visual property of alignment. As a consequence, small-magnitudes and large-magnitudes are mostly separate within the MDS space. Within this space, the Chinese numerals show three distinct groupings, [一, 二], [三, 五] and [四, 六, 七, 八, 九]. It is unclear whether the largest group might be better classified as two or three sub-groups, for example, [四], [六, 七] and [八, 九]. These

[4] The rotation and direction of items within the MDS solution, relative to the x- and y-axis, is arbitrary. It is only important that these dimensions are orthogonal to one-another.

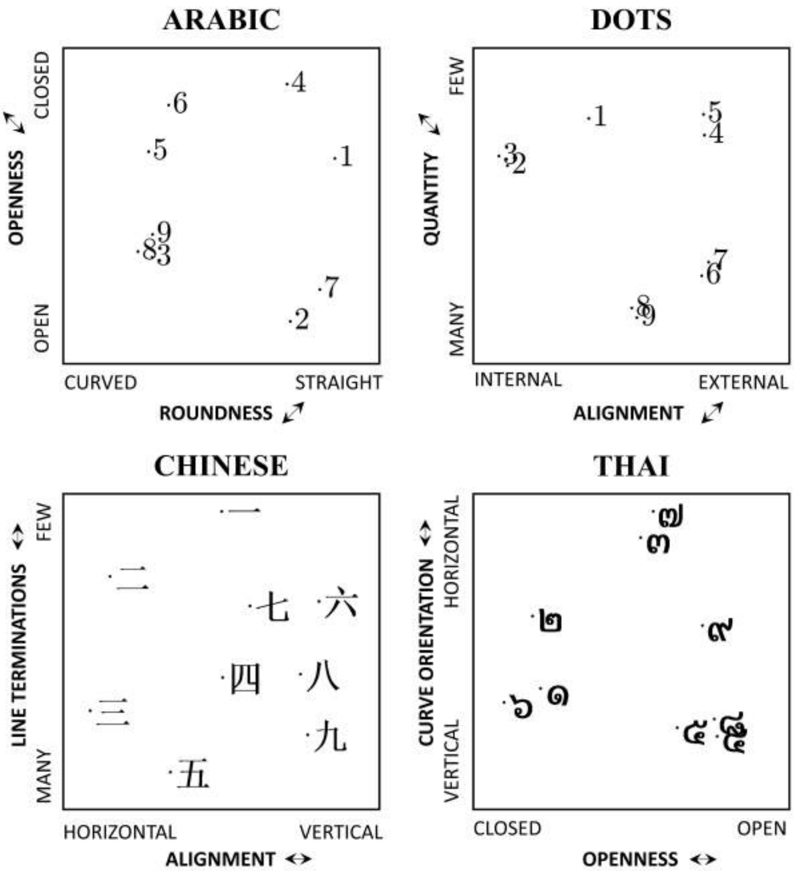

Figure 6. Individual differences scaling (indscal) solution for all participants best represented by a two-dimensional MDS space, displayed separately for each numeric-type. Dimensional labels and directionality (arrows) are displayed on the y-axis and x-axis.

results show an apparent effect of perceptual similarity on the mental representations of Chinese numerals.

The Thai indscal MDS solution (Figure 6) appears to be arranged across dimensions of openness (x-axis) and curve orientation (y-axis; i.e., whether item curvature is horizontally or vertically aligned). Thai numerals show four distinct groupings: numerals with a vertical curve (numerals [3, 7]), numerals with a horizontal curve (numerals [4, 5, 8]), numerals with a closed shape (numerals [1, 2, 6]), and the solo group of item nine. These results show a clear effect of perceptual similarity on the mental representations of Thai numerals.

Based on visual inspection, indscal MDS solutions provided an accurate representation of individual participant MDS results. Arabic, Chinese and Thai numerals were represented within the mental space across dimensions of perceptual similarity. Non-symbolic dots were represented within the mental space using dimensions of numerical and perceptual similarity. Indscal analysis is useful for identifying latent MDS dimensions, however, does not provide a formal measure of item-clustering.

Deciding which items group together and which items are independent is a difficult process. For example, visual inspection of the Arabic indscal solution suggests items ‘1’ and ‘4’ may cluster together or may be independent. Similarly, Chinese numerals [四, 六, 七, 八, 九] may form one group, or three. The ‘strength’ with which two numerals cluster, may determine the likelihood of their confusion within the mental space. To characterize the strength of item-clusters in each individual, and across two- and three-dimensional MDS solutions, we applied to the data a variant of K-means clustering analysis.

MDS clustering

K-means is an iterative clustering technique used to identify item groupings within dense data sets. A number of randomly located centroids (K) are updated iteratively until the data set can be partitioned into ‘K’ non-overlapping clusters. This method works well for large, dense data sets, however, experiences a notable limitation with small data-sets.

Identifying the correct number of centroids is difficult for small data sets. Two, three or four centroids may be adequate for a sample of nine items. However, cluster selection methods developed for large data sets will generally favor higher centroid counts, (e.g., five or six centroids), at a risk to over-fitting the data.

To overcome this limitation, we ran K-mean cluster algorithms using 2–6 centroids, on each bias-free MDS solution. On each iteration of ‘K’, we recorded which items clustered together (Figure 7.a) to produced a measure of cluster frequency (Figure 7.b). For illustration, in the data presented in Figure 7.b, the digits ’1’ and ’2’ were clustered together three times (across the the fine clustering scenarios, K=2, K=3, ... K=6, illustrated in Figure 7.a), whereas the digits ’1’ and ’4’ were clustered together only one time. Of course, each digit is always clustered with itself, resulting in the maximum value of five along the main diagonal.

Within each numeric-type, cluster frequencies were summed across participants and represented by proportion (see Figure 8.a). Separate heatmaps were calculated for two-dimensional and three-dimensional participants (supplementary materials, Figures S3.1–S3.2). Being comparable, these results were collapsed into Figure 8.a. This method was robust to the number of MDS dimensions, as clusters could be calculated in either two- or three-dimensional space. For direct comparison to the previously presented group indscal results, a separate cluster heatmap was generated using the two-dimensional indscal results (Figure 8.b).

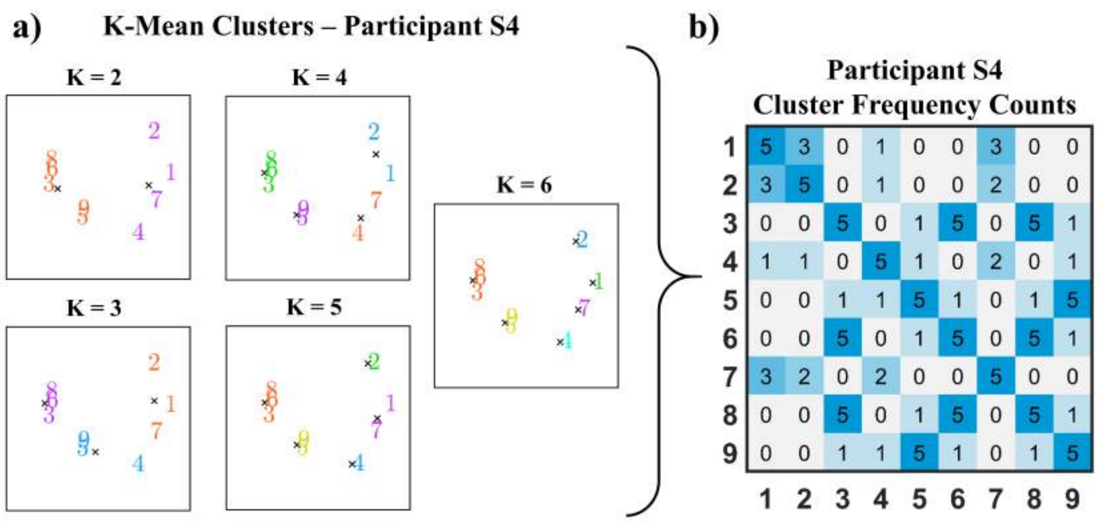

Figure 7. a) K-mean cluster solutions for 2–6 clusters, for a single participant. K-mean cluster centers (centroids) are illustrated by ‘x’ markers, and groupings are denoted by color. b) Cluster frequency heatmap for the same data. Darker colors indicate items which most frequently cluster together.

The top-left of Figure 8.a displays the proportion by which items clustered across individuals, for Arabic numerals. Across individuals, the strength of items clusters generally aligned with the group indscal results (top-left, Figure 8.b). In line with the indscal MDS solution, items with similar perceptual properties, for example, the items [3, 5, 6, 8, 9] share the perceptual property of ‘roundness’, while [2, 3] share the property of ‘openness’; frequently clustered across individuals. Cluster patterns displayed no effect of numerical proximity (neighbouring items clustered infrequently). These results

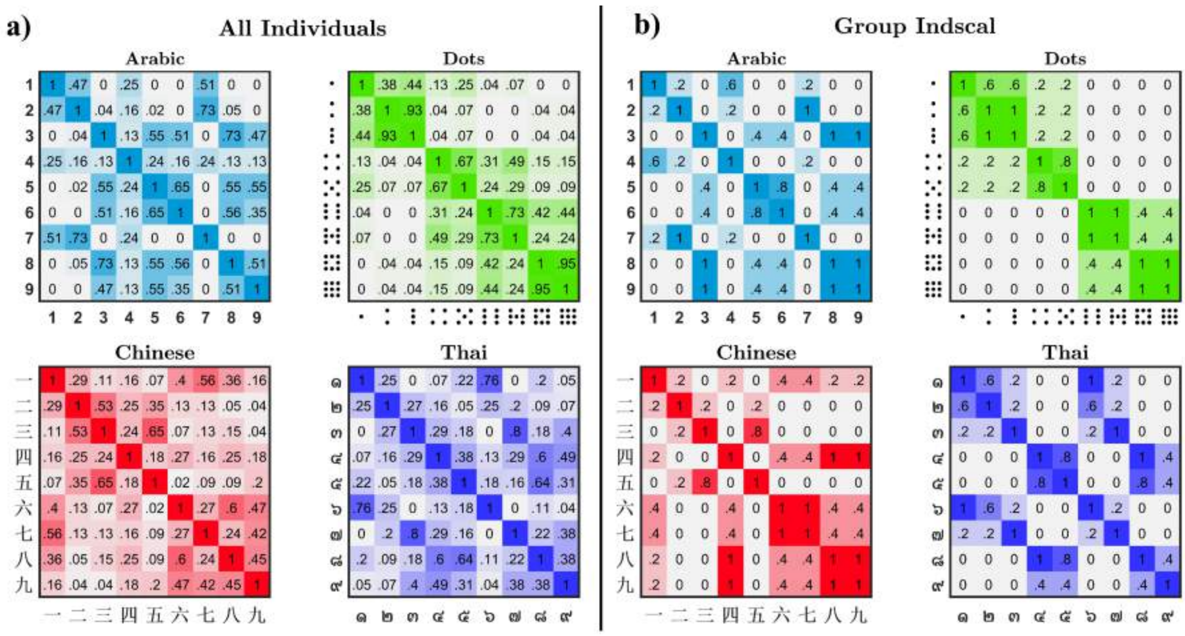

Figure 8. a) Proportional cluster-frequency heatmap for all eleven participants (including both two- and three-dimensional MDS solutions), across 2–6 K-mean clusters. b) Group indscal two-dimensional MDS (Figure 6) cluster-frequency heatmap, across 2–6 K-mean clusters. Larger proportions (darker colored squares) indicate items which most frequently cluster together.

support the indscal analysis, and suggest that at the individual level, Arabic numerals were clustered strongly by the perceptual properties of 'curvature' and 'openness'.

The top-right panel of Figure 8.a displays the proportion of item-clustering for non-symbolic dots. The left-to-right diagonal pattern of results, radiating outwards towards zero in the opposite corners, suggests items clustered by numerical proximity. Yet, items close in numerical proximity cluster together in staggered item-sets. For example, items [2, 3], [4, 5] and [6, 7] cluster, but rarely [3,4], [5,6] or [7,8]. This pattern of results is made clearer by the group indscal plot (Figure 8.b). This staggered pattern of results is not accounted for by numerical proximity, but rather, perceptual similarities.

Staggered clusters, such as 4 and 5 dots, share similar perceptual characteristics, and differ only by the location of a single, central dot. Mental distance between dot representations are likely confounded by both perceptual similarity and numerical proximity; minimal changes such as adding one dot result in minimal changes to both quantity and visual appearance, and likewise adding a large number of dots to a display results in substantial changes to both quantity and visual appearance. As such, it could be that our observed clusters are due to i) only perceptual, ii) perceptual *and* numerical, or iii) only numerical similarities. Results from the indscal MDS solution suggested items were confused along dimensions of quantity (numerical) and alignment (perceptual). As such, it seems likely this staggered pattern reflects a combination of numerical and perceptual similarities.

The bottom-left panel of Figure 8.a displays cluster frequencies for Chinese numerals across individual participants. These results do not align with numerical-proximity, (e.g., non-symbolic dots heatmap), and suggests items clustered by perceptual similarity. Noisy, low-frequency item-clusters are common for Chinese numerals, reflecting the unfamiliar nature of this numeric set with our cohort. Moderate cluster frequencies are present between numerals [二, 三, 四, 五] and [六, 七, 八, 九]. The items within these groups are — unbeknownst to our participants — numerically contiguous. This clustering might reflect the logographic nature of the Chinese numeric-set and the perceptual similarities within smaller and larger magnitudes.

With increases in magnitude, Chinese numerals shift from horizontal to vertical alignment, creating perceptual similarities within smaller and larger magnitudes. Furthermore, smaller magnitudes are generally represented by fewer line-features, (i.e., line-endings), than larger magnitudes. Subsequently, perceptual similarities are strongest within smaller and larger magnitudes. This accounts for the observed indscal MDS results (Figure 8.b) and cluster frequency results. Together, these analyses suggest that at the individual level, Chinese numerals are strongly influenced by the perceptual properties of 'alignment' and 'line-endings'.

The bottom-right panel of Figure 8.a displays cluster frequencies for Thai numerals. Similar to Chinese numerals, Thai cluster patterns do not align with numerical proximity and display an abundance of low-frequency clusters. This may reflect the unfamiliar nature of the numeric-set. Across individuals, and at the group level (Figure 8.b), high cluster frequencies are apparent between item numbers [1,6], [3,7], [4,8], [4,5] and [5,8]. Notably, these items share perceptual features of 'roundness' and 'curvature orientation'. Supplementing these findings with the previous indscal MDS analysis, results suggest that at the individual level, Thai numerals clustered strongly by the perceptual properties of 'curvature orientation' and 'roundness'.

Supplementing indscal analysis with cluster frequency heatmaps, our results indicate perceptual similarity strongly influenced the confusion of Arabic, Chinese and Thai numerals. Furthermore, perceptual similarity also influenced the confusion of non-symbolic dots, but could be confounded with numerical distance in this set, as explained above. Determining the fidelity of these claims is difficult without a benchmark model for comparison. To this end, we now present simulated results from a simple ideal observer analysis.

Ideal Observer Analysis

The ideal observer analysis is a simple template matching process that compares numeric stimuli, pixel-by-pixel, to generate a confusion matrix. The ideal observer is *not* a model of human performance, but rather, a benchmark against which we may compare the performance of human observers (e.g., Gold, Bennett, & Sekuler, 1999; Eidels & Gold, 2014). The 'ideal observer' compares a noisy numeric stimulus to all possible templates, for example comparing a noisy '1' stimulus to the numerals '1–9'. The template with the best cross-correlational match over many iterations, with randomly sampled noise, is selected as the 'ideal observer response'. Normally distributed noise ($\mu = 0$, $\sigma = [1.065, .12, 1.127, 1.463]$ for Arabic, dot, Chinese and Thai numerals, respectively) is added to each numeric stimulus, until the ideal observer's accuracy resembles the average accuracy of the participants. This process was repeated 10,000 times, per numeric-stimulus, per numeric-type, generating four confusion matrices.

To afford a direct comparison to the collected participant data, Luce's choice model was applied to the simulated ideal observer data. Figure 9.a displays the bias-free MDS solutions generated by the ideal observer. Figure 9.b displays the corresponding K-mean cluster frequency heatmaps. These Figures provide a benchmark of performance, given numeric-stimuli were only confused by perceptual similarities.

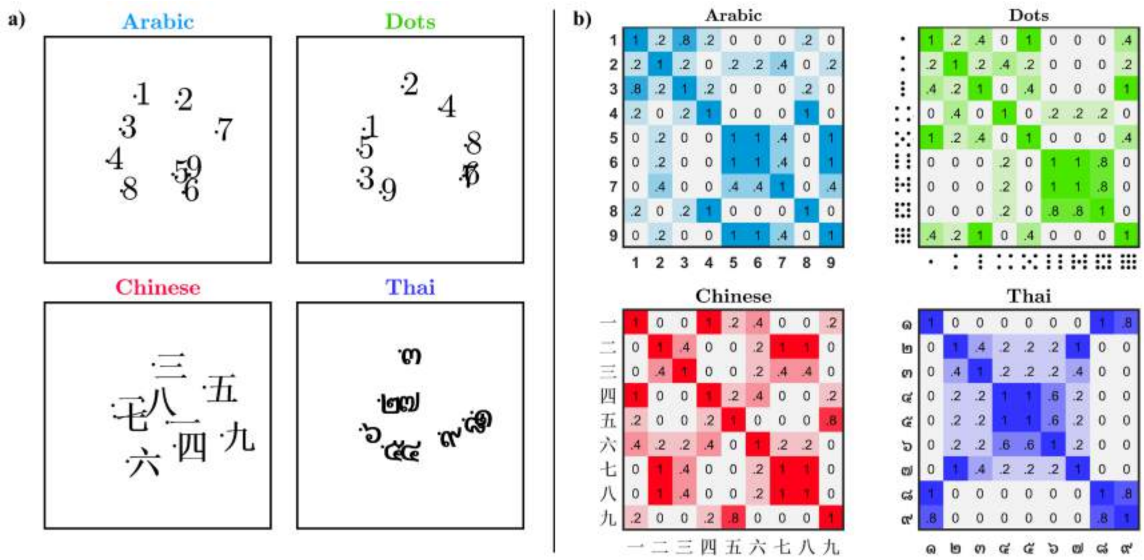

Figure 9. a) Ideal observer analysis bias-free MDS solutions, generated separately for each numeric-type. Non-symbolic dots are displayed as Arabic numerals in the MDS plot for clarity to the reader. b) Ideal observer K-mean cluster frequency heatmaps.

Comparing Arabic indscal MDS results (Figure 6.a) to the Arabic ideal observer MDS results (Figure 9.a), we observe differences between item proximities and co-occurring item groups [2, 7] and [5, 6]. Comparing cluster frequency heatmaps, we find co-occurring item-clusters [1, 2], [2, 7], [5, 6], [5, 9] and [6, 9], suggesting participants confused these items due to perceptual similarities. Other item-clusters did not co-occur even though Arabic numerals appeared to be confused along dimensions of perceptual similarity.

Indscal MDS results for non-symbolic dots differed greatly from the ideal observer, and only shared item groups [1, 5] and [6, 7]. Cluster frequency heatmaps were comparable for items [1, 3], [6, 7] and [6, 8], yet the remaining item clusters were markedly different. Participants appeared to represent non-symbolic dots along dimensions of perceptual *and* numerical similarity. The differences between participant and ideal observer MDS and cluster frequency results may reflect the impact of numerical proximity on the mental space.

The Chinese indscal MDS solution shared similarities with the ideal observer MDS solution. Chinese numerals [一, 二, 七], [三, 五], [四, 九] group in both participant and ideal observer MDS solutions. Cluster frequency heatmaps were similar for items [一,

六], [二, 三], [四, 九], [四, 六], [四, 六] and [六, 七] ; these items all share distinct horizontal line features. While many item clusters were observed in both participant and ideal observer results, the pattern consistent with a logographic numeric-set was not observed by the ideal observer. As with our comparison of Arabic numeral results, it appears the ideal observer is only sensitive to a limited set of perceptual similarities.

The Thai indscal MDS solutions shared similarities with the ideal observer MDS solution. Two MDS groups are distinctly apparent in both solutions, items with a vertical curve (item numbers 4 and 5) and items with a horizontal curve (item numbers 2, 3 and 7). These groups are reflected in the cluster frequency heatmaps (Figure 9). The ideal observer analysis shows very little noise in item-clustering, suggesting that item clusters were determined by highly salient (and comparable) perceptual features.

Discussion

In the current study, participants were asked to identify a noisy symbol (numeral) using a stimulus response wheel. A staircase procedure ensured identification accuracy was approximately 60% for all participants, regardless of numeric-type. Stimulus accuracy positively correlated with response-frequency, across all participants and numeric-types. The application of Luce's choice model negated the effect of response-bias from the multidimensional scaling solutions. MDS and cluster frequency analyses were used to determine the dimensions upon which items were represented in the mental space. Arabic, Chinese and Thai numerals were represented by dimensions of perceptual similarity, and non-symbolic dots were represented by dimensions of numerical proximity and perceptual similarity. MDS and cluster patterns generated by an ideal observer were similar for Chinese and Thai numerals, although, differed greatly for Arabic and non-symbolic dot numerals.

Response Bias

Response-bias had a significant effect on identification accuracy and the multidimensional scaling solutions. Luce's choice model removed response-bias from individual MDS solutions. This altered relative item-proximities and created more even-weightings between items. Indscal analysis collapsed results across bias-free MDS solutions, and allowed the interpretation of bias-free similarity dimensions within the mental space. To the best of our knowledge, this study provides the first ever bias-free representation of the mental space, for familiar and unfamiliar numeric-sets.

Multidimensional Scaling

The majority of participants in Arabic, Chinese and Thai numeric-types, and all participants in the non-symbolic dot numeric-type, were best characterised by two-dimensional MDS solutions. Where participants displayed a third MDS dimension, MDS and K-mean cluster frequency results were comparable, and no category label could be easily applied to the third similarity dimension. As such, the following will focus on the two dimensional MDS results.

Arabic symbols. In line with past findings (Godwin et al., 2014), Arabic numerals appeared to be arranged in MDS space by the perceptual dimensions of 'roundness' and 'openness'. Against predictions, familiarity with the Arabic numerals did not produce numeric-confusions. Participant MDS and K-mean cluster frequency heatmaps displayed limited similarities with the ideal observer analysis.

Items [1, 2], [2, 7], [5, 6], [5, 9] and [6, 9] were similarly represented by both participants and the ideal-observer. Item '6' and '9' are identical once rotated, and '5', '6' and '9' share features of curvature. Items '2' and '7' share similar diagonal midsections, and a horizontal feature. These similarities relate to the 'roundness' of the items and not the openness of their form.

The ideal observer analysis was not sensitive to similarities of 'openness'. Openness relates to the concave 'absence' within an item, and not an extant feature, for example, a straight-line. As openness may be poorly captured by a pixel-by-pixel comparison of similarity, many Arabic numerals that clustered in the participant data were not clustered in the ideal observer analysis.

Non-symbolic Dots. All participants displayed two-dimensional MDS solutions for non-symbolic dots and appeared to be arranged along dimensions of 'quantity' and 'alignment'. MDS plots displayed a rotational ordering, with items progressing from smaller-to-larger magnitudes — a possible representation of the mental number-line. Items [2, 3], [4, 5], [6, 7] and [8, 9] reliably clustered together. This might reflect numerical proximity and the numerical distance effect operating within the mental space. However, the staggered item-clusters may also be caused by perceptual similarities. Sadly, perceptual similarity and numerical distance are confounded in dot stimuli; adding (or subtracting) one dot from a given display results in a relatively small change to both numerosity and visual appearance, and likewise adding many dots changes substantially both numerosity and visual appearance. Future studies could potentially disentangle this confound, perhaps by orthogonally manipulating the size and quantity of the dots, to eliminate or at least minimize their co-variation.

Staggered cluster-sets only differed by the presence/absence of a single central item and MDS patterns displayed an effect of dot alignment. This suggests an effect of perceptual similarity. A simple template-matching ideal observer did not produce the staggered cluster pattern displayed by participants. A model as simple as the one we applied is only sensitive to low-level visual similarity driven by spatial overlap, and has no knowledge of numerosity. However, because numerosity and perceptual similarity co-vary in the dot set we have tested, it is not possible to separate effects of numerical and perceptual distances on the mental representations of these stimuli. Future work could focus on manipulations that minimize the co-variation.

The MDS dimension of alignment could be unique to the current dot stimuli. For example, this dimension may disappear if items were arranged in a circular pattern or in a different canonical form (e.g., dice patterns). Similarly, it is unclear whether the dimension of quantity would be displayed if dots were presented in randomised locations. Assessing how different canonical forms and randomised dot patterns affect the MDS space is another clear direction of future research.

Chinese symbols. In line with our predictions, Chinese MDS dimensions appeared to be arranged by perceptual dimensions of 'line terminations' (as similarly found in letters, Fiset et al., 2008) and 'alignment' (horizontal vs vertical). K-mean cluster frequency patterns depict large cluster groups between item-numbers 2–5 and 6–9. This reflects the logographic nature of the numeric-set, an effect captured by the shift from horizontal (< 5) and vertical (> 5) alignment. Although the ideal observer replicated many Chinese numeral cluster patterns, the logographic cluster pattern was not. Instead, the ideal observer focused upon similarities in horizontal line features.

Although Chinese numerals were unfamiliar to the tested cohort, the numeric-set displayed intuitive similarities between items of similar magnitude. This logographic feature may be useful numeric property. For example, an intuitive relationship between symbol and magnitude might help when initially learning the numeral system (Hung et al., 1992). Additionally, logographic numerals might aid the precision of numeric-communication within a Chinese speaking cohort (e.g., mistaking 二 for 三 is less costly than mistaking 6 for 9).

Thai symbols. In line with our predictions, Thai MDS solutions were arranged by perceptual dimensions of 'curve orientation' and 'openness' (as found in Arabic numerals, Godwin et al., 2014). K-mean cluster frequency patterns depict many low-frequency clusters, suggesting uncertainty in participant responses. Yet, regular cluster-patterns were displayed between items sharing similar perceptual features — a finding echoed by the ideal observer analysis. As predicted, Thai numerals did not display a logographic cluster pattern. Together, these findings show a clear effect of perceptual similarity on the mental space for the unfamiliar Thai numeric set.

Future research

The current study tested an English cohort, comparing multidimensional scaling solutions for familiar (Arabic and symbolic-dots) and unfamiliar (Chinese and Thai) numeric-sets. In future work, we propose to test this experimental design within a Chinese speaking cohort.

Arabic and Chinese symbols are common within Chinese speaking countries. As such, the mental representation of Arabic digits may be similar between cohorts, while Chinese symbols may be represented differently (e.g., Yeh et al., 2003). These differences might reflect familiarity with the numeric-set, (i.e., expertise), and an effect of numerical similarity. We would expect Thai symbols and symbolic dots to be represented similarly by both cohorts. However, with such different backgrounds, experiences and languages, this prediction is far from a forgone conclusion.

To disentangle perceptual from semantic effects in the mental space, we also propose two additional experiments: a perceptual matching task, and a semantic matching task. Following a similar spatial arrangement method to Godwin et al. (2014) and Goldstone (1994), we may ask participants to arrange the four numeric-sets into clusters that represent their i) perceptual similarities, and ii) semantic similarities. This method may i) further validate the perceptual results we observe in this task, and ii) examine the effect of semantic similarity on the mental space.

Conclusions

People often confuse the identity of numeric symbols. These confusions may be of little consequence, (e.g., confusing '$6' vs '$9'), or a major inconvenience (e.g., confusing '2' vs '7' eggs in a cake mix!). Past research has examined the mental dimensions of numeric item-sets, however, these results were always confounded by participant response-bias. We have presented the first bias-free mental representations of familiar (Arabic and dots) and unfamiliar (Chinese and Thai) numeric-sets. We also compared symbolic and non-symbolic mental representations of quantity. Our findings show Arabic, Thai and Chinese symbols are represented by dimensions of perceptual similarity within the mental space. Representation of non-symbolic dots could be affected by either perceptual similarity or numerical proximity, or both, however, co-variation precludes a clear inference. A clear path forward from the current study is to replicate this work in Chinese or Thai speaking cohort.

From mathematics to recipes, speed-signs to phone-numbers, our ability to perceive and communicate symbolic-quantities is critical to daily life. Understanding why fundamental cognitive mechanisms fail and confuse symbolic quantities is an important topic of human cognition. Aside from extending our understanding of numerical cognition, the findings of this study have applications in the development of future numeric fonts and item-sets. Such work must consider i) the perceptual dimensions upon which items differ, ii) whether items should convey implicit value, (i.e., be logographic), and iii) how these factors may improve the rate of symbolic learning and minimize numeric confusions.

References

Agrillo, C., Dadda, M., Serena, G., & Bisazza, A. (2008). Do fish count? spontaneous discrimination of quantity in female mosquitofish. *Animal cognition*, *11*(3), 495–503.

Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, *35*(3), 283–319.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. OUP USA.

Eidels, A., & Cassey, P. (2016). The mental representation of roman letters: Revisiting townsend's 1971 letter-identification data. In *Mathematical models of perception and cognition volume ii* (pp. 90–106). Routledge.

Eidels, A., & Gold, J. (2014). Measuring single-item identification efficiencies for letters and 3-d objects. *Behavior research methods*, *46*(3), 722–731.

Fiset, D., Blais, C., Ethier-Majcher, C., Arguin, M., Bub, D., & Gosselin, F. (2008). Features for identification of uppercase and lowercase letters. *Psychological science*, *19*(11), 1161–1168.

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*(1-2), 43–74.

Gilmore, G., Hersh, H., Caramazza, A., & Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics*, *25*(5), 425–431.

Godwin, H. J., Hout, M. C., & Menneer, T. (2014). Visual similarity is stronger than semantic similarity in guiding visual search for numbers. *Psychonomic Bulletin & Review*, *21*(3), 689–695.

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision research*, *39*(21), 3537–3560.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*(4), 381–386.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, *53*(3-4), 325–338.

Groenen, P., & Borg, I. (2014). Past, present, and future of multidimensional scaling. *Visualization and verbalization of data*, 95–117.

Hawkins, R. X., Houpt, J. W., Eidels, A., & Townsend, J. T. (2016). Can two dots form a gestalt? measuring emergent features with the capacity coefficient. *Vision research*, *126*, 19–33.

Hefner, R. (1959). Warren s. torgerson, theory and methods of scaling. new york: John wiley and sons, inc., 1958. pp. 460. *Behavioral Science*, *4*(3), 245–247.

Hung, D. L., Tzeng, O. J., & Tzeng, A. K. (1992). Automatic activation of linguistic information in chinese character recognition. In *Advances in psychology* (Vol. 94, pp. 119–130). Elsevier.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, *29*(2), 115–129.

Luce, R. D. (1963). Detection and recognition. In D. Luce (Ed.), *Handbook of mathematical psychology* (pp. 1–103). John Wiley & Sons.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological review*, *100*(2), 254.

Menninger, K. (2013). *Number words and number symbols: A cultural history of numbers*. Courier Corporation.

Mueller, S. T., & Weidemann, C. T. (2012). Alphabetic letter identification: Effects of perceivability, similarity, and bias. *Acta psychologica*, *139*(1), 19–37.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, *47*(1), 90–100.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*, *50*(2), 530–556.

Pepperberg, I. M., & Gordon, J. D. (2005). Number comprehension by a grey parrot (psittacus erithacus), including a zero-like concept. *Journal of Comparative Psychology*, *119*(2), 197.

Pleskac, T. J. (2015). Decision and choice: Luce's choice axiom. *International encyclopedia of the social & behavioral sciences*, 895–900.

Pomerantz, J. R., & Portillo, M. C. (2011). Grouping and emergent features in vision: Toward a theory of basic gestalts. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(5), 1331.

Shakkour, W. (2014). Cognitive skill transfer in english reading acquisition: Alphabetic and logographic languages compared. *Open Journal of Modern Linguistics*, *4*(04), 544.

Shepard, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika*, *27*(2), 125–140.

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

Townsend, J. (1971). Alphabetic confusion: A test of models for individuals. *Perception & Psychophysics*, *9*(6), 449–454.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological methods*, *18*(3), 368.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature reviews neuroscience*, *5*(6), 495.

Woodruff, G., & Premack, D. (1981). Primative mathematical concepts in the chimpanzee: proportionality and numerosity. *Nature*, *293*(5833), 568.

Yeh, S.-L., Li, J.-L., Takeuchi, T., Sun, V., & Liu, W.-R. (2003). The role of learning experience on the perceptual organization of chinese characters. *Visual Cognition*, *10*(6), 729–764.

Appendix A

Luce's choice model

Luce's (1963) choice model describes identification responses as probabilistic outcomes driven by the similarity of a stimulus to the others in the choice set, as well as a response-bias parameter — one for each stimulus. By estimating the parameters of the model, researchers can examine the theoretically meaningful similarity scores free from the effect of response-bias that can contaminate the observed data. Formally, the probability of making response j when presented with stimulus i can be expressed as:

$$\mathrm{C_{ij}} = \frac{\eta_{\mathrm{ij}}\beta_{\mathrm{j}}}{\sum_{\mathrm{k=1}}^{\mathrm{N}} \eta_{\mathrm{ik}}\beta_{\mathrm{k}}} \tag{1}$$

where C_{ij} is the theoretical similarity matrix for i = 1, 2...N, j = 1, 2...N. The similarity parameter η is symmetrical along the matrix diagonal i.e., $\eta_{ij} = \eta_{ji}$, and $\eta_{ii} =$ 1 for all i. In the current study, we will employ nine unique numerals, resulting in [N(N + 1)/2] - 1 = 44 free parameters to be estimated from the data.

We estimated the bias and similarity parameters of Luce's (1963) choice model using the combination of a custom Differential-Evolution Markov chain Monte-Carlo (DE-MCMC) process and maximum likelihood estimation (Myung, 2003). We initialised each of the 50 chains by estimating parameter values from Townsend's (1971) approximation of Luce's model:

$$\eta_{\mathrm{ij}} = \sqrt{\frac{\mathrm{P(R_i|S_j)P(R_j|S_i)}}{\mathrm{P(R_i|S_i)P(R_j|S_j)}}} \tag{2}$$

$$\beta_{\mathrm{j}} = \frac{1}{\mathrm{N}} \sum_{\mathrm{k=1}}^{\mathrm{N}} \sqrt{\frac{\mathrm{P(R_j|S_j)P(R_k|S_j)}}{\mathrm{P(R_j|S_k)P(R_k|S_k)}}} \tag{3}$$

where R is the response probability given stimulus S; then adding uniformly sampled noise. On each iteration, each chain proposed updated parameter estimates by weighting the previous estimates with the estimates of two randomly selected chains using the weighting formula outlined by Turner, Sederberg, Brown, and Steyvers (2013). The log-likelihood of these new parameters and the previous ones were computed by generating an expected confusion matrix (using the estimated parameters and Luce's choice model) and comparing to the observed data, with the parameters that maximised the log-likelihood being kept. After 500 iterations the parameters from the chain with the highest log-likelihood were used for further analysis.

Appendix B

Experimental accuracy by contrast level and participant

The following appendix examines accuracy during experimental trials, over the five levels of contrast. We show that our manipulation of contrast appropriately influenced accuracy, that accuracy was relatively stable across blocks, and that accuracy was close to 60% for all participants, across conditions of numeric-type (Arabic, Chinese, Thai and dot numerals).

During experimental trials, stimuli were presented at five signal-levels, one step below the critical contrast value (level 1: hardest), and three steps above (levels 3, 4 and 5: easiest). As shown in Figure B1.a., across numeric-types, mean accuracy increased linearly with the visibility of the contrast levels, being lowest at level 1 (μ = .32, σ = .02) and highest at level 5 (μ = .8, σ = .03). On average, accuracy was highest for Chinese numerals (μ = .6, σ = .21), then Arabic (μ = .59, σ = .19) and non-symbolic dots (μ = .59, σ = .19), and finally, Thai numerals (μ = .54, σ = .19).

A repeated-measures ANOVA found a significant main effect of contrast level on accuracy ($F(4, 40) = 447.914$, $p <.001$, $\eta^2 = .99$), but not a main effect of numeric-type on accuracy ($F(3, 30) = 2.134$, $p = 0.12$, $\eta^2 = .18$). There was no interaction effect between numeric-type and contrast level on accuracy ($F(12, 120) = 0.951$, $p =.5$, $\eta^2 = .09$). Post-hoc pair-wise *t*-tests using the Bonferroni correction revealed significant differences between all combinations of contrast level ($p < .001$). By contrast, pair-wise *t*-tests showed no difference in accuracy between numeric-types, except between familiar items, Arabic numerals and non-symbolic dots ($p < .05$). All simple effects are reported in the supplementary materials, Tables S1.1 and S1.2. These results indicate our chosen signal levels appropriately influenced response accuracy. However, there appears to be no effect of numeric familiarity on response-accuracy. We will revisit this line of inquiry shortly.

Figure B1.b. depicts mean accuracy across experimental blocks, for each numeric-type. Mean accuracy was comparable between numeric-types, and increased marginally with block number, being lowest at block 1 (μ = .50, σ = .14) and highest at

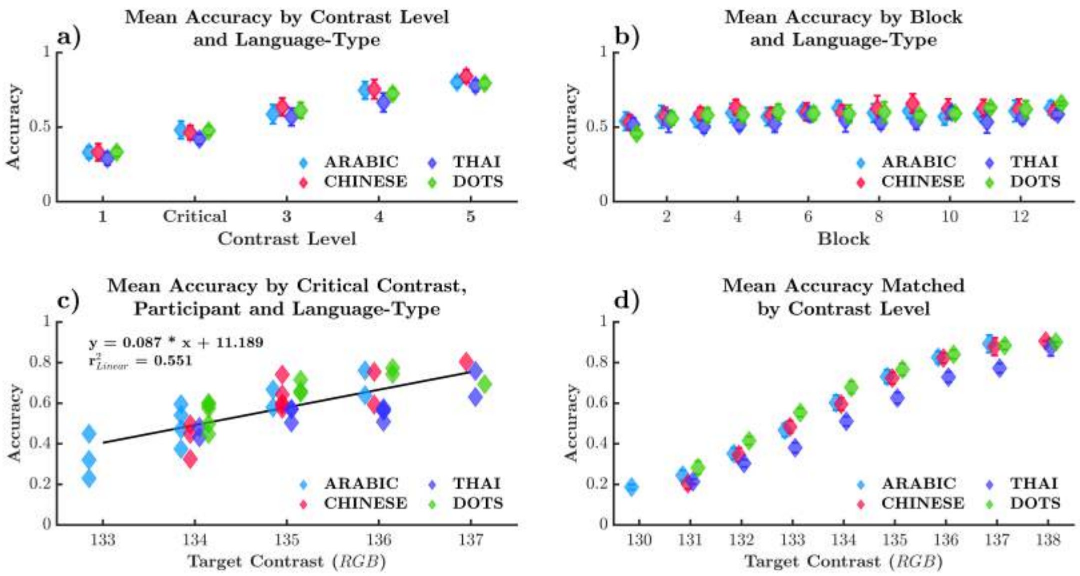

Figure B1. a) Mean accuracy across five signal contrast-levels, and four numeric-types. b) Mean accuracy across each experimental block. c) Mean accuracy for each participant by critical contrast level. d) Mean accuracy matched by contrast-level, across numeric-types. Error bars represent the standard-error of the mean.

block 13 (μ = .62, σ = .13). A repeated-measures ANOVA found a significant main effect of block on accuracy (F(12, 120) = 14.733, p <.001, η^2 = .6), and no main effect of numeric-type on accuracy (F(3, 30) = 2.139, p =0.12, η^2 = .18). There was no interaction effect of numeric-type and block on accuracy (F(36, 360) = .975, p =.51, η^2 = .09). Post-hoc pair-wise analysis revealed significant differences in accuracy between early and late experimental blocks. Block 1 differed significantly from blocks 5–13 (p < .01), block 2 from blocks 12–13 (p < .01) and block 3 from blocks 9, 11 and 13 (p < .05). Simple effects are reported in the supplementary materials, Table S1.3. These results suggest a small practice effect, slightly boosting accuracy in later blocks.

Figure B1.c. presents mean experimental accuracy across critical contrast levels, separated by participant and numeric-type. A linear regression found a significant positive relationship between critical contrast and mean accuracy (r^2 = .551), suggesting a dependency between contrast and accuracy. To disentangle the effect of numeric-type and contrast on accuracy, we assessed accuracy matched across RGB values from each participant's five signal-contrast levels (see B1.d).

Figure B1.d. presents mean accuracy matched across participant's five contrast-levels, separated by numeric-type. For example, if for Arabic numerals, participant S1 responded to RGB contrast values 130–134 and participant S2 responded to RGB contrast values 134–137, their accuracy at contrast value 134 would be averaged and depicted in Figure B1.d.

Figure B1.d. displays a positive relationship between contrast and matched accuracy. Matching accuracy for contrast levels when all numeric-types were presented, (i.e., excluding contrast values 130 and 138), accuracy was highest for non-symbolic dots (μ = .63, σ = .22), then Arabic numerals (μ = .59, σ = .24), then Chinese numerals (μ = .58, σ = .25) and lowest for Thai numerals (μ = .51, σ = .21).

We completed a two-way between-subjects ANOVA to assess the effect of numeric-type and contrast-level on matched accuracy (Figure B1.d). We found a main effect of numeric-type ($F(3, 185) = 15.606$, $p < .001$, $\eta^2 = 0.04$), and a main effect of contrast-level ($F(6, 185) = 148.814$, $p < .001$, $\eta^2 = 0.79$) on accuracy. There was no interaction effect between contrast level and numeric-type on accuracy ($F(18, 185) = 0.003$, $p = .99$, $\eta^2 = 0.01$). Post-hoc pair-wise t-tests displayed significant differences between all contrast values ($p < .001$), except between the highest RGB values, 136 and 137 (all pair-wise tests are reported in the supplementary materials, Table S1.. Post-hoc t-tests displayed a significant differences in accuracy between all numeric-types ($p < .05$), except for comparisons between Chinese and Thai, and Arabic and Thai numerals (all pair-wise tests reported in the supplementary materials, Table S1.. These results show a clear effect of contrast-level on accuracy. After accounting for contrast level, trends indicate that accuracy was higher for familiar items (Dots and Arabic) compared to unfamiliar items (Chinese and Thai), however, this was not borne out by the simple effects.

Supplementary Material S1.

Simple effects: t-tests

Table S1.1

Post-hoc comparisons between contrast levels. Level 1 being the lowest signal contrast level (hardest) and level 5 being the highest (easiest). Level 2 is elsewhere referred to as the critical contrast level.

		Mean Difference	SE	t	Cohen's d	p_{bonf}
Level 1	Level 2	-0.113	0.009	-12.34	-3.720	$< .001$
	Level 3	-0.241	0.014	-17.13	-5.166	$< .001$
	Level 4	-0.355	0.016	-21.97	-6.623	$< .001$
	Level 5	-0.440	0.016	-26.89	-8.106	$< .001$
Level 2	Level 3	-0.128	0.009	-13.97	-4.211	$< .001$
	Level 4	-0.242	0.011	-21.72	-6.550	$< .001$
	Level 5	-0.327	0.015	-22.55	-6.798	$< .001$
Level 3	Level 4	-0.114	0.006	-18.67	-5.629	$< .001$
	Level 5	-0.199	0.009	-22.75	-6.860	$< .001$
Level 4	Level 5	-0.085	0.008	-10.29	-3.104	$< .001$

Table S1.2

Post-hoc comparisons between numeric-types.

		Mean Difference	SE	t	Cohen's d	p_{bonf}
ARABIC	CHINESE	-0.085	0.051	-1.678	-0.506	0.746
	THAI	-0.049	0.063	-0.766	-0.231	1.000
	DOTS	-0.120	0.036	-3.353	-1.011	0.044
CHINESE	THAI	0.037	0.058	0.640	0.193	1.000
	DOTS	-0.035	0.041	-0.839	-0.253	1.000
THAI	DOTS	-0.072	0.044	-1.616	-0.487	0.822

Table S1.3
Post-hoc comparisons of accuracy by block number.

		Mean Difference	SE	t	p_{bonf}
Block1	Block2	-0.032	0.014	-2.257	1.000
	Block3	-0.059	0.015	-3.992	0.199
	Block4	-0.072	0.015	-4.857	0.052
	Block5	-0.066	0.013	-5.209	0.031
	Block6	-0.082	0.013	-6.336	0.007
	Block7	-0.078	0.011	-6.756	0.004
	Block8	-0.080	0.013	-6.188	0.008
	Block9	-0.096	0.012	-8.144	$< .001$
	Block10	-0.092	0.012	-7.834	0.001
	Block11	-0.096	0.014	-6.623	0.005
	Block12	-0.096	0.014	-6.747	0.004
	Block13	-0.115	0.017	-6.841	0.004
Block2	Block3	-0.026	0.013	-1.961	1.000
	Block4	-0.040	0.014	-2.799	1.000
	Block5	-0.034	0.017	-2.004	1.000
	Block6	-0.049	0.013	-3.660	0.342
	Block7	-0.045	0.012	-3.891	0.234
	Block8	-0.048	0.016	-2.955	1.000
	Block9	-0.063	0.015	-4.334	0.116
	Block10	-0.060	0.016	-3.749	0.296
	Block11	-0.064	0.016	-3.870	0.243
	Block12	-0.064	0.009	-7.064	0.003
	Block13	-0.082	0.013	-6.105	0.009
Block3	Block4	-0.014	0.012	-1.115	1.000
	Block5	-0.007	0.009	-0.806	1.000
	Block6	-0.023	0.009	-2.597	1.000

Table S1.3 continued from previous page

		Mean Difference	SE	t	p_{bonf}
	Block7	-0.019	0.008	-2.338	1.000
	Block8	-0.021	0.010	-2.127	1.000
	Block9	-0.037	0.007	-5.200	0.031
	Block10	-0.034	0.011	-2.964	1.000
	Block11	-0.037	0.006	-5.826	0.013
	Block12	-0.038	0.008	-4.837	0.053
	Block13	-0.056	0.008	-7.089	0.003
Block4	Block5	0.006	0.014	0.456	1.000
	Block6	-0.009	0.012	-0.772	1.000
	Block7	-0.005	0.010	-0.527	1.000
	Block8	-0.008	0.010	-0.791	1.000
	Block9	-0.023	0.010	-2.409	1.000
	Block10	-0.020	0.014	-1.468	1.000
	Block11	-0.024	0.011	-2.203	1.000
	Block12	-0.024	0.009	-2.758	1.000
	Block13	-0.042	0.013	-3.359	0.566
Block5	Block6	-0.016	0.011	-1.430	1.000
	Block7	-0.012	0.011	-1.104	1.000
	Block8	-0.014	0.009	-1.518	1.000
	Block9	-0.030	0.010	-3.108	0.865
	Block10	-0.027	0.010	-2.591	1.000
	Block11	-0.030	0.007	-4.418	0.101
	Block12	-0.030	0.012	-2.495	1.000
	Block13	-0.049	0.011	-4.255	0.131
Block6	Block7	0.004	0.006	0.689	1.000
	Block8	0.002	0.009	0.162	1.000
	Block9	-0.014	0.010	-1.468	1.000

Table S1.3 continued from previous page

		Mean Difference	SE	t	p_{bonf}
	Block10	-0.011	0.011	-1.020	1.000
	Block11	-0.014	0.009	-1.640	1.000
	Block12	-0.015	0.009	-1.636	1.000
	Block13	-0.033	0.007	-4.881	0.050
Block7	Block8	-0.003	0.007	-0.341	1.000
	Block9	-0.018	0.008	-2.217	1.000
	Block10	-0.015	0.012	-1.200	1.000
	Block11	-0.018	0.008	-2.332	1.000
	Block12	-0.019	0.008	-2.482	1.000
	Block13	-0.037	0.008	-4.783	0.058
Block8	Block9	-0.016	0.008	-1.911	1.000
	Block10	-0.012	0.012	-1.019	1.000
	Block11	-0.016	0.005	-2.959	1.000
	Block12	-0.016	0.011	-1.506	1.000
	Block13	-0.035	0.011	-3.231	0.703
Block9	Block10	0.003	0.010	0.344	1.000
	Block11	-2.525e-4	0.007	-0.034	1.000
	Block12	-5.051e-4	0.009	-0.056	1.000
	Block13	-0.019	0.011	-1.684	1.000
Block10	Block11	-0.004	0.011	-0.312	1.000
	Block12	-0.004	0.011	-0.335	1.000
	Block13	-0.022	0.012	-1.815	1.000
Block11	Block12	-2.525e-4	0.010	-0.025	1.000
	Block13	-0.019	0.009	-2.119	1.000
Block12	Block13	-0.018	0.007	-2.729	1.000

Table S1.4
Post-hoc comparisons of accuracy matched by contrast level, for RGB contrast values 131–136.

		Mean Difference	SE	t	Cohen's d	p_{bonf}
131	132	-0.118	0.028	-4.227	-1.572	$< .001$
	133	-0.235	0.027	-8.851	-2.504	$< .001$
	134	-0.361	0.026	-13.731	-3.528	$< .001$
	135	-0.476	0.027	-17.916	-5.024	$< .001$
	136	-0.568	0.029	-19.850	-7.684	$< .001$
	137	-0.621	0.034	-18.528	-8.921	$< .001$
132	133	-0.117	0.021	-5.654	-1.240	$< .001$
	134	-0.243	0.020	-11.925	-2.394	$< .001$
	135	-0.358	0.021	-17.265	-3.757	$< .001$
	136	-0.450	0.023	-19.309	-5.617	$< .001$
	137	-0.503	0.029	-17.277	-6.344	$< .001$
133	134	-0.125	0.019	-6.764	-1.150	$< .001$
	135	-0.240	0.019	-12.700	-2.306	$< .001$
	136	-0.333	0.022	-15.312	-3.505	$< .001$
	137	-0.386	0.028	-13.839	-3.947	$< .001$
134	135	-0.115	0.018	-6.233	-1.054	$< .001$
	136	-0.208	0.021	-9.728	-2.037	$< .001$
	137	-0.261	0.028	-9.452	-2.462	$< .001$
135	136	-0.092	0.022	-4.262	-0.968	$< .001$
	137	-0.146	0.028	-5.224	-1.478	$< .001$
136	137	-0.053	0.030	-1.778	-0.675	1.000

Table S1.5
Post-hoc comparisons of accuracy matched by contrast level, across numeric-types.

		Mean Difference	SE	t	Cohen's d	p_{bonf}
ARABIC	CHINESE	-0.052	0.018	-2.857	-0.261	0.029
	DOTS	0.074	0.019	3.891	0.379	$< .001$
	THAI	-0.009	0.019	-0.472	-0.043	1.000
CHINESE	DOTS	0.126	0.019	6.779	0.674	$< .001$
	THAI	0.043	0.019	2.309	0.213	0.132
DOTS	THAI	-0.083	0.019	-4.271	-0.420	$< .001$

Supplementary Material S2.

Scree analysis of bias-free MDS stress values

Scree analysis compares the multidimensional stress values (y-axis) against the number of MDS dimensions (x-axis). Scree analysis, such as this, is a subjective measure. A useful heuristic for identifying the correct number of dimensions is to look for the 'elbow' where an increase in dimensions does not meaningfully improve stress values. This elbow has been identified by a marker in each plot.

Scree Plots

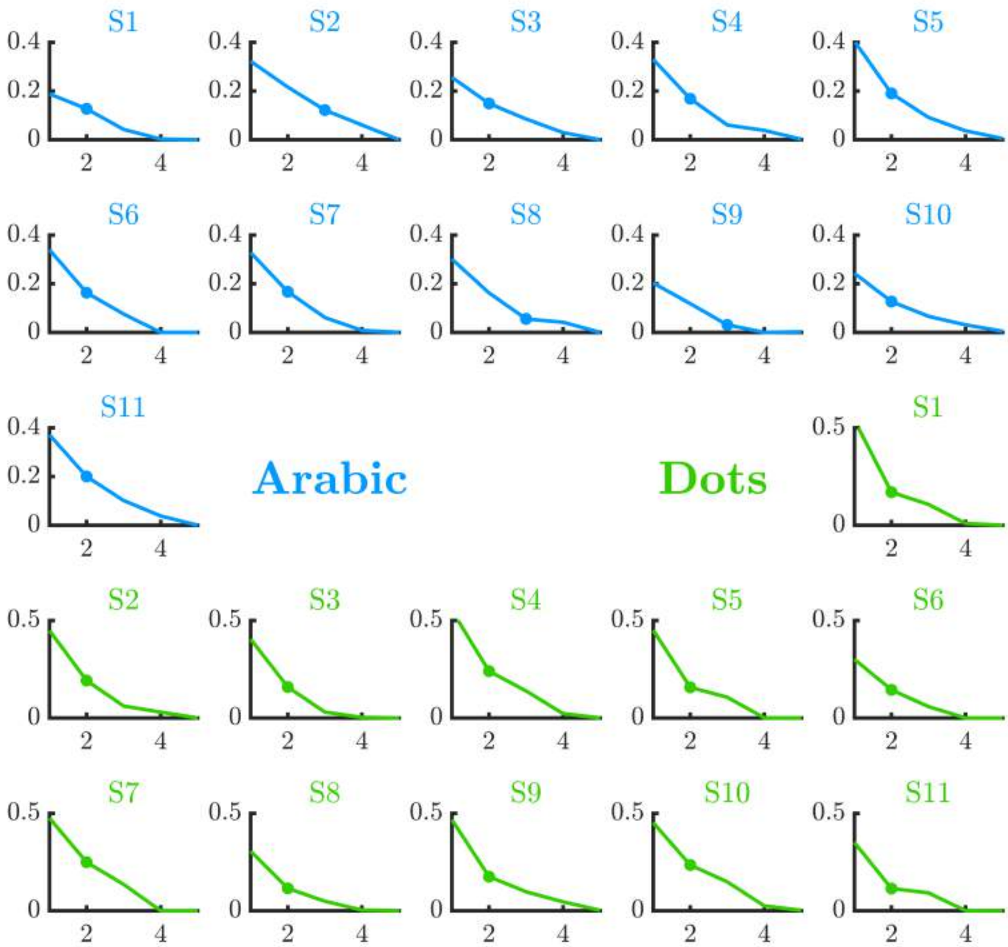

Figure S2.1. Bias-free MDS scree plots for Arabic digits (blue) and symbolic dots (green). The y-axis displays stress values, and the x-axis the number of dimensions. Markers identify the optimal number of dimensions in each scree plot.

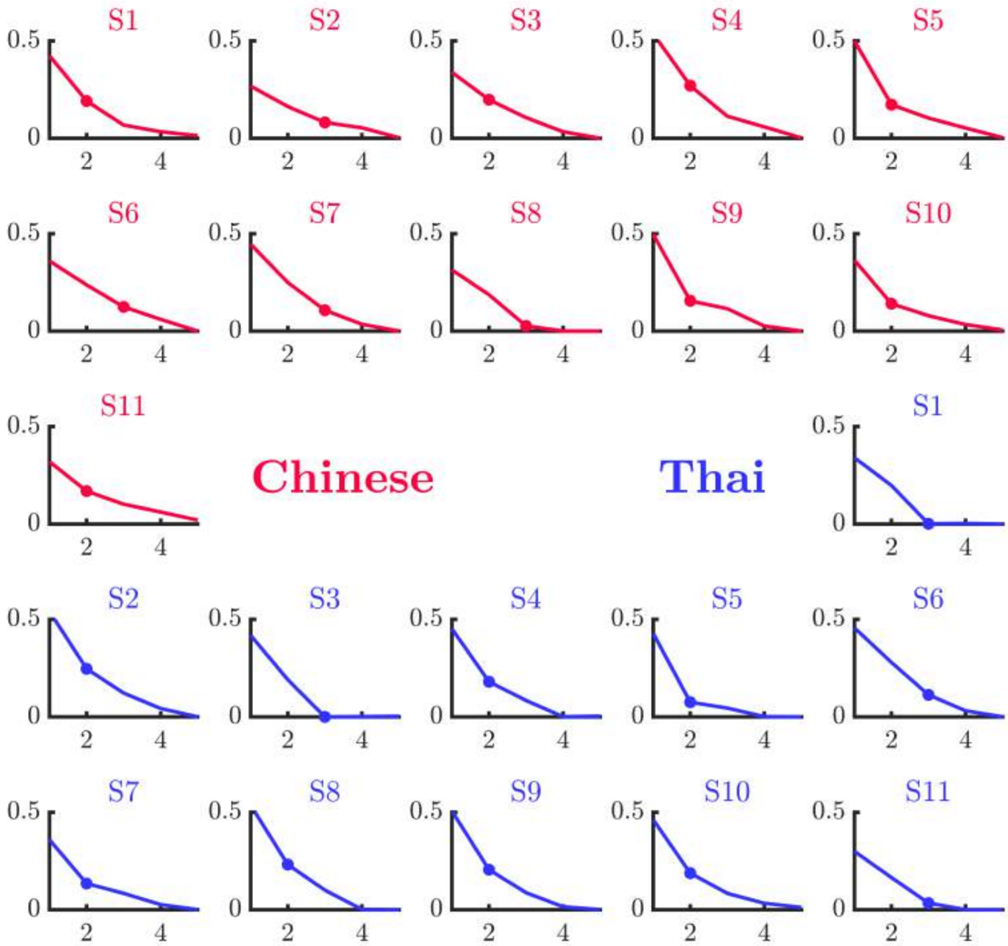

Figure S2.2. Bias-free MDS scree plots for Chinese (red) and Thai (purple) symbols. The y-axis displays stress values, and the x-axis the number of dimensions. Markers identify the optimal number of dimensions in each scree plot.

Individual MDS solutions

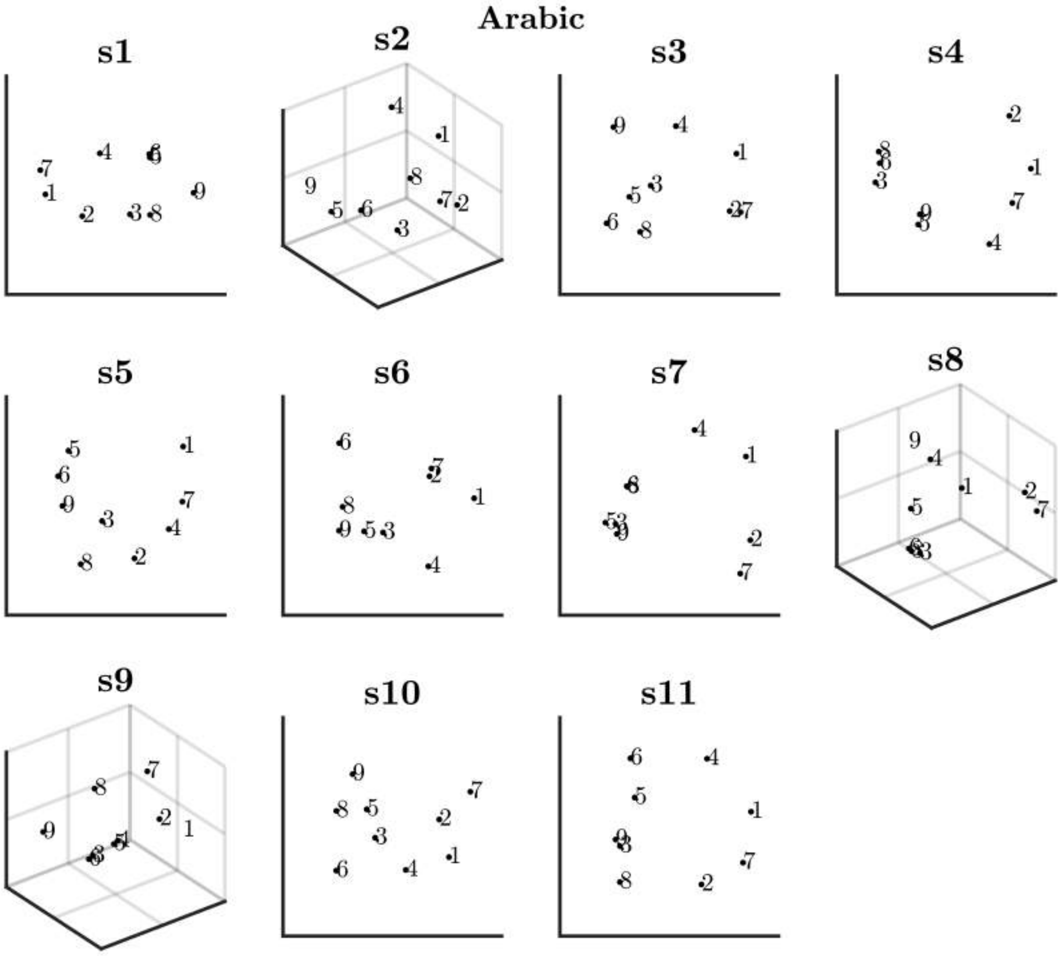

Figure S2.3. Individual bias-free MDS solutions for the Arabic digits.

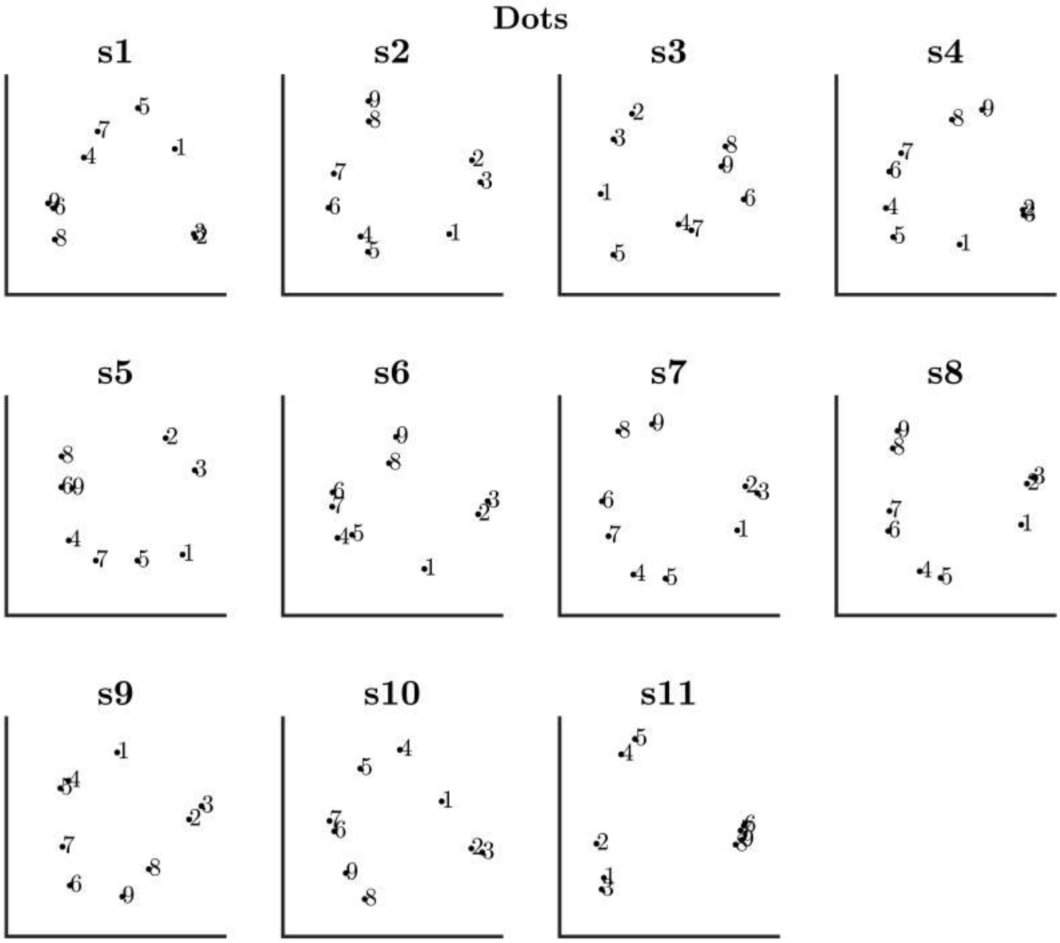

Figure S2.4. Individual bias-free MDS solutions for symbolic dots. Dots are represented by Arabic numbers for simplicity.

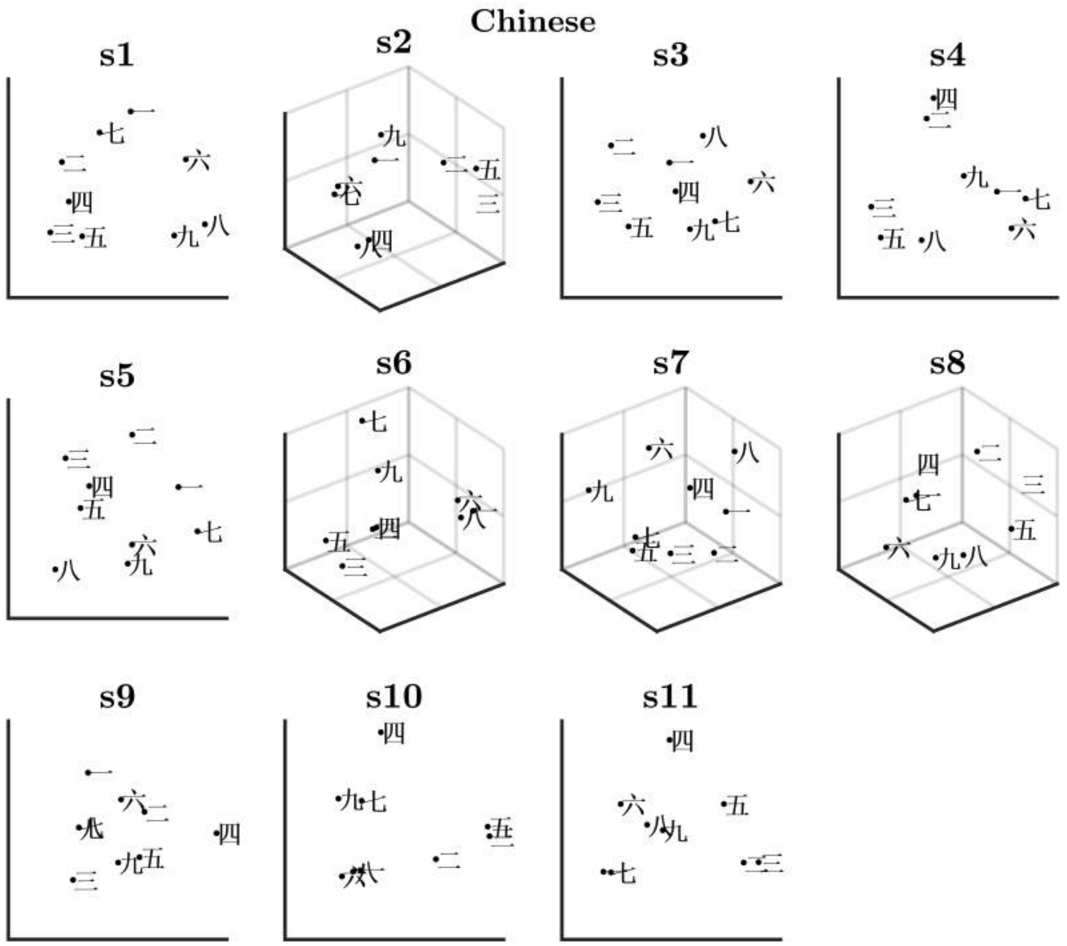

Figure S2.5. Individual bias-free MDS solutions for Chinese symbols.

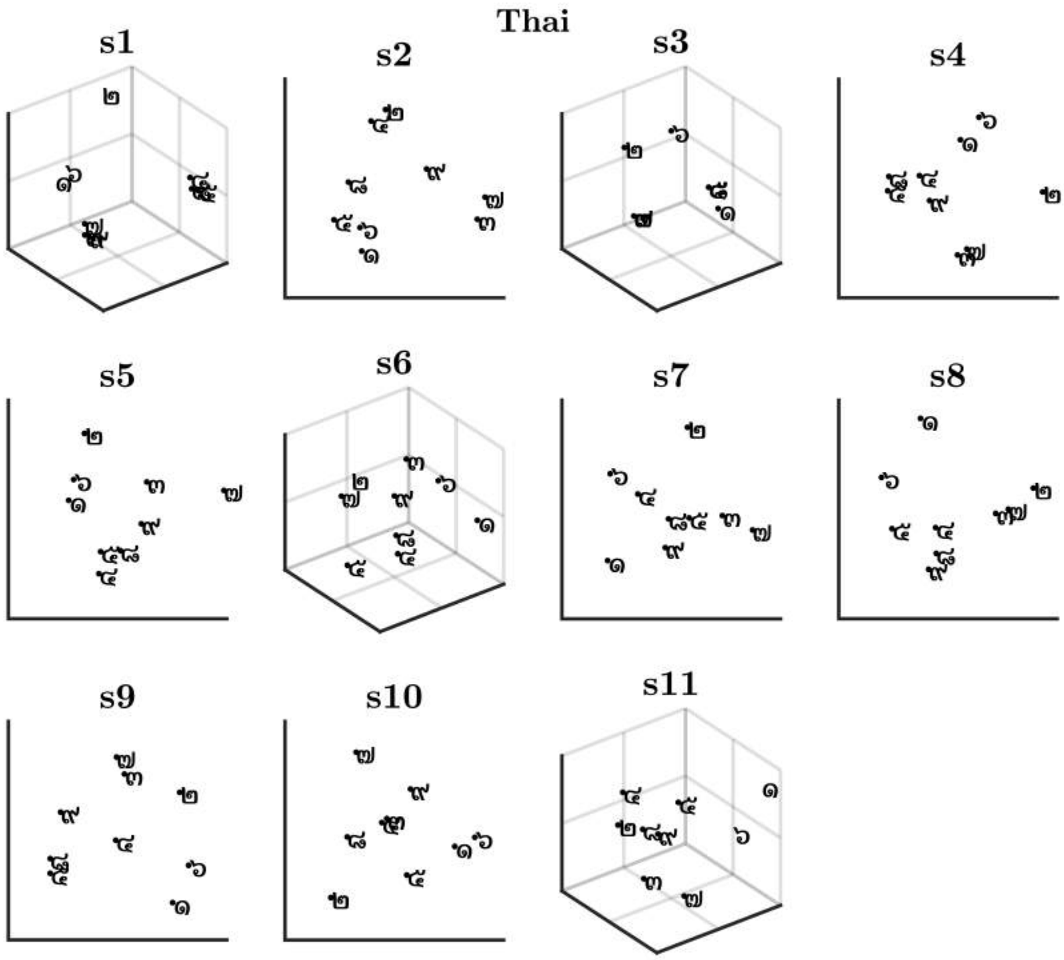

Figure S2.6. Individual bias-free MDS solutions for the Thai symbols.

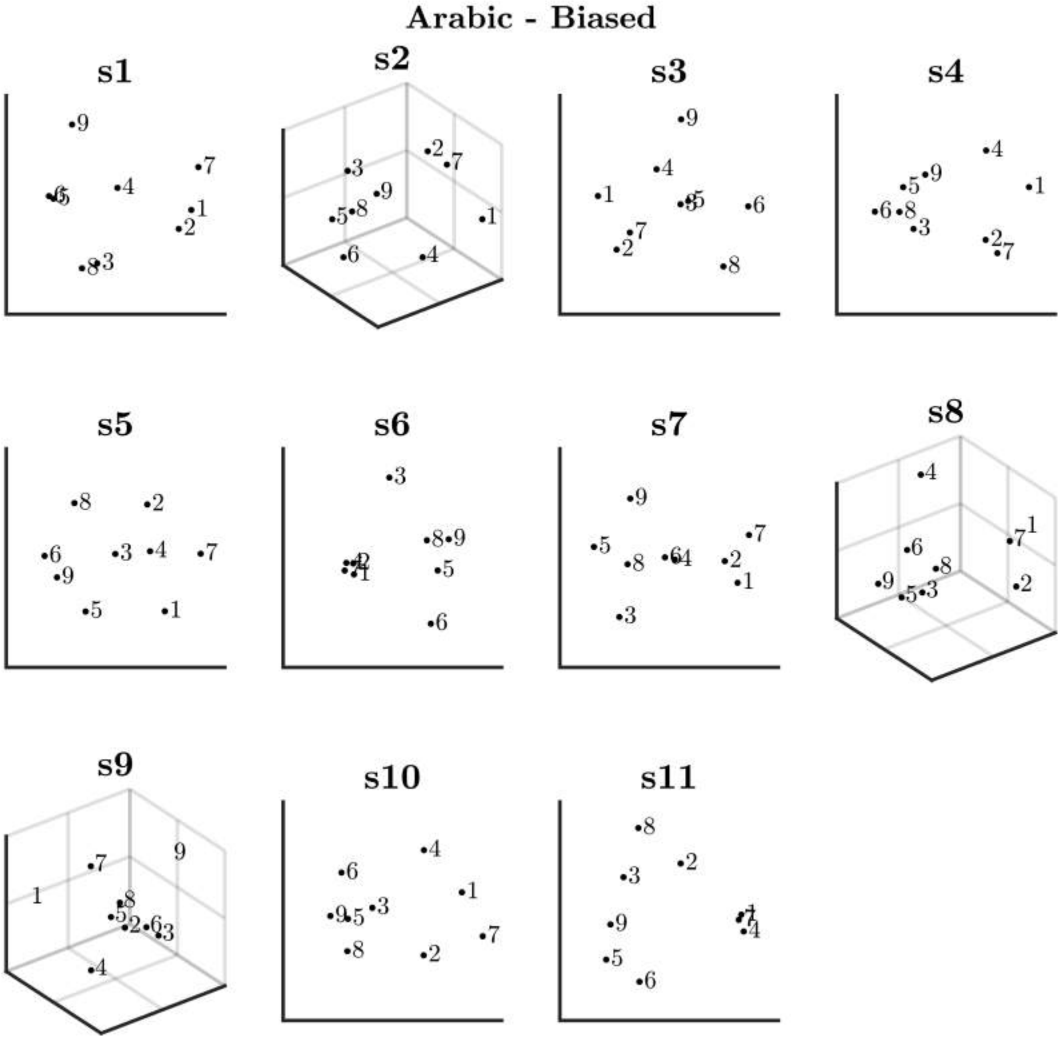

Figure S2.7. Individual biased MDS solutions for the Arabic digits.

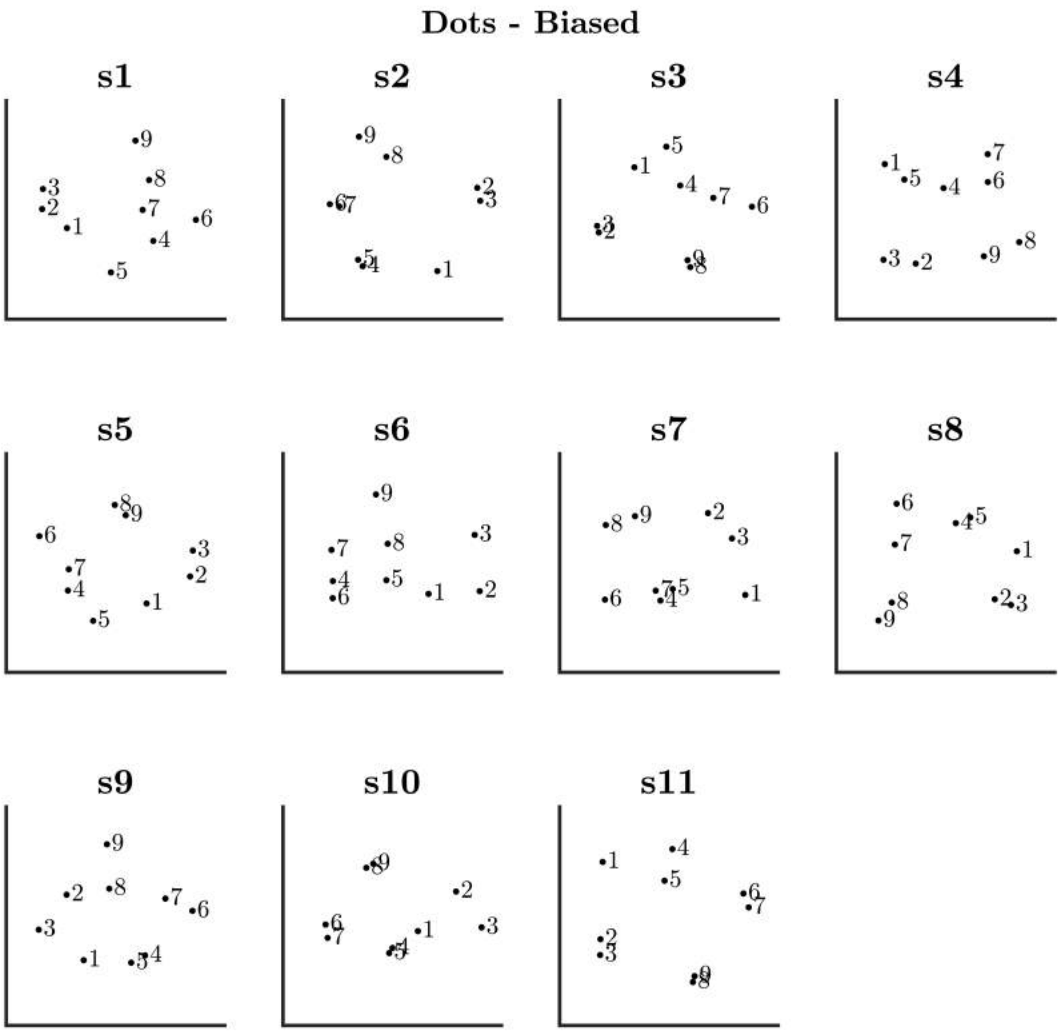

Figure S2.8. Individual biased MDS solutions for symbolic dots. Dots are represented by Arabic numbers for simplicity.

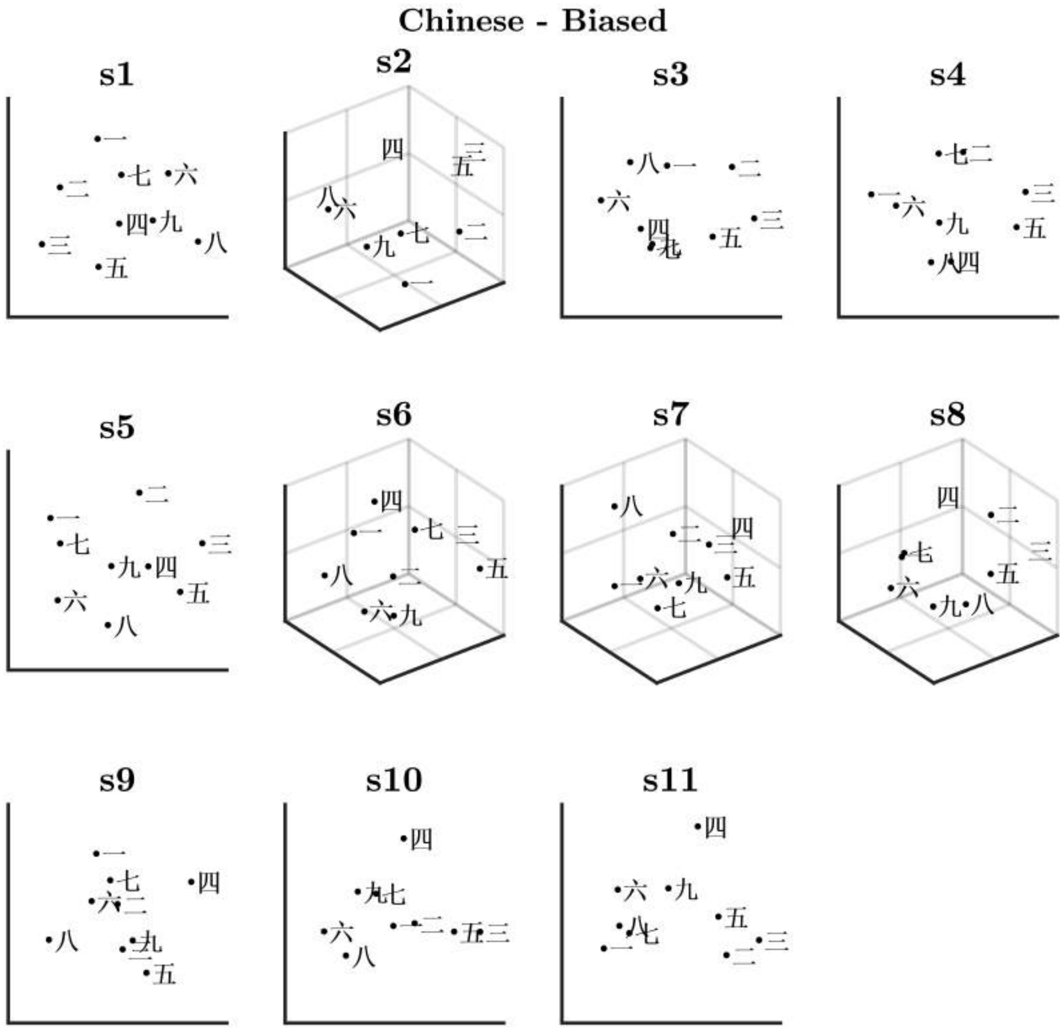

Figure S2.9. Individual bias-free MDS solutions for Chinese symbols.

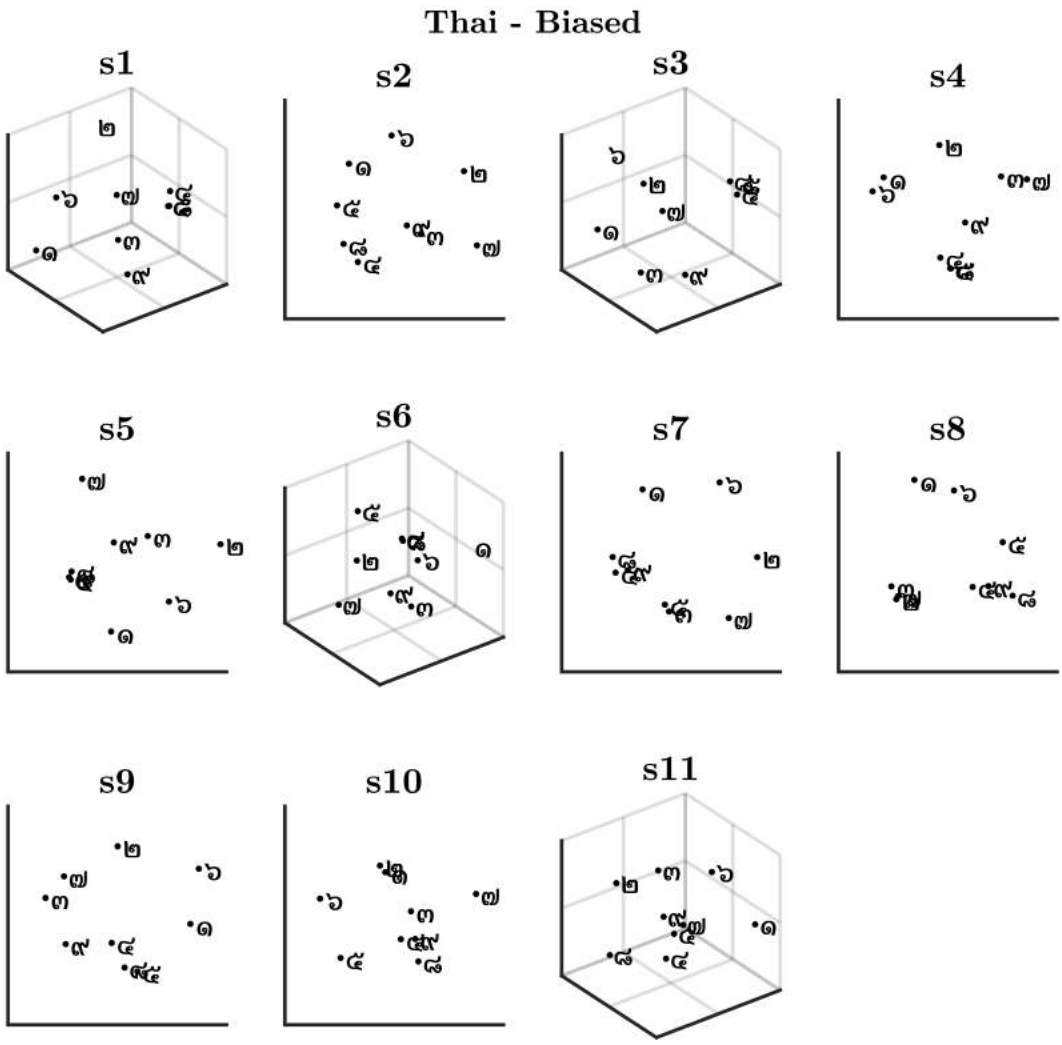

Figure S2.10. Individual bias-free MDS solutions for Thai symbols.

Supplementary Material S3.

MDS cluster frequency heatmaps

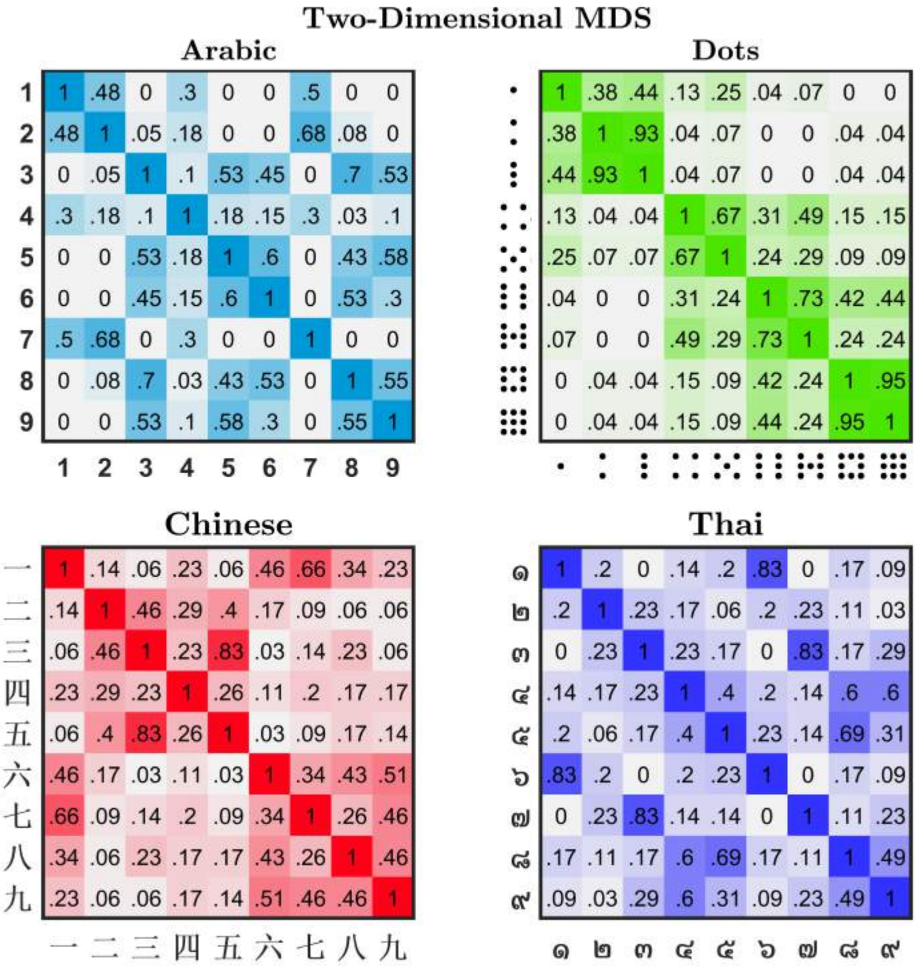

Figure S3.1. Proportional cluster-frequency heatmap for participants with two-dimensional MDS solutions, across 2–6 K-mean clusters. Larger proportions (darker colored squares) indicate items which most frequently cluster together.

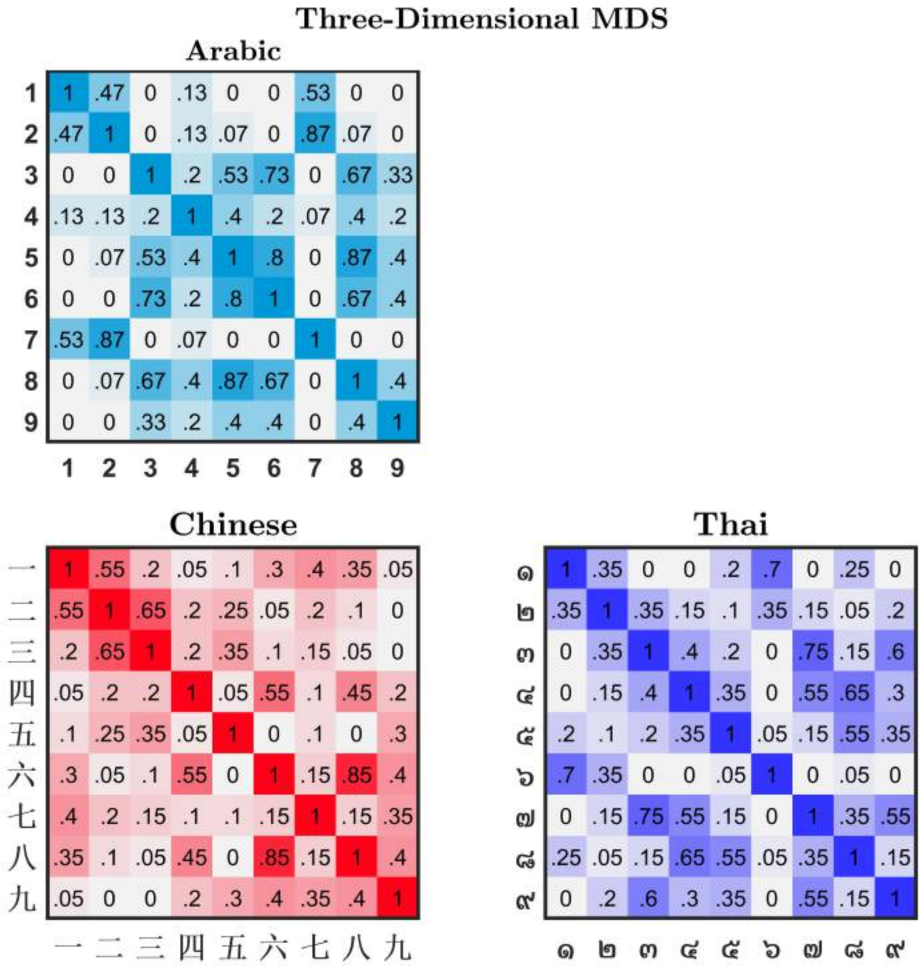

Figure S3.2. Proportional cluster-frequency heatmap for participants with three-dimensional MDS solutions, across 2–6 K-mean clusters. Larger proportions (darker colored squares) indicate items which most frequently cluster together.

Supplementary Material S4.

Three dimensional group indscal solutions

Figure S4.1 displays the group indscal MDS and K-mean cluster frequency results for those participants identified with three MDS dimensions. No participants displayed a third MDS dimension in the symbolic-dot numeric-type. MDS and cluster frequency plots are comparable between three-dimensional and two-dimensonal indscal results. Arabic items displayed similar clusters, however, item ‘9’ shifts from being grouped with items ‘3’ and ‘8’, to being grouped with items ‘5’ and ‘6’. Chinese results are comparable between two- and three-dimensional plots, except in the three-dimensional plot, item 四 shifts away from all items along the third-dimension. Finally, similar results were observed in the three-dimensional Thai MDS and cluster-frequency plots, except that item-numbers [1,6] move away from all other items along the third-dimension.

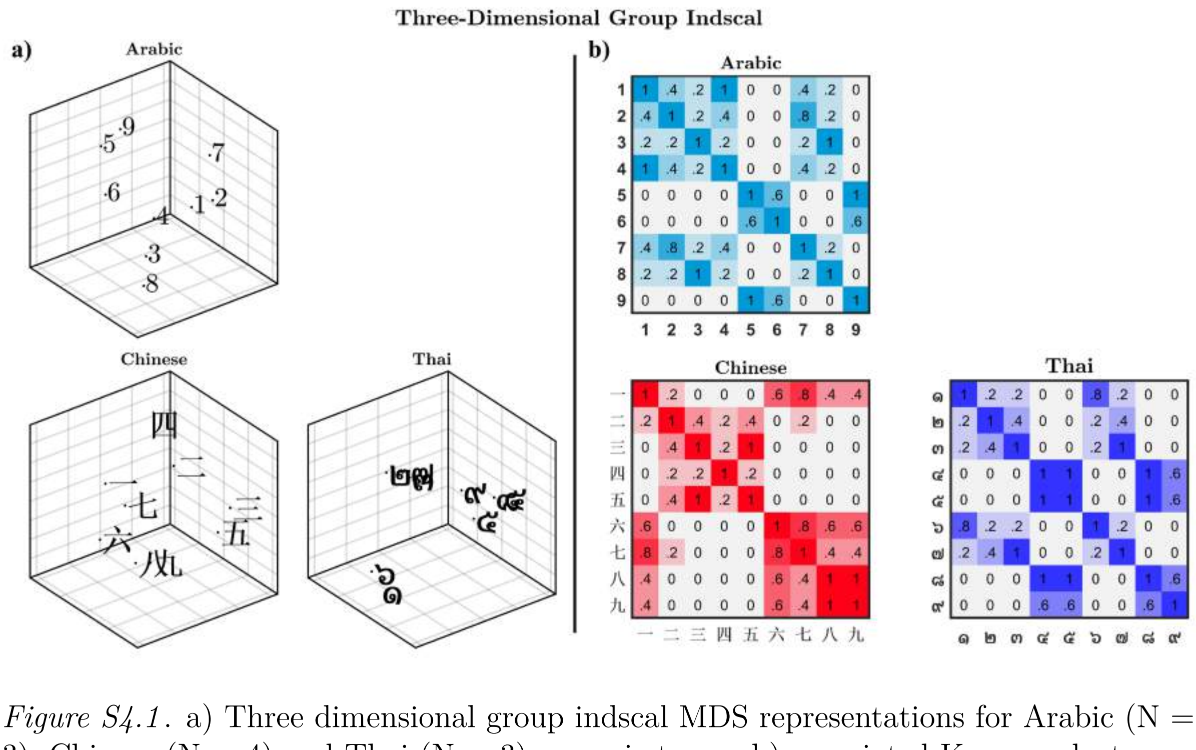

Figure S4.1. a) Three dimensional group indscal MDS representations for Arabic (N = 3), Chinese (N = 4) and Thai (N = 3) numeric-types. b) associated K-mean cluster frequency heat maps for three dimensional indscal MDS solutions.