Journal of Cognitive Psychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pecp21

JAM-boree: An application of observation oriented modelling to judgements of associative memory

Kathrene D. Valentine [a] & Erin M. Buchanan [a]
[a] Missouri State University , Springfield , MO , USA

Journal of Cognitive Psychology, 2013
Vol. 25, No. 4, 400–422, http://dx.doi.org/10.1080/20445911.2013.775120

JAM-boree: An application of observation oriented modelling to judgements of associative memory

Kathrene D. Valentine and Erin M. Buchanan

Missouri State University, Springfield, MO, USA

Null hypothesis significance testing is criticised for emphatically focusing on using the appropriate statistic for the data and an overwhelming concern with low *p*-values. Here, we present a new technique, Observation Oriented Modeling (OOM), as an alternative to traditional techniques in the social sciences. Ten experiments on judgements of associative memory (JAM) were analysed with OOM to show data analysis procedures and the consistency of JAM results across several types of experimental manipulations. In a typical JAM task, participants are asked to rate the frequency of word pairings, such as LOST-FOUND, and are then compared to actual normed associative frequencies to measure how accurately participants can judge word pairs. Three types of JAM tasks are outlined (traditional, paired, and instructional manipulations) to demonstrate how modelling complex hypotheses can be applied through OOM to this type of data that would be conventionally analysed with null hypothesis significance testing.

Keywords: Association; Judgements; Memory; Statistics; Techniques.

Before psychology was recognised as a field of science, researchers used associative memory tasks to examine memory structure, learning, and even to try to detect crimes (Münsterberg, 1908). Association is defined as the first word that pops into the head when a target is presented (Spence & Owens, 1990). For instance, the terms ROCK and ROLL are associated through their combined use to describe a genre of music. An operational definition of association comes from the free association task (Nelson, McEvoy, & Dennis, 2000). Participants are given cue words (BOOK) and asked to list the first word that comes to mind (READ). These responses are then averaged over many participants to create a probability of each cue word triggering a target response word, as available in Nelson, McEvoy, and Schreiber's (2004) free association norms. Forward strength (FSG) is the probability of a cue word eliciting a target word, such as the likelihood of READ after being given BOOK. Backward strength (BSG) is the probability of the target word eliciting the cue word, or how many times BOOK is given to the READ cue.

In the judgements of associative memory task (JAM), participants are given a combined cue–target pair, instead of only cue words (Maki's, 2007a, nomenclature; Koriat, Fiedler, & Bjork, 2006). They are instructed to estimate the number of people out of a 100 that would list the target word when shown the cue word. For example, when shown LOST-FOUND, participants should estimate that approximately 75 people would list FOUND when given LOST. Participants are given a variety of cue–target pairings to estimate and their scores are compared against

Correspondence should be addressed to Erin M. Buchanan, Department of Psychology, Missouri State University, 901 S. National Ave., Springfield, MO 65897, USA. E-mail: erinbuchanan@missouristate.edu

The authors would like to thank Drs. James Grice and William S. Maki, and our reviewers, for their advice and corresponding reviews on this manuscript.

the Nelson et al. database norms. These judgements of word-pair frequency turn out to be alarmingly incorrect, comparable to the difficulties seen in the judgements of learning literature. In a judgement of learning task, subjects are asked to estimate how likely they are to remember a paired combination on a subsequent test. Koriat and Bjork (2005, 2006) have shown that participants usually overestimate their likelihood of remembering information for the future test.

While the paired associates task is nothing new (Garskof & Forrester, 1966; Kamman, 1968), Maki's (2007a) research explored and defined participant issues with associative judgements. Over seven experiments, he showed that participants overestimate associative strength with different materials, rating scales, semantics, and when presented with viable target alternatives. The JAM function is the regression equation of forward strength predicting participant judgements; JAM = Intercept + Slope * FSG. Perfect alignment of the judgement process to associative strength would display slope values close to 1 and very low intercepts (close to zero). What is striking about Maki's results is the regularity of intercept and slope values. The intercepts range from approximately 43 to 60 points, while slope values range from approximately 0.2 to 0.45, with one high extreme of 0.56. Similar work by Nelson, Dyrdal, and Goodmon (2005) also portrays a pattern of very low slope values, along with high overestimation intercept scores.

The intercept is an overall bias in estimation of the relation between word pairs, while slopes indicate the sensitivity to the distinction in word-pair frequencies (Maki, 2007b). Traditionally, testing bias and sensitivity would be performed using null hypothesis significance testing (NHST) with regression analyses. Normed database information is used to predict participant judgement scores (as described earlier), and these values would be tested against a not predictive (zero) null hypothesis separately. To support that participants are at least somewhat sensitive to the difference between low and high frequency strengths, the null hypothesis of zero slope would need to be rejected. However, if we wanted to examine if participants are *not* biased, the null hypothesis for intercept values would need to be accepted and not rejected, which is not well supported in journal article publication (Christopher & Brannick, 2012; Masicampo & Lalande, 2012; Rosenthal, 1979). Further, if we wished to test both low bias and high sensitivities (or any other combinations) simultaneously, few traditional analyses with NHST could accommodate such a composite hypotheses set.

The technique suggested in this article as a solution to both NHST pitfalls and learning complicated model testing procedures is Observation Oriented Modeling (OOM; Grice, 2011). OOM is a statistical technique developed by Grice (2011; see Grice, Barrett, Schlimgen, & Abramson, 2012, for recent publication), which focuses on individual observations instead of averages (which would be necessary for NHST on JAM). This approach is based on the philosophical ideas of Aristotle and Aquinas, namely that of philosophical realism, where a phenomenon's effects conform to their cause. In line with this type of thinking, researchers should focus on seeing things not as having one direct cause and effect, but as an integrated model where many factors are combined with an effect as the outcome. This line of thinking also eliminates many statistical assumptions and definition of population parameters to focus instead on the observed individual data. Grice emphasises that we rarely know the actual population, and thus should instead focus on our methodologies, our obtained data, and replication; especially given recent studies on the lack of replicability in psychology (Spies et al., 2012).

Here, we aim to examine the consistency of the output of the memory judgement process, specifically the bias and sensitivity of participants across 10 different JAM studies. Two models were compared: a perfect model where participants are accurately able to discern the difference between low and high frequency relations with no to little bias, and a low sensitivity model of shallow slopes and medium bias on judgement estimation based on the findings of Maki (2007a, 2007b), Koriat and Bjork (2005), Korait et al. (2006), and Nelson et al. (2005) described earlier. In both models, a band of scores (akin to a confidence interval) was used for sensitivity and bias instead of a point estimate, which is line with the APA task force emphasis on confidence intervals (Wilkinson & The APA Task Force, 1999). Participant scores were then overlaid and matched to each model, outlining the hypothesised consistency of judgements and how OOM can be applied to data conventionally analysed with NHST.

DATA ANALYSIS APPROACH

Dependent variable calculations

For each experiment listed here, the judgement of associative memory slope and intercept was calculated separately for each participant as our dependent variable. All experiments were scaled to a 100-point scale, where 0 points would indicate no associative relation, and 100 points would indicate complete associative relation. Participants rated word-association strength using a 10-point Likert type scale. The scale included markers indicating that 0 = 0–9 people, 1 = 10–19 people, etc. (described later). These single digit ratings were then multiplied by 10, and the mean value of the range (4.5 points) was added to simulate a 100-point scale (as described in Maki, 2007a, Exp. 1b). The Nelson et al. (2004) free association database was our reference to determine the normed associative value for each word pair. Several other context-based relation variables are available in the psycholinguistic literature, such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), BEAGLE (Jones, Kintsch, & Mewhort, 2006), and TOPICS (Griffiths, Steyvers, & Tenenbaum, 2007). However, we believe that these variables are inappropriate for dependent variable calculation for several reasons. First, while free association and a context analysis of text are clearly related, Maki and Buchanan (2008) have shown that LSA, BEAGLE, and TOPICS can be more accurately defined as thematic variables that assess the complex relations processed in written text. Second, all experimental stimuli presented in this paper were created using the Nelson et al. free association database. Therefore, it is only logical to use these norms as the yardstick to measure participant performance. Our instructions for experiments also clarifies that the forward strength values from the free association norms are more applicable. As shown in Appendix A, participants are given a miniature free association task as part of the experiment. Then they are shown how to rate word pairs for their associative strength, in which participants guess at how a free association task would turn out if 100 college students were polled. Although thematic variables likely predict participant scores on these tasks, participants are rating word pairs without the context of other information; thus, we find the Nelson norms to be best suited for our calculations.

These values were then used to predict participant judgements of the word pairs, in a simple linear regression. The unstandardised B and intercept values were used as dependent variables in each experiment. Maki (2007b) describes slope values (*B*) as the sensitivity of a participant to the distinction between low and high frequency relations, and the (*a*) intercept value as the bias of the participant when making ratings.

Previous analyses on these experimental data sets were conducted using the free association norms to predict participant judgement (regression and NHST) in order to examine (1) if participants can rate word pairs better than chance (i.e., slope of zero), (2) relationship of semantic and associative judgements in the same study and database predictive ability of those judgements, and (3) effects of instructional changes on associative judgements. Here, we've grouped the experiments into the same categories for clarity, but changed the focus of analyses into understanding the pattern of associative judgements *across* all these studies. Later, we outline our statistical technique and how it may be applied to our data for several purposes. One, we believe that OOM is simplistic and that would make it appealing to many researchers. Two, instead of analysing if participant ability is greater than chance, we expect to find consistency across these experimental manipulations, indicating a stable model of expected judgement ability.

Observation oriented modeling

Observation oriented modeling (OOM) allows researchers to analyse experimental data at the individual participant level and define expected outcomes for multiple variables at once. Here, we have combined the slope and intercept values for participants to examine if these match a poor judgement process or, alternatively, a strong judgement process. To clarify the use of the modelling software, we have included an example guide, which demonstrates how the results from Experiment 1 were obtained (Appendix B).

In order to assess how our participants' slopes and intercepts matched our expected prediction, we must first enter and define the observed data. As per program instructions, two observations should be found in each range, and data ranges should be combined if cell counts are low. Therefore, after examining datasets and consulting the studies mentioned in the introduction, slope and intercept values were broken into five intervals: no sensitivity (negative slopes to 0.19), low sensitivity (B = 0.20–0.39), medium sensitivity (B = 0.40–0.59), high sensitivity (B = 0.60–0.79), and perfect sensitivity (B = 0.80–1.00); no bias (intercepts of 0.00–19.99), low bias (a = 20.00–39.99), medium bias (a = 40.00–59.99), high bias (a = 60.00–79.99), and very high bias (a = 80.00–100.00) (see Appendix B). These ranges also give the opportunity for participants' observations to vary and still be representative of their overall capabilities.

Next, a Concatenated Observations Analysis was examined to test each JAM study's slope and intercept values. This analysis was chosen for its ability to define and test a hypothesised pattern of slopes and intercepts, using previous research as a guide for this confidence interval (Maki, 2007a, 2007b; Nelson et al., 2005; Appendix B shows how these values are combined). Additionally, the concatenated observation analysis provides a distinct advantage over traditional NHST. In NHST, the slopes and intercepts could each be individually tested for significance from zero, indicating some form of bias or sensitivity in judgements. In OOM, the two components of the JAM function can be combined and analysed together to see how participants match a medium bias and low sensitivity model (as seen previously) or alternatively, a no bias and high sensitivity model (perfect judgements). Two different models of JAM slope–intercept combinations were defined. First, a low sensitivity model was analysed with a pattern of low slopes, from 0.20 to 0.39, and medium intercepts, from 40.00 to 59.99. Then, a perfect sensitivity model was tested with high slopes, from 0.80 to 1.00, and low intercepts, from 0.00 to 19.99, were defined (see Appendix B). For both of these models, a randomisation test was requested which allows the researcher to receive a c-value for the analysis (indicating observed data uniqueness).

These two analyses produced two sets of results, which included matches, complete matches, and c-values. To examine specified cause, both matches and complete matches are considered. Matches are the proportion of observations that align with our designated pattern on at least one dimension (for this analysis, either the specified slope or intercept). Alternatively, complete matches are the proportion of observations that match the designated pattern on both dimensions. C-values are also integral to understanding results. These values are obtained by a process much like bootstrapping but instead of selecting data points from a data set with replacement, the information in one's dataset is taken and observations are randomised. These randomised datasets are then compared to the hypothesised model. If the randomised datasets fit the pattern more often than the actual data, then c-values will be high, thus showing that the hypothesised model fit was not unique. As with traditional p-values, low c-values are desirable, though it is vital to note that c-values do not adhere to the strict cutoff scores that researchers use with p-values. All of these values can be seen in the text output screen in Appendix B. These match values can then be interpreted and compared directly to each other. When looking at the expected low sensitivity theoretical pattern, we can see that Experiment 1 had a match value of 47% with a c-value of 0.00 and a complete match value of 27% with a c-value of 0.00. When looking at the perfect sensitivity pattern, we can see that Experiment 1 had a match value of 0% with a c-value of 1.00 and a complete match value of 0% with a c-value of 1.00. When comparing these results, we can conclude that the low sensitivity model better fit data in the experiment than the perfect sensitivity model. Namely, individuals showed a bias in their estimation of word pair relationships (with the high intercepts) and insensitivity to the differences between word pair frequencies (low slope).

Using the concatenated observation analysis with the confidence intervals we selected, dozens of possible models could be hypothesised. The low sensitivity model in this paper was selected because of previous work from Maki (2007a, 2007b), and Nelson et al. (2005), wherein nearly all slope and intercept functions fell into this range. Therefore, the match values of our data to their previous results will indicate consistency of JAM functions across experiments, which is a focus of our work. While c-values would indicate if our hypothesis is as probable as a randomisation of the data, we have also included a perfect sensitivity model as a comparison to show hypothesised low match percentages and high c-values. The perfect sensitivity model is often used as a target benchmark (see solid lines in Figures 1, 2, 3) for judgements, as judgements would be very sensitive (high slopes) and unbiased (low intercepts). Several of the experiments presented here were designed to improve judgements into these ranges, and consequently, this model is presented for both purposes: a comparison while presenting OOM techniques and a comparison for possible judgement parameters given experimental effectiveness.

GENERAL METHOD

Participants and materials

For all experiments, participants and materials have been placed in Table 1 for comparison across studies. All studies were approved by the Institutional Review Board at the university. All stimuli were selected from the Nelson et al. (2004) free association norms.

Procedure

Participants in all studies were given packets of instructions and word pairings to associatively rate. First, participants read instructions on the definition of associative memory (see Appendix A) and read about the free association test. For example, participants were told that CATS and DOGS were related because of their common use in the phrase "it's raining cats and dogs". When all participants had finished reading the introduction, a free association test was

TABLE 1
Materials for all experiments

				Stimuli statistics	
Experiment	*N*	*N word-pairs*	*Selection criteria*	*FSG*	*BSG*
Traditional judgements 1 Traditional judgements 2 Traditional judgements 3 Traditional judgements 4	74 48 31 57	96	Selected to create four categories based on forward (FSG) and backward (BSG) strength: low-low, low-high, high-low, and high-high with 24 pairs in each FSG × BSG combination.	Low $M = 0.06$ (0.01–0.19) High M = 0.63 (0.51–0.78)	Low $M = 0.05$ (0.01 – 0.20) High M = 0.64 (0.51–0.80)
Combined judgements 5 (Foster & Buchanan, 2012)	46 Blocked 54 Mixed	100	25 cue words were selected with at least four target pairings (e.g., computer: game, keyboard, mouse, program). The four strongest target words were selected creating 100 pairs (25 cues × 4 targets each)	$M = 0.17$ (0.01–0.81)	$M = 0.11$ (0.00–0.94)
Combined judgements 6 Buchanan (2013)	20 Group 1 34 Group 2 21 Group 3 21 Group 4 42 Group 5	216 Group 1–4 120 Group 5	Selected with a wide range of associative and semantic relationships using Maki, McKinley, and Thompson's (2004) semantic dictionary norms	Group 1–4 $M = 0.12$ (0.01–0.92) Group 5 $M = 0.22$ (0.01–0.78)	Group 1–4 $M = 0.03$ (0.00–0.62) Group 5 $M = 0.01$ (0.00–0.12)
Combined judgements 7 Buchanan (2010)	154	202	Selected with a wide range of associative and semantic relationships using Maki et al.'s (2004) semantic dictionary norms	$M = 0.21$ (0.07–0.78)	$M = 0.17$ (0.00–0.80)
Instruction judgements 8	64 Control 64 Debias	96	Selected to create four categories based on FSG and BSG: low-low, low-high, high-low, and high-high with 24 pairs in each FSG × BSG combination.	Low $M = 0.06$ (0.01 – 0.19) High M = 0.63 (0.51–0.78)	Low $M = 0.05$ (0.01–0.20) High M = 0.64 (0.51–0.80)
Instruction judgements 9	139 Control 160 Debias				
Instruction judgements 10	27 Control 26 Control load 24 Debias 25 Debias load		Practice items were changed to emphasise the effect of backward strength on word-pair ratings with special instructions to ignore high backward ratings (i.e., steak-sirloin)		

Participants were recruited from the undergraduate research pool across several large universities (Texas Tech University, The University of Mississippi, and Missouri State University) with the permission of each university's respective Internal Review Board. All word stimuli were selected from the Nelson et al. (2004) free association norms.

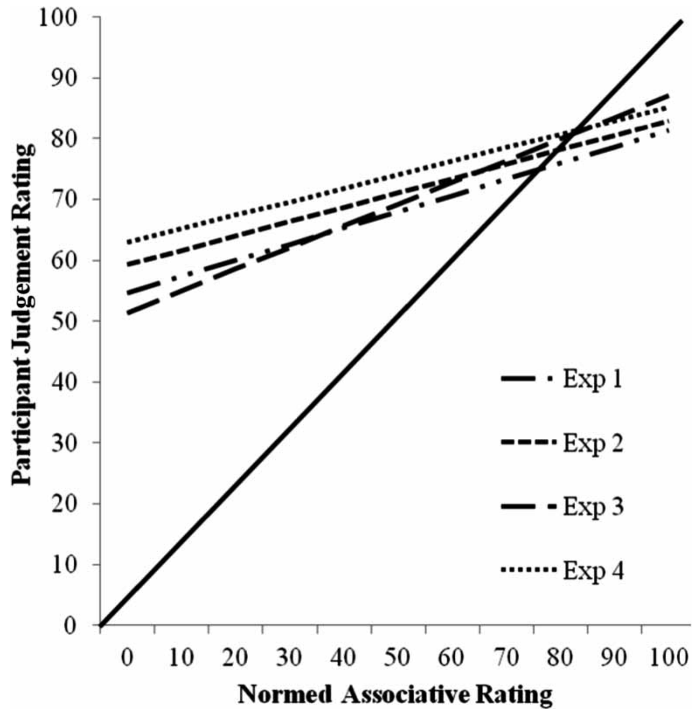

Figure 1. JAM function for traditional experiments. Solid black line would indicate perfectly attuned judgements.

presented. Participants wrote down the first words that came to mind when shown the words LOST, OLD, and ARTICLE. The experimenter then explained the most common targets associated with each cue word (i.e., FOUND, NEW/YOUNG, CLOTHING/NEWSPAPER/MAGAZINE) by asking participants to give a show of hands for each main target word. The third page in the experiment packets contained a sample-rating task. Participants were told to rate the number of people out of 100 who would list the second word in each pair if they have been given the first word in each pair. As seen in Appendix A, participants were asked to rate the number of college students who would have said APPLE if they were first shown PEAR. A 0–9 Likert type scale was shown as 0 (0–9 people), 1 (10–19 people), 2 (20–29 people), etc. After participants made sample ratings, they completed the 96 experimental pairs on their own. The combined and instructional experiments discussed after traditional judgements shared the same basic instructional set provided in Appendix A, with different experimental manipulations to stimuli or judgements. Procedures for those manipulations are listed with each experiment type.

Traditional JAM experiments

Results

Figure 1 depicts the average slopes and intercepts for each experiment. Mean and standard deviation values for each experimental slope and intercept are listed in Table 2. The average slopes for the traditional style JAM experiments ranged between $B = 0.22$ and $B = 0.36$, and the intercept values were fairly high, ranging from $a = 51.41$ to $a = 62.97$. These values match previous research in judgements of memory, and should fit a low sensitivity model with low slope values and high intercept values. Table 2 shows the match and complete match values for a perfect judgement model and a low sensitivity judgement model. Our data best fit a model with slight discernment between low and high frequency related pairs and a medium bias factor when making these judgements. None of the individual participant slope–intercept combinations fit a perfect model (even when only matching one factor at a time), and even though some variability exists in our findings, the low sensitivity model was shown to be better than a random arrangement of the dataset with very low c-values.

TABLE 2
Average slope and intercept values separated by experiment

			Matches		Complete matches	
Experiment	*Slope*	*Intercept*	*Perfect model*	*Low sensitivity*	*Perfect model*	*Low sensitivity*
Traditional judgements 1	0.27 (0.14)	54.69 (14.59)	$0.00^{1.00}$	$0.47^{0.00}$	$0.00^{1.00}$	$0.270^{0.00}$
Traditional judgements 2	0.24 (0.12)	59.29 (12.11)	$0.00^{1.00}$	$0.54^{0.00}$	$0.00^{1.00}$	$0.42^{0.00}$
Traditional judgements 3	0.36 (0.15)	51.41 (15.76)	$0.06^{1.00}$	$0.53^{0.00}$	$0.00^{1.00}$	$0.32^{0.00}$
Traditional judgements 4	0.22 (0.16)	62.97 (13.54)	$0.04^{1.00}$	$0.38^{0.00}$	$0.00^{1.00}$	$0.16^{0.00}$
Combined judgements 5 Blocked	0.10 (0.21)	57.11 (10.33)	$0.01^{1.00}$	$0.36^{0.00}$	$0.00^{1.00}$	$0.07^{0.29}$
Combined judgements 5 Mixed	0.21 (0.18)	53.56 (11.89)	$0.00^{1.00}$	$0.47^{0.00}$	$0.00^{1.00}$	$0.26^{0.00}$
Combined judgements 6 Group 1	0.26 (0.20)	37.04 (15.36)	$0.07^{0.99}$	$0.33^{0.05}$	$0.00^{1.00}$	$0.10^{0.17}$
Combined judgements 6 Group 2	0.23 (0.30)	41.93 (14.87)	$0.07^{1.00}$	$0.34^{0.01}$	$0.00^{1.00}$	$0.09^{0.14}$
Combined judgements 6 Group 3	0.37 (0.22)	48.36 (8.40)	$0.05^{1.00}$	$0.64^{0.00}$	$0.00^{1.00}$	$0.43^{0.00}$
Combined judgements 6 Group 4	0.15 (0.26)	42.63 (17.12)	$0.05^{1.00}$	$0.36^{0.01}$	$0.00^{1.00}$	$0.19^{0.01}$
Combined judgements 6 Group 5	0.14 (0.29)	40.51 (14.80)	$0.02^{1.00}$	$0.29^{0.04}$	$0.00^{1.00}$	$0.05^{0.50}$
Combined judgements 7	0.15 (0.27)	64.41 (11.60)	$0.01^{1.00}$	$0.29^{0.00}$	$0.00^{1.00}$	$0.12^{0.00}$
Instructional judgements 8 Control	0.33 (0.14)	55.20 (12.46)	$0.00^{1.00}$	$0.51^{0.00}$	$0.00^{1.00}$	$0.27^{0.00}$
Instructional judgements 8 Debias	0.30 (0.13)	47.58 (11.62)	$0.00^{1.00}$	$0.56^{0.00}$	$0.00^{1.00}$	$0.36^{0.00}$
Instructional judgements 9 Control	0.31 (0.13)	59.53 (12.24)	$0.00^{1.00}$	$0.51^{0.00}$	$0.00^{1.00}$	$0.24^{0.00}$
Instructional judgements 9 Debias	0.44 (0.17)	43.46 (12.71)	$0.02^{1.00}$	$0.40^{0.00}$	$0.01^{1.00}$	$0.18^{0.00}$
Instructional judgements 10 Control	0.30 (0.19)	50.71 (13.72)	$0.00^{1.00}$	$0.54^{0.00}$	$0.00^{1.00}$	$0.37^{0.00}$
Instructional judgements 10 Control load	0.27 (0.18)	53.70 (15.86)	$0.00^{1.00}$	$0.44^{0.00}$	$0.00^{1.00}$	$0.23^{0.00}$
Instructional judgements 10 Debias	0.51 (0.26)	34.94 (16.10)	$0.23^{0.36}$	$0.27^{0.14}$	$0.17^{0.01}$	$0.13^{0.07}$
Instructional judgements 10 Debias load	0.32 (0.23)	49.79 (14.15)	$0.04^{1.00}$	$0.40^{0.00}$	$0.00^{1.00}$	$0.16^{0.01}$

Slope values are unstandardised *B*-values, where a slope of 1 would be perfect judgement ability. Intercept values are scaled to 100, where an intercept of 0 would be perfect judgement ability. Numbers are average values across participants, with standard deviations in parentheses. C-values are denoted in superscript.

JAM combined with semantic judgements experiments

Procedure

In the following experiments, participants were given instructions about associative judgements and used the same 0- to 9-point Likert scale as described in the first set of experiments and Appendix A. In these experiments, participants were also asked to judge the semantic feature overlap between word pairs. For example, participants were given CHEDDAR and CHEESE and asked to rate what percentage of features the paired concepts shared using the same Likert scale where 0 indicated 0–9% of features, 1 designated 10–19% of features, etc. Participants were given judgements in separate blocks, which were counterbalanced for order effects. Only associative judgements are analysed for JAM functions. The different experimental manipulations on judgements are described next for consideration.

Combined judgement experiment 5. As seen in Table 1, word pairs were created so that cue words (the first word in each pair) repeated four times with different target words (second words). Participants were randomly assigned to judge word pairs in a mixed or blocked condition. In the mixed condition, participants received cue–target sets in a random order, while in the blocked condition, participants would be shown each four pair cue set together. Therefore, in the mixed condition a participant might see COMPUTER, then DARK as cue words, whereas a blocked participant would see all four COMPUTER targets before DARK was shown.

Combined judgement experiment 6. This experiment consisted of five different judgement groups (between subjects), where different aspects of semantic and associative priming were manipulated. The judgement task was paired with a rapid visual serial presentation (RSVP) task, which was used to investigate priming from a single judgement word (Buchanan, 2013). In an RSVP task, participants are shown a series of masking stimuli mixed with a target word to identify. Here, participants were first shown a word pair at the centre of the screen that they would be judging for association. After viewing this pair, the RSVP task consisted of a series of very quick (64 ms) masking stimuli ($&*@%) and a primed word mixed together. After viewing all RSVP slides, participants entered the target word from the slides. Once word entry was complete, participants were shown the original judgement pair and asked to rate the pair on the normal Likert scale. Group 3 was the only exception to this procedure, and participants rated the judgement pairs before the RSVP task and word entry. All five groups completed both semantic and associative judgements (priming data is not analysed here, group differences were based on priming manipulations).

Combined judgement experiment 7. This experiment closely mirrored traditional judgement experiments; however, judgements were made on a computer instead of pencil and paper. Participants were also instructed to make judgements as quickly as possible, while still reading both words in the pair. Participants and ratings were excluded if they were faster than the lexical decision time for both words combined using the English Lexicon Project (Balota et al., 2007).

Results

Table 1 and Figure 2 show the average slope and intercept values for these experiments. The slope ranges for these experiments was broader than for associative judgements alone, $B = 0.10$ to $B = 0.37$, which may indicate that the corresponding task manipulations (multiple cue–target pairs, RSVP, timed judgements) or addition of semantic judgements negatively affected judgements. The intercept values range from a low $a = 37.04$ bias factor to a high bias factor $a = 64.41$. Again, as shown in Table 2, participants do not fit a perfect model judgement process, in either partial or complete matches. However, in these experiments, participants only partially fit a low sensitivity-medium bias judgement model. The match values indicate that participants are either slightly sensitive or biased, and the c-values portray that these values are better than random data arrangement. However, when the slope and intercept values are combined, some of the experiments show poor complete match values. An examination of participant means from Table 1 implies that participant sensitivity values were lowered in these experiments. Blocking word pairs did not help participants in Maki's (2007b) experiments, and also similarly changed our participant ability. The RSVP task also appears to have changed participant judgements, except for Group 3 who judged pairs before the masked stimuli appeared. In fact, the only similar complete match values are for experiments that most closely mirrored

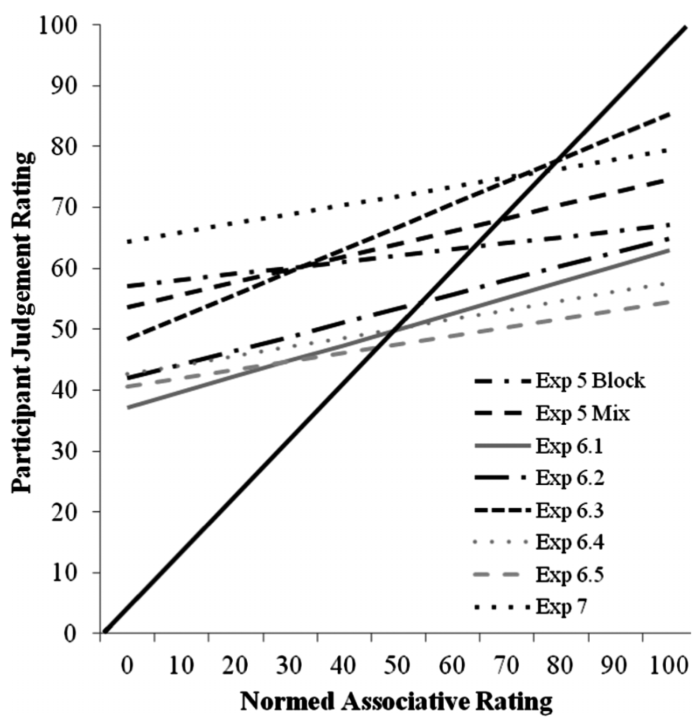

Figure 2. JAM function for combined judgement experiments. Solid black line would indicate perfectly attuned judgements.

the traditional experiments: Combined judgement experiment 5 mixed condition, Combined judgement experiment 6 Group 3 (judgement first condition), and Combined judgement experiment 7 (speeded judgement condition). These results may indicate that the addition of a semantic judgement task, speeded instructions, or priming manipulations limit or distract our ability to estimate judgements from memory.

Experiments on JAM instructions

Procedure

All three experiments included instructions described in the Traditional Judgements section with descriptions of associative relations, free association, and judgements (see Appendix A). The 0–9 Likert scale for judgements was used in this experiment set. The control groups matched the traditional experiments exactly.

Instructional judgements experiment 8. Experimental debias instructions were developed to investigate Koriat and Bjork's (2006) debiasing instructions on associative judgements. Their study was intended to reduce judgements of learning bias for pairs with deceivingly high backward strengths. Here, we informed participants that backward strength caused them to overrate word pairs in the hopes of limiting the influence of backward association. Participants read definitions of associative relations and then completed a free association task. Next, participants were shown the judgement scale and asked to complete practice judgements (as indicated in Appendix A). Debiasing instructions were given aloud after this practice session. The experimenter asked participants to give a show of hands for high judgements (7, 8, and 9) and explained the real ratings for practice pairs (which were much lower). Then the experimenter explained backward association and wrote the forward and backward associations for practice pairs on a chalkboard to demonstrate how backward association causes overestimation of judgements. For example, the pair STEAK-SIRLOIN has a very low forward relation (0 on the Likert scale) but a very high backward relation (8). Participants were told that this high backward relation influences their judgements upwards. Participants then completed the judgement pairs with a reminder at the top of each page to "ignore the backward association".

Instructional judgement experiment 9. A second version of the debiasing instructions was created, paired with Maki's (2007b) error feedback training. Participants were stepped through instructions and practice judgements as described earlier for Instructional experiment 8. The next page of the experimental packets included a chart of correct ratings with a blank for participants to fill in their practice ratings. After filling in their practice ratings, a difference score was calculated. For the pair STEAK-SIRLOIN, participants filled in their rating, saw that the rating was supposed to be 0, and then subtracted to show how their judgement was influenced upwards. The experimenter then asked for a show of hands; how many participants' ratings were too high. Participants then compared ratings with a neighbour. Packets were handed out so that every other participant had reverse practice pairings. The experimenter then explained backward association and how flipping judgement pairs changed ratings (as described earlier). Several examples were discussed while participants compared ratings. After these instructions, participants completed experimental pairs on their own with the same backward association reminder as before.

Instructional judgements experiment 10. This experiment contained two manipulations: debiasing instructions and a memory load. The instructions were exactly the same as Instructional judgements experiment 9, including error feedback and comparison with a neighbour. The memory task was added when participants were completing judgement pairs. Participants were told to remember the first word of each judgement pair and prompted to enter them every three to seven judgements (similar to a working memory task). This experiment was completed on a computer due to the addition of the memory component. Four groups were tested: regular control associative judgements, control judgements with the load task, debiasing instruction judgements, and debiasing instruction judgements with the load task.

Results

These experiments investigated the role of debiasing instructions on judgement improvement. If instructions were able to improve judgement ability, low sensitivity models would show very few match and complete match values because slopes would increase and intercepts would decrease (thus lowering our matches),

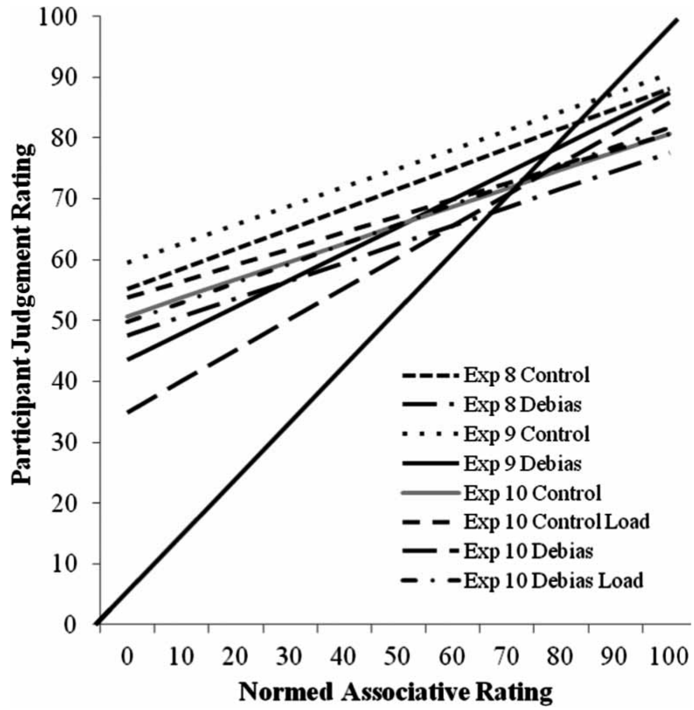

Figure 3. JAM function for instructional judgement experiments. Solid black line would indicate perfectly attuned judgements.

and increasing matches for perfect sensitivity. The load task was designed to investigate if task attention to judgement instructions required additional memory processes, and therefore any load groups should show a return to basic heuristic processing and match the low sensitivity model. Lastly, control groups did not receive any debias instruction manipulation and were expected to match the low sensitivity model.

In general, control groups showed low sensitivity for associative judgements ($B = 0.27$ to $B = 0.33$; see Table 1) with medium bias intercepts ($a = 50.71$ to $a = 59.53$). The debias instruction groups showed a mix of slope values, indicating that some of the debias instruction manipulations may have been effective at increasing slopes ($B = 0.30$ to $B = 0.51$) and lowering intercepts ($a = 34.94$ to $a = 49.79$). Further, groups with load manipulations matched a low sensitivity model as well as control groups, indicating that judgements were not improved to a perfect sensitivity model with instructions if participants were distracted by a memory task. Figure 3 shows all JAM functions for this experiment set overlaid. Table 2 shows that the low sensitivity model matched all experiments and had moderate complete match values for all experiments (13% to 37%) and those values were better than a random organisation of the dataset. Interestingly, in instructional judgement experiment 10, the experimental group showed a higher complete match proportion (17% vs. 13%) with a perfect sensitivity model, which was better than chance. The debiasing instruction groups did tend to show lower match/complete match values, which was expected given that instructions should decrease intercepts and increase slope values, especially in the third iteration of these experiments (instructional judgement experiment 10, debiasing instruction group). However, even with overt attention to an influencing factor on judgements (backward strength), participants still matched a model of shallow slopes and high intercepts, indicating that this judgement ability may be difficult to train.

GENERAL DISCUSSION

This application was prepared to uniquely examine judgement consistency across experimental manipulations, and to showcase a new data analysis technique that allowed us to inspect slopes and intercepts in combination, rather than separately by using NHST. As with Maki's (2007a) assertion, we find the slope and intercept values to be replicable and consistent across experiments and manipulations. Further, we have expanded the findings of predictability to new experimental manipulations and designs. Table 1 and Figures 1, 2, and 3 demonstrate that, although some experimental alterations can change regression values, judgements are highly insensitive and overly biased. Two models of judgement processing were examined: perfect and low sensitivity. Obviously, many different combinations of data ranges and slope–intercept combinations could have been investigated, just as many different structural equation models can be programmed with different paths. Our combinations were based on the range of prior experimental findings for the low sensitivity model, similar to confidence interval testing. The perfect sensitivity model was not theoretically expected to completely match all participant data, given that studies have shown our poor ability of metacognitive judgements. Indeed, very few participants matched these values, especially when both slope and intercept were combined for a complete match. The model was included, however, to portray that both factors of the judgement process are poorly tuned.

The consistent match to this low sensitivity-medium bias model indicates a more permanent state of the judgement process: wherein experimental design does not affect ability, judgements are stable, and while conscious management can change judgements slightly (debiasing experiments), an overload on the memory system will revert judgements back to a normal state. Participants are unable to distinguish between high and low frequency context pairings and overestimate their relationships. However, these results do not simply indicate that the underlying associative memory hierarchy is somehow defunct or completely individualised. If memory structure were not at least similar across participants, one would expect very little context-based priming and computer models of association to fail miserably. The associative boost is a well-known phenomenon, and associative priming is very robust (Buchanan, 2010; Ferrand & New, 2004; Thompson-Schill, Kurtz, & Gabrieli, 1998). Maki (2008) and Nelson et al. (2005) have both modelled associations with success.

Instead, these results indicate that conscious judgement processes cannot tap into or readily interpret unconscious memory structure. Other task processes can inhibit judgements further (memory loads and differential task demands: semantic judgements) or slightly enhance judgements (instruction), but the basic output of judgements of memory remains the same. As discussed, if participants were able to judge memory connections in a simple "low versus high" manner, the proportion of matches would have been much higher in the perfect model (i.e., participants would match on at least slope values). Instead, we believe that a better model of the JAM task is the proposed low-sensitivity model, which we were able to portray using OOM's concatenated observations analysis.

Lastly, we have demonstrated how the reliability of judgements across paradigm manipulations can be examined through OOM in a simple and easy to analyse procedure. OOM has many advantages over NHST by not focusing on assumptions or specific population parameters, but instead allows the researcher to define and support a specific hypothesis. With OOM's easy to use program and the ability to test many types of hypothesis including analyses similar to regression and ANOVA, we suggest researchers investigate its use as an alternative to the more stringent NHST that commonly appears in the literature. In the future, we hope this technique can be implemented in the social sciences to increase the focus of our studies on individuals and to develop complex integrated models that will shed more light on our various fields.

Original manuscript received March 2012
Revised manuscript received January 2013
First published online May 2013

REFERENCES

Balota, D., Yap, M., Cortese, M., Hutchison, K., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English Lexicon project. *Behavior Research Methods*, *39*, 445–459. doi:10.3758/BF03193014

Buchanan, E. (2010). Access into memory: Differences in judgments and priming for semantic and associative memory. *Journal of Scientific Psychology*, *1*, 1–8.

Buchanan, E. (2013). *Separating associations from semantics: differences in priming and judgments.* Manuscript submitted for publication.

Christopher, J., & Brannick, M. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128. doi:10.1037/a0024445

Ferrand, L., & New, B. (2004). *Semantic and associative priming in the mental lexicon*. Hauppauge, NY: Nova Science.

Foster, L., & Buchanan, E. (2012). Is insensitivity limited to one word pairing or four? *Journal of Psychological Inquiry*, *17*, 17–25.

Garskof, B., & Forrester, W. (1966). The relationships between judged similarity, judged association, and normative association. *Psychonomic Science*, *6*, 503–504.

Grice, J. W. (2011). *Observation oriented modeling: Analysis of cause in the behavioral sciences* (1st ed.). San Diego, CA: Academic Press.

Grice, J. W., Barrett, P. T., Schlimgen, L. A., & Abramson, C. I. (2012). Toward a brighter future for psychology as an observation oriented science. *Behavioral Sciences*, *2*, 1–22.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244. doi:10.1037/0033-295X.114.2.211

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*, 534–552. doi:10.1016/j.jml.2006.07.003

Kamman, R. (1968). A study of the properties of associative ratings and the role of association in word-word learning. *Journal of Experimental Psychology*, *78*, 1–16.

Koriat, A., & Bjork, R. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 187–194. doi:10.1037/0278-7393.31.2.187

Koriat, A., & Bjork, R. (2006). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 1133–1145. doi:10.1037/0278-7393.32.5.1133

Koriat, A., Fiedler, K., & Bjork, R. (2006). Inflation of conditional predictions. *Journal of Experimental Psychology*, *135*, 429–447. doi:10.1037/0096-3445.135.3.429

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240. doi:10.1037/0033-295X.104.2.211

Maki, W. S. (2007a). Judgments of associative memory. *Cognitive Psychology*, *54*, 319–353. doi:10.1016/j.cogpsych.2006.08.002

Maki, W. S. (2007b). Separating bias and sensitivity in judgments of associative memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 231–237. doi:10.1037/0278-7393.33.1.231

Maki, W. S. (2008). A database of associative strengths from the strength-sampling model: A theory-based supplement to the Nelson, McEvoy, and Schreiber word association norms. *Behavior Research Methods*, *40*, 232–235. doi:10.3758/BRM.40.1.232

Maki, W. S., & Buchanan, E. M. (2008). Latent structure in measures of associative, semantic and thematic knowledge. *Psychonomic Bulletin and Review*, *15*, 598–603. doi:10.3758/PBR.15.3.598

Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods*, *36*, 421–431. doi:10.3758/BF03195590

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below .05. *Quarterly Journal of Experimental Psychology*, *65*, 2271–2279. doi:10.1080/17470218.2012.711335

Munsterberg, H. (1908). *On the witness stand*. New York, NY: Doubleday.

Nelson, D., Dyrdal, G., & Goodmon, L. (2005). What is preexisting strength? Predicting free association, similarity ratings, and cued recall probabilities. *Psychonomic Bulletin and Review*, *12*, 711–719. doi:10.3758/BF03196762

Nelson, D., McEvoy, C., & Dennis, S. (2000). What is free association and what does it measure? *Memory and Cognition*, *28*, 887–899. doi:10.3758/BF03209337

Nelson, D., McEvoy, C., & Schreiber, T. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods: Instruments and Computers*, *36*, 402–407. doi:10.3758/BF03195588

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. doi:10.1037/0033-2909.86.3.638

Spence, D., & Owens, K. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, *19*, 317–330. doi:10.1007/BF01074363

Spies, J. R., Nosek, B. A., Bartmess, E., Lai, C., Galak, J., Cohn, M., … Kappes, H. (2012, April 1). *In Open Science Framework: Reproducibility Project*. Retrieved October 18, 2012, from http://openscienceframework.org/project/EZcUj/wiki/home

Thompson-Schill, S., Kurtz, K., & Gabrieli, J. (1998). Effects of semantic and associative relatedness on automatic priming. *Journal of Memory and Language*, *38*(4), 440–458. doi:10.1006/jmla.1997.2559

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *The American Psychologist*, *54*, 594–604. doi:10.1037/0003-066X.54.8.594

APPENDIX A. JAM INSTRUCTIONS FOR ALL EXPERIMENTS

Introduction to word association experiment

This experiment is concerned with the structure of human associative memory. This knowledge is structured in some way and the mental structure is thought to come about through a process of associative learning. For example, consider the word (and concept of) DOG. We often see the word DOG appear in the same context as the word CAT. "It's raining cats and dogs." "I have two dogs, but my neighbour has a cat." And so on. By experiencing the words CAT and DOG together many times, we develop an association (a mental connection) between them. With lots of this kind of associative learning experience during our lives, we develop a very large and very complex associative memory.

Psychologists are interested in understanding the structure of associative memory and have several ways of investigating it. One method of investigating associative memory is known as a test of "free association". In free association tests, participants like you are given a series of words and are asked to respond to each word by writing the first word that pops into mind.

**When you have finished reading this introduction,
please wait for further instructions from the experimenter.**

**DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO
BY THE EXPERIMENTER!!!**

Here is an example of a free association test. In the space provided, please write the first word that comes to your mind in response to each of the following words.

	Your responses:
LOST	
OLD	
ARTICLE	

**When you have finished writing your responses,
please wait for further instructions from the experimenter.**

**DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO
BY THE EXPERIMENTER!!!**

In this experiment you will be asked to determine how many students out of 100 would respond to a cue word with a specific target word during a free association task. Below is an example of the kind of rating scale you will be using. When you make your ratings, make them by circling the most appropriate number next to each word pair.

Sample rating form

Assume 100 college students from around the nation gave responses to each CUE word. How many of these 100 students do you think would have given the RESPONSE word?

Mark the ratings below using the following scale.

Rating	0	1	2	3	4	5	6	7	8	9
No. of people	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89	90–100

Cue	**Response**	**Circle your rating**
PEAR	APPLE	**0 1 2 3 4 5 6 7 8 9**
BANJO	GUITAR	**0 1 2 3 4 5 6 7 8 9**
TOILET	BOWL	**0 1 2 3 4 5 6 7 8 9**
CLAM	SHELL	**0 1 2 3 4 5 6 7 8 9**
TANGERINE	ORANGE	**0 1 2 3 4 5 6 7 8 9**
SHIELD	SWORD	**0 1 2 3 4 5 6 7 8 9**
NYLONS	HOSE	**0 1 2 3 4 5 6 7 8 9**
PISTOL	GUN	**0 1 2 3 4 5 6 7 8 9**

When you have finished writing your responses, please wait for further instructions from the experimenter.

DO NOT TURN THIS PAGE UNTIL TOLD TO DO SO BY THE EXPERIMENTER!!!

APPENDIX B. EXAMPLE ANALYSIS USING OBSERVATION ORIENTED MODELING SOFTWARE

Observation oriented modeling: How to run a concatenated observations analysis

Data imputation. To begin working with the OOM software we must first import the data we will be using. Much like SPSS, the basic Data View window of OOM shows observations as rows and variables as columns. You can either copy and paste in data from Excel or SPSS, or enter data by hand into this window.

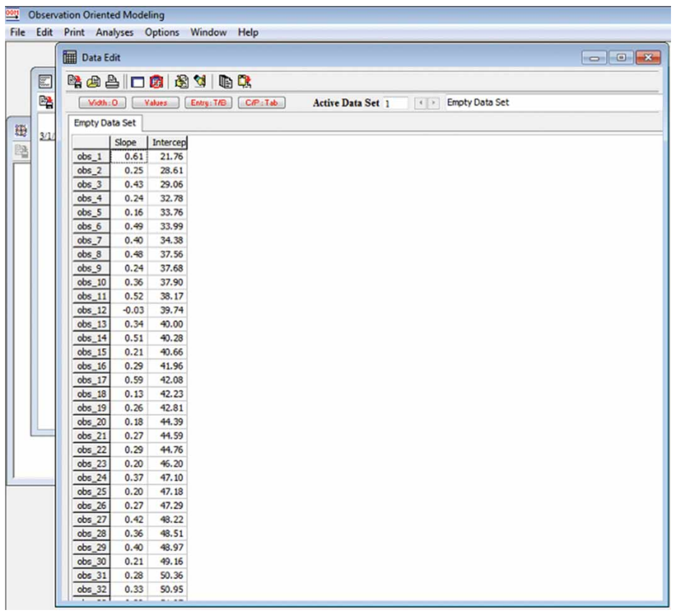

Figure 4. Data view example in OOM.

Instead of having a Variable view like SPSS has, OOM has an option that allows you to define your variables. By clicking on the Define Ordered Observations icon (see Figure 5) or selecting Edit → Define Ordered Observations, you will open a window that will allow you to name your variables (which you can do by copying and pasting the labels from the variable view of SPSS or by typing in the labels).

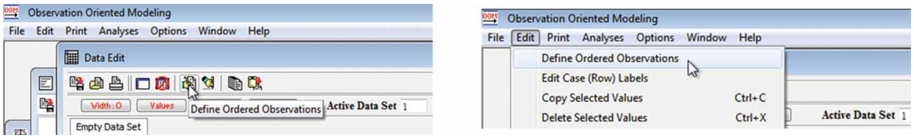

Figure 5. How to define variables for use in OOM.

Here we have defined our variables of slope and intercept in the form of ranges as follows (Figure 6):

- *Slope*: No sensitivity (negative slopes to 0.19), low sensitivity (0.20–0.39), medium sensitivity (0.40–0.59), high sensitivity (0.60–0.79), and perfect sensitivity (0.80–1.00)
- *Intercept*: No bias (intercepts of 0.00–19.99), low bias (20.00–39.99), medium bias (40.00–59.99), high bias (60.00–79.99), and very high bias (80.00–100.00).

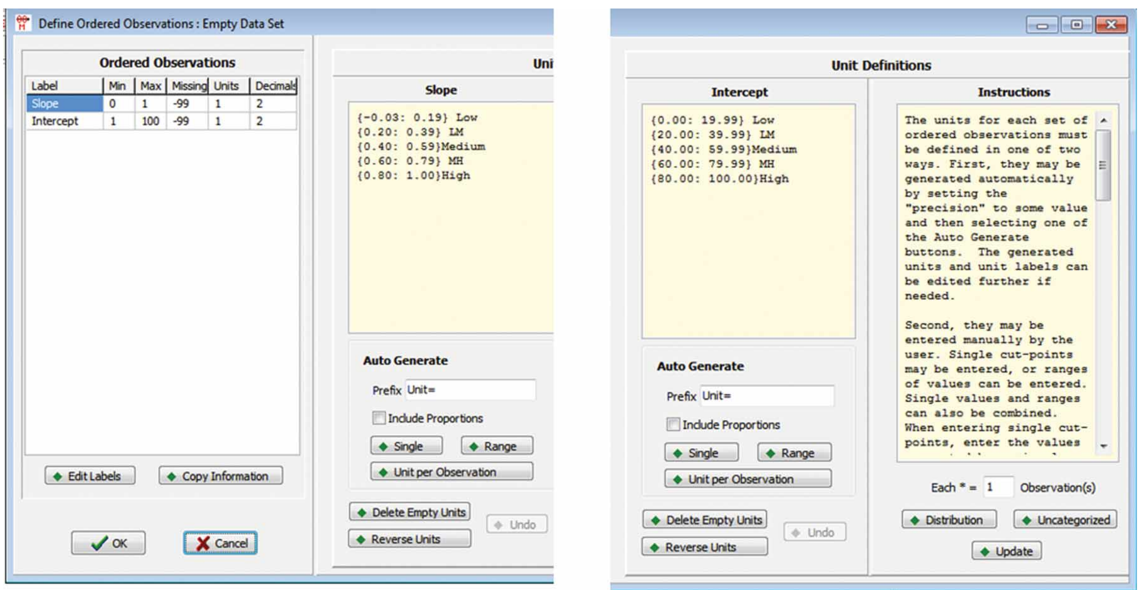

Figure 6. Variable range definition for slope and intercept.

By clicking on the Distribution button you will be able to see how your observations are distributed as you have defined them (similar to a histogram) (Figure 7). By clicking on the Uncategorized button you will be able to see what observations, if any, have not been able to be categorised by your definitions. There should be no uncategorised data, and ideally at least three observations in each range/option.

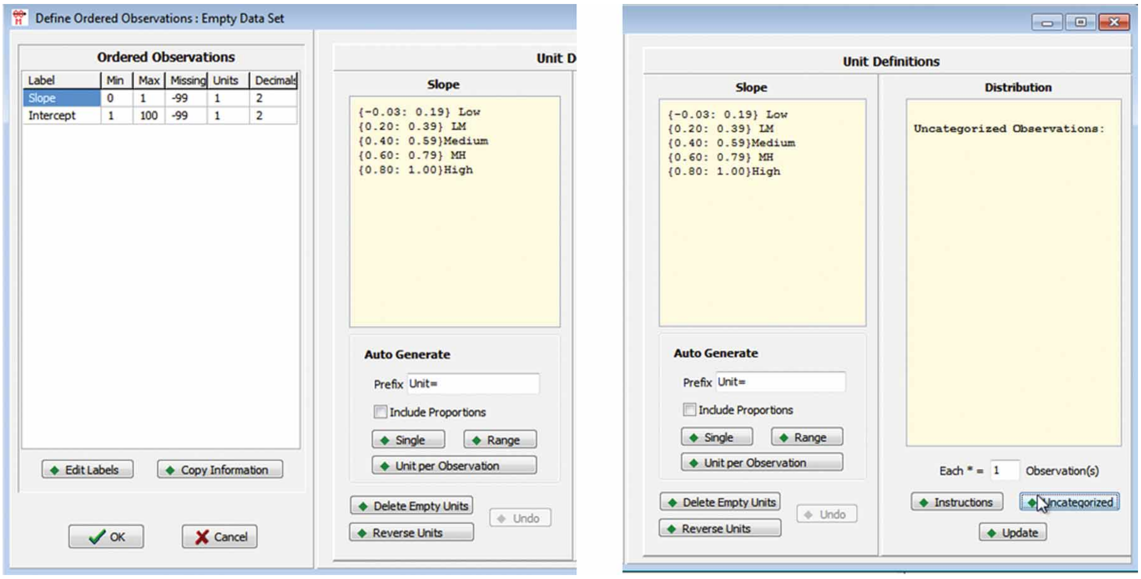

Figure 7. On the left, the distribution of observations for slope, and on the right, uncategorised observations for the slope.

Data analysis. Once your data has been defined, you can now consider which analysis you would like to perform. For the purposes of this paper we have used only the Concatenated Observations Analysis. To perform this analysis you can select Edit → Pattern Analysis → Concatenate Observations (Figure 8).

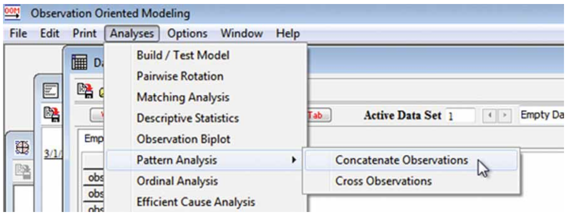

Figure 8. Choosing the Concatenate Observations analysis from the drop-down menu.

The Concatenate Observation window will open after this drop-down selection. In this window, all of your observations will appear in the list on the left (Figure 9, left). By selecting the variables for your analysis, you can move them over to the list on the right with the arrow button (Figure 9, right).

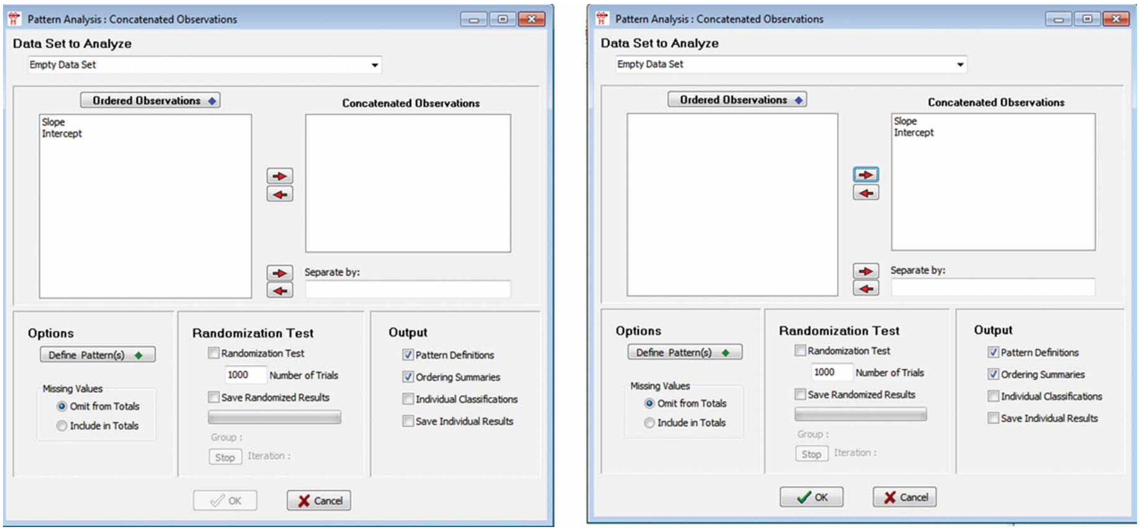

Figure 9. On the left, the Concatenate Observations analysis window, and on the right, the variables being moved over and selected for the analysis.

To create your hypothesised model, click the Define Pattern button. By clicking in the appropriate boxes for your variable combinations, you will define the pattern that you expect to see in your data. For Experiment 1 we first chose a low sensitivity (slope) and medium bias (intercept) pattern to test. By clicking these boxes they become green, designating that they are to be compared to each observation (Figure 10). If you happen to click on a box in error, right-clicking on the same box will deselect the box.

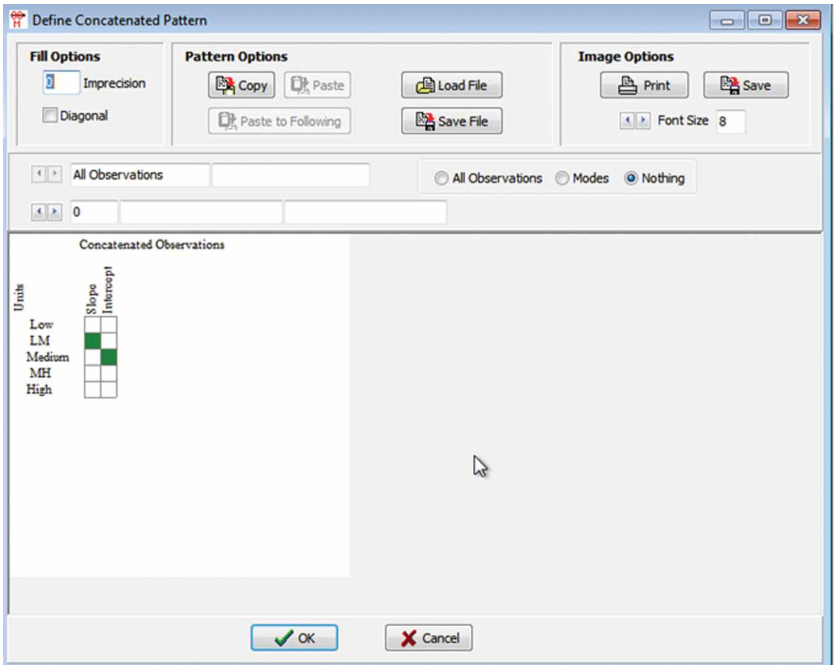

Figure 10. Pattern definition for the analysis with a low-medium slope value and a medium intercept value.

Then we will click OK. The next step is to select the box under Randomization Test titled Randomization Test. Ensure that the Trials box contains the number 1000 (Figure 11). To run the Concatenated Observation Analysis, click OK.

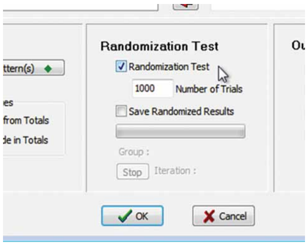

Figure 11. Selecting a randomisation test.

Output. The Text Output window will open and will contain text relating to the analysis (Figure 12). The output displays the observed proportion of matches (those individuals who match on either the designated slope or the designated intercept) and its related c-value (the proportion of times that the data in the data set, if randomised, met or exceeded that proportion of matches), and the observed proportion of complete matches (those individuals who match on both the designated slope and the designated intercept) and its related c-value.

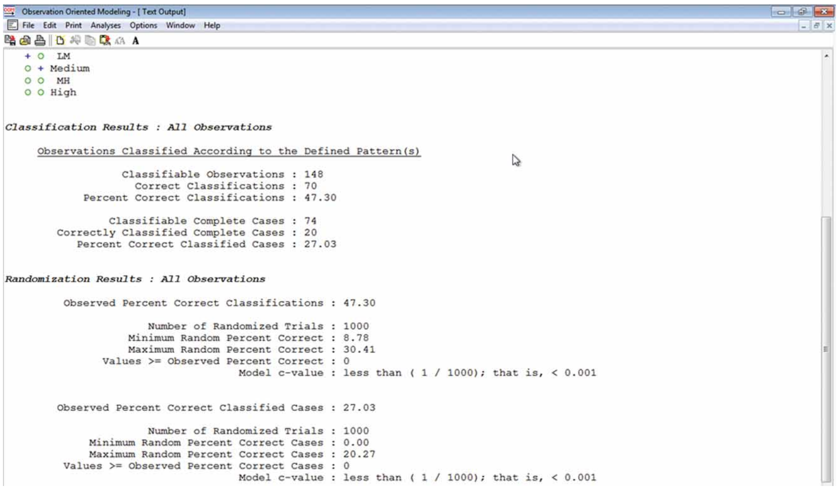

Figure 12. Concatenated Observations analysis output from the first expected pattern in the output window.

To analyse our comparison perfect sensitivity model, we will redefine the expected pattern under define pattern. This model was defined as pattern of high sensitivity (slope) and no bias (intercept) (Figure 13).

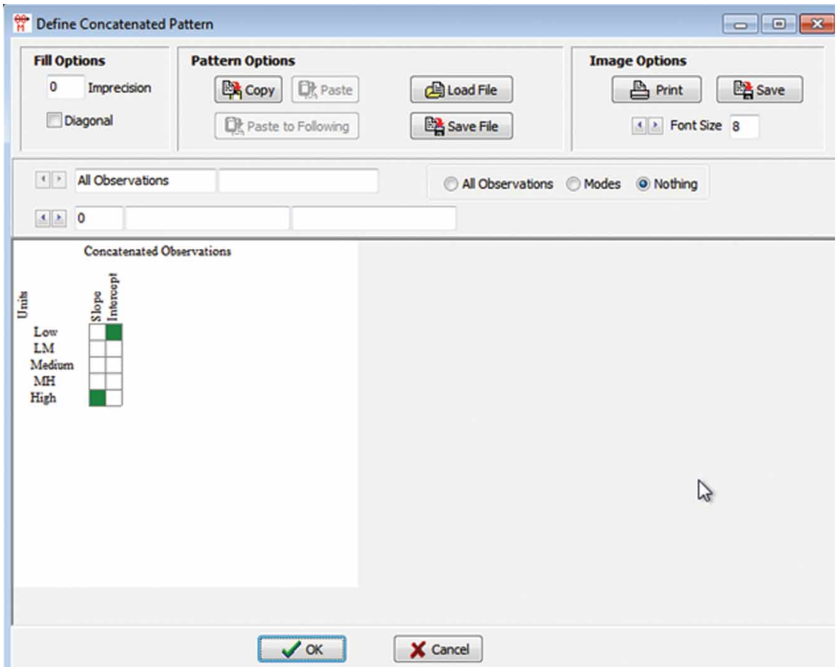

Figure 13. Pattern definition for the analysis with a high slope value and a low intercept value.

Again, be sure that the Randomization Test is selected and the Trials box contains the number 1000. Click OK to run the analysis and the Text Output window should again open (Figure 14).

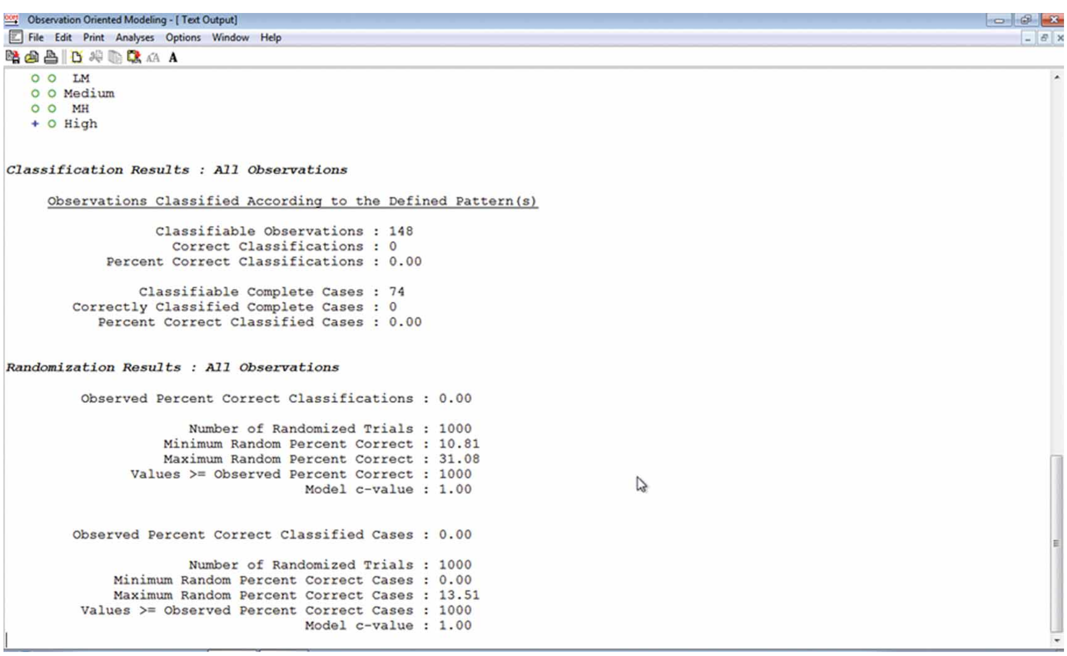

Figure 14. Concatenated Observations analysis output from the second expected pattern in the output window.

Again, we will see the Observed proportion of matches and complete matches, as well as their corresponding c-values. We can now compare the match outcomes with these outputs and determine which pattern fits our observed data. Because the low sensitivity model had a higher proportion of match and complete matches and low c-values, we can conlclude that our observations reflect more of a low sensitivity (slope) and medium bias (intercept) pattern than the perfect expected pattern.