RICE UNIVERSITY

Visual Displays: The Continuing Investigations of the Highlighting Paradox

by

Frank Tamborello

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE

Master of Arts

APPROVED, THESIS COMMITTEE:

Michael D. Byrne
Assistant Professor, Psychology

David M. Lane
Associate Professor,
Psychology, Statistics, and Management

James L. Dannemiller
Lynette S. Autrey Professor, Psychology

HOUSTON, TEXAS
MAY 2006

ABSTRACT

Visual Displays: The Continuing Investigations of the Highlighting Paradox

by

Frank Tamborello

Previous research has suggested that making certain items visually salient, or highlighting, can speed performance in a visual search task. But designers of interfaces cannot always easily anticipate a user's target, and highlighting items other than the target can be associated with performance decrements. Three experiments were performed which demonstrated that people's performance in a visual search task is differentially sensitive to highlighting's predictiveness of target location. That sensitivity depends upon the proportion of instances in which highlighting actually predicts target location. A cognitive model constructed using the ACT-R architecture inferred that people evaluate and adjust their visual search behavior at a very small level of the task.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

Certain properties of our visual processing system make it so that certain features, if easily distinguishable from the rest of the visual field, can be especially salient. This visual salience can be harnessed in the form of highlighting to make the visual search component of an information search task more efficient. Yet it is not always easy for the designers of visual interfaces to anticipate the particular target a user may be searching for in a particular context, and it is not clear what the detriments of invalid highlighting may be. The research proposed in this paper aims to examine issues of highlighting validity's effects on a visual search task, and how people respond to varying levels of highlighting validity.

Most investigators agree that certain fundamental features of visual stimuli (such as color, brightness, and movement) are processed in parallel relatively early in the visual pathway of humans (Treisman & Gelade, 1980; Wolfe, 1994). Given a visual search task, the item that can be distinguished by one of those basic features tends to "pop out" from the field of other stimuli. For example, when searching through a field of green T's, the time to find a red T remains roughly the same no matter how many distractor green T's are present. If, however, the target item can only be distinguished by a conjunction of features (such as a certain color and brightness combination), then visual search will be slower and more effortful.

Wolfe's Guided Search 2.0 (1994) theory of visual attention postulates a bottom-up process which filters stimuli through broadly-tuned "categorical" channels. Each of these channels processes a certain dimension of visual stimuli, such as color or orientation. The output of these channels, based on local differences in the stimuli, is then integrated with top-down task demands to compute a feature map for each stimulus dimension. The top-down commands to the feature maps activate locations possessing specific categorical attributes relevant to task demands, such as “activate ‘red’ objects.” The weighted sum of these feature maps forms the activation map, with the weightings being based upon task demands.

Attention deploys limited capacity resources to locations in order of decreasing activation. Visual salience, then, is this combination of bottom-up local difference calculations and top-down, task demand-driven commands. Wolfe actually instantiated his Guided Search 2.0 theory as a computational model of multiple types of visual search tasks. And although Guided Search 2.0 does take into account “top-down” guidance of visual search, it says nothing about learning. That is, how might a human learn about the relative helpfulness of visual cues, and how might that learning be reflected in human performance?

Highlighting by color can be an effective means to harness the computational power of the visual system to aid visual search. Fisher and Tan (1989) performed two experiments assessing the effects of highlighting types and validity on search times. Subjects searched for a target digit in a horizontal array of five digits, one of which was the target. The target was always the digit 1, 2, 3, or 4 (chosen at random), and the distractors were always the digits 5 through 9. Experiment 1 had four highlighting conditions: control (no highlighting), highlighting by color, by reverse video, and by blinking. Additionally, when highlighting was present, the target was highlighted on 50% of the trials, and one of the distractors was highlighted the other 50% of the time. The term “validity” will henceforth be used to describe the proportion of trials on which the highlighting correctly indicated the target. Their experiment 2 was identical, except that highlighting was 100% valid.

Fisher and Tan reported that highlighting by the most effective means, color, did not help participants find the target faster than they found it in control trials when highlighting validity was at 50% (highlighting by other means made subjects slower). Yet when highlighting was 100% valid, subjects did find the target digit 192 ms faster in the color condition than in the control condition. Furthermore, when highlighting validity was 100%, subjects were 90 ms faster on trials with valid color highlighting than they were on trials with valid color highlighting at 50% validity. The authors speculated that the difference in performance occurred because subjects did not always initially attend to the highlighted digit in Experiment 1, but did so in Experiment 2 when they saw that highlighting was more predictive of the items' status as target or distractor. Apparently participants had been making estimates of the relative costs of attending to the highlighting or disregarding it. It was not clear from Fisher and Tan's results what would occur at other levels of validity and so an experiment replicated and extended their results.

Given a paradigm like Fisher and Tan's digit search task, people operate in an environment with probabilistic cues. Shanks, Tunney and McCarthy (2002) described probability matching as a choice anomaly that is heavily context-dependent. What they meant is that probability matching is a case where people deviate from optimal models of choice in the decisions that they make, and these deviations depend upon the probability structures in their environment. People tend to match the probabilities of responding one way or another to the probabilities of cues in their environment. Though people are strongly predisposed toward probability matching behavior, given enough compulsion they can be induced to behave optimally. What Shanks, Tunney and McCarthy showed was that both outcome feedback and substantial reward are necessary to compel people to behave optimally. Sampling, even at a detriment to payoff outcome, seems to be a strongly-ingrained behavior.

In a task like Fisher and Tan's digit search task, subjects receive minimal feedback and reward. Assuming a correct response, the only feedback and reward they receive for attending to the highlighting when it is valid or ignoring it when it is invalid is a faster time to complete the trial. Even so, slow trial completion times average approximately 1,200 ms while fast trial completion times average approximately 500 ms. The 700 ms that can be saved if the subject finds the target on his or her first shift of visual attention may not seem like much incentive, if it is even noticed by the subject at all. If the faster trial completion time is not noticed by the subject, then there is no feedback for response time at all because there is no overt feedback on trials in which the subject issues a correct response.

There are several questions concerning the helpfulness of highlighting that could be explored. How sensitive are people to highlighting validity? To what extent can people adapt to changing validity levels? What search strategies do people employ, and how do the answers to the previous two questions come to bear on the development and adoption of those strategies? Finally, it would be informative to have a model of visual search with highlighted displays that can make predictions about human performance. The research proposed in this thesis seeks to answer these questions.

EXPERIMENT 1A

Fisher and Tan’s (1989) study only implemented two levels of highlighting validity, 50% and 100%. It was useful to see that, overall, 50% validity offered no improvement in visual search performance over no highlighting. But it would be more useful to derive a predictive function of validity’s effect. We therefore wanted to test validity’s effect at more levels of validity such that we might derive such a predictive tool. Experiment 1A is essentially a replication of Fisher and Tan’s paradigm, except that only color highlighting was examined and it was examined at additional levels of validity. The effects of position will not be a major focus since Fisher and Tan found, unsurprisingly, that subjects were fastest for targets in the middle and roughly equally slow on both ends. Bear in mind that subjects began each trial with a center fixation.

Method

Participants

One hundred Rice University undergraduates (62% female) participated to fulfill experiment participation requirements for their psychology classes. The participants had a mean age of 19.2 years (1.3). There were 20 subjects per condition.

Design

Experiment 1A used a mixed design, with the same within-subject variable, trial type (standard, valid highlighting, invalid highlighting) as in Fisher and Tan’s study. Additionally, subjects were assigned to one of five validity proportion conditions: 50%, 67.5%, 75%, 82.5%, or 100% highlighting validity.

Procedure

The subjects' task was to, as quickly as possible without making any mistakes, find the number in the display that was less than five and immediately press the corresponding key on the number row at the top of the keyboard. At the start of the trial, subjects viewed a crosshair for 500 ms at the intended fixation point, in the center of the computer screen. They subsequently viewed a horizontal array of five different numerals. The numerals were printed in 14-point Times New Roman font on a 17-inch CRT computer monitor at a resolution of 1024 by 768 pixels. One numeral from the potential target set, {1 2 3 4} was chosen at random, while four distractors from the distractor set {5 6 7 8 9} were also chosen at random. The target and distractors were sorted randomly. The array would disappear upon the subject's key press, and one second later the next trial would begin. In the event of an incorrect response, the computer beeped and paused the experiment for two seconds. This time penalty discouraged simple guessing.

Results and Discussion

Experiment 1B was an extension of 1A, so results and discussion for both experiments will be presented together.

EXPERIMENT 1B

Preliminary results from Experiment 1A made it desirable to test additional levels of validity such that the entire range of possibilities of 0% to 100% might be addressed by a model. Therefore, Experiment 1B replicated 1A except that it used validity levels of 0%, 12.5%, 25%, and 37.5%.

Method

Participants

Eighty Rice University undergraduates (51% of whom were female) participated to fulfill experiment participation requirements for their psychology classes. The participants had a mean age of 19.2 years (1.3). There were 20 subjects per condition.

Results and Discussion

Outliers were removed prior to statistical analysis, and this was done both for single trials and entire subjects. An outlier trial was defined as a trial in which the response time was more than three standard deviations from the subject's overall mean. Those response times were replaced with the subject's mean response time. Each subject's mean response time per condition was similarly screened against the total mean response time for all subjects, per condition. Any subject whose mean response time was more than three standard deviations from the total mean response time for all subjects in more than one condition was considered an outlier subject. Four such subjects were found, three from experiment 1A and one from experiment 1B. Their data were removed from further analysis.

Figure 1 summarizes the means for each validity condition by trial type. The mean response time for the 100% validity condition (640 ms) was 86 ms faster than the mean of mean response times for all other conditions (726 ms). ANOVA with contrast revealed that this effect was reliable, $F(1, 167) = 12.51$, $p = 0.001$. It is worth mentioning that subjects in the 0% validity condition were not reliably slower than subjects in the other conditions, contrast *F*(1, 167) = 0.07, *p* = 0.799. Excluding the 0% and 100% validity conditions, omnibus ANOVA yielded no effect of validity condition, *F*(6, 129) = 1.24, *p* = 0.292.

Subjects responded to trials with valid highlighting on average 51 ms faster than trials with no highlighting, for which they were 75 ms faster than trials with invalid highlighting (linear *F*(1, 129) = 418.32, *p* < 0.001). Furthermore, there was an interaction of trial type and validity condition such that as validity increased, subjects became faster on valid trials and slower on invalid trials, *F*(12, 258) = 16.22, *p* < 0.001 (see Figure 1). But there was no main effect of validity condition on response times for the standard trial type, *F*(6, 129) = 1.14, *p* = 0.341. One framework for understanding this validity by trial type interaction is the notion of sensitivity, as reflected in response times, to the valid and invalid trial types in the context of the validity proportion. Since subjects in the 100% validity condition received no invalid trials, and vice versa for the 0% validity condition, further analysis will exclude these two conditions.

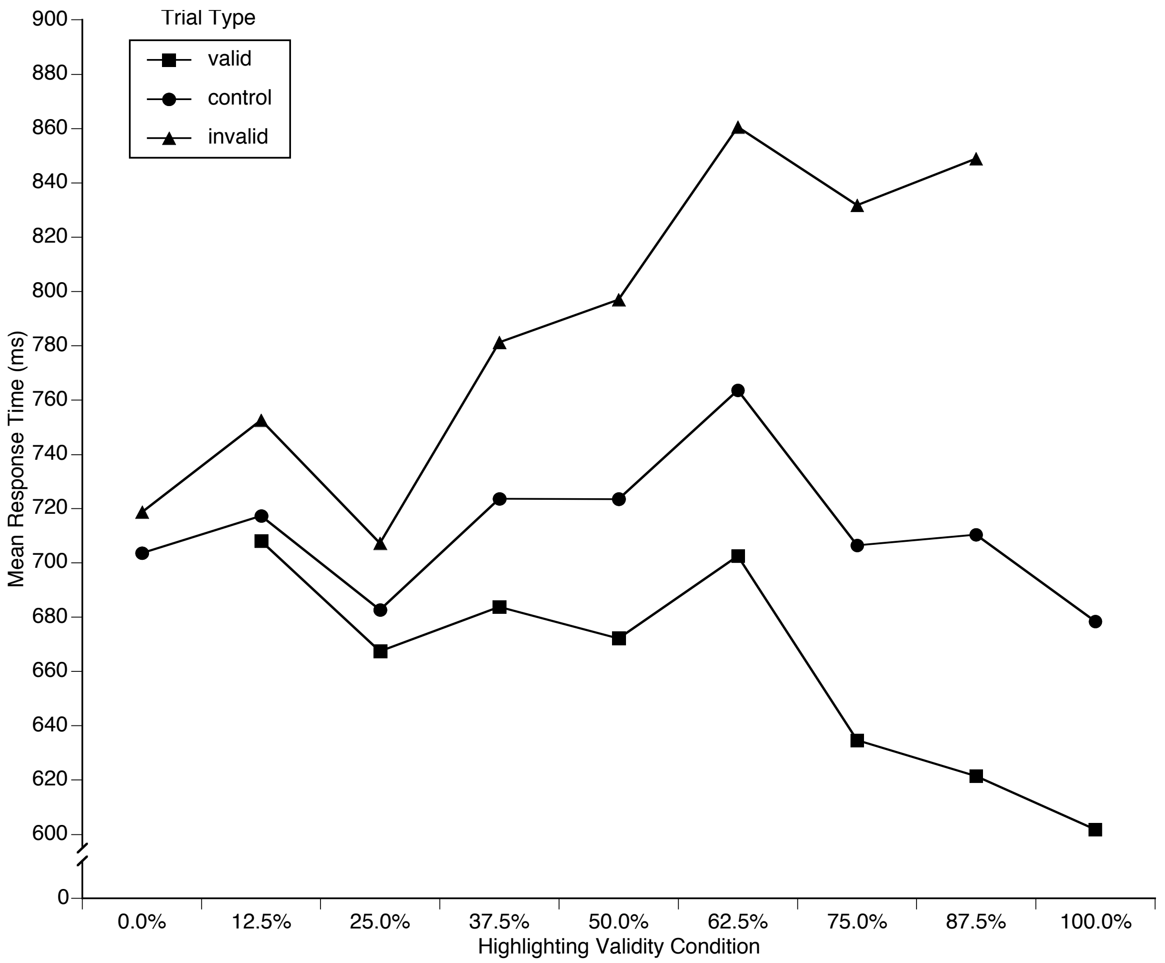


Figure 1. Mean response time as a function of validity condition.

Subjects were differentially sensitive to highlighting validity, depending upon which validity condition they were in. Sensitivity was defined as the response time for invalid trials minus the response time for valid trials. Note in Figure 1 how the lines for valid and invalid trials diverge going from the 0% validity condition to the 100% validity condition. Furthermore, participants sensitized within the first block of 64 trials. Figure 2 plots the effect of block and validity condition on sensitivity. Within block 1, subjects in the 87.5% validity condition were 201 ms more sensitive to trial validity than were subjects in the 12.5% condition, an over four-fold difference. One-way ANOVA revealed an effect of validity condition on sensitivity, *F* (6, 129) = 7.91, *p* < 0.001. Furthermore, linear contrast indicated an increasing effect of validity condition such that subjects were more sensitive to trial type the more often they encountered validly highlighted trials, *F*(1, 129) = 45.362, *p* < 0.001. Furthermore, this trend held for all trial blocks, linear *F* (1, 129) = 54.06, *p* < 0.001, though sensitivity did decline overall from block one to block six, linear *F* (1, 129) = 12.90, *p* < 0.001.

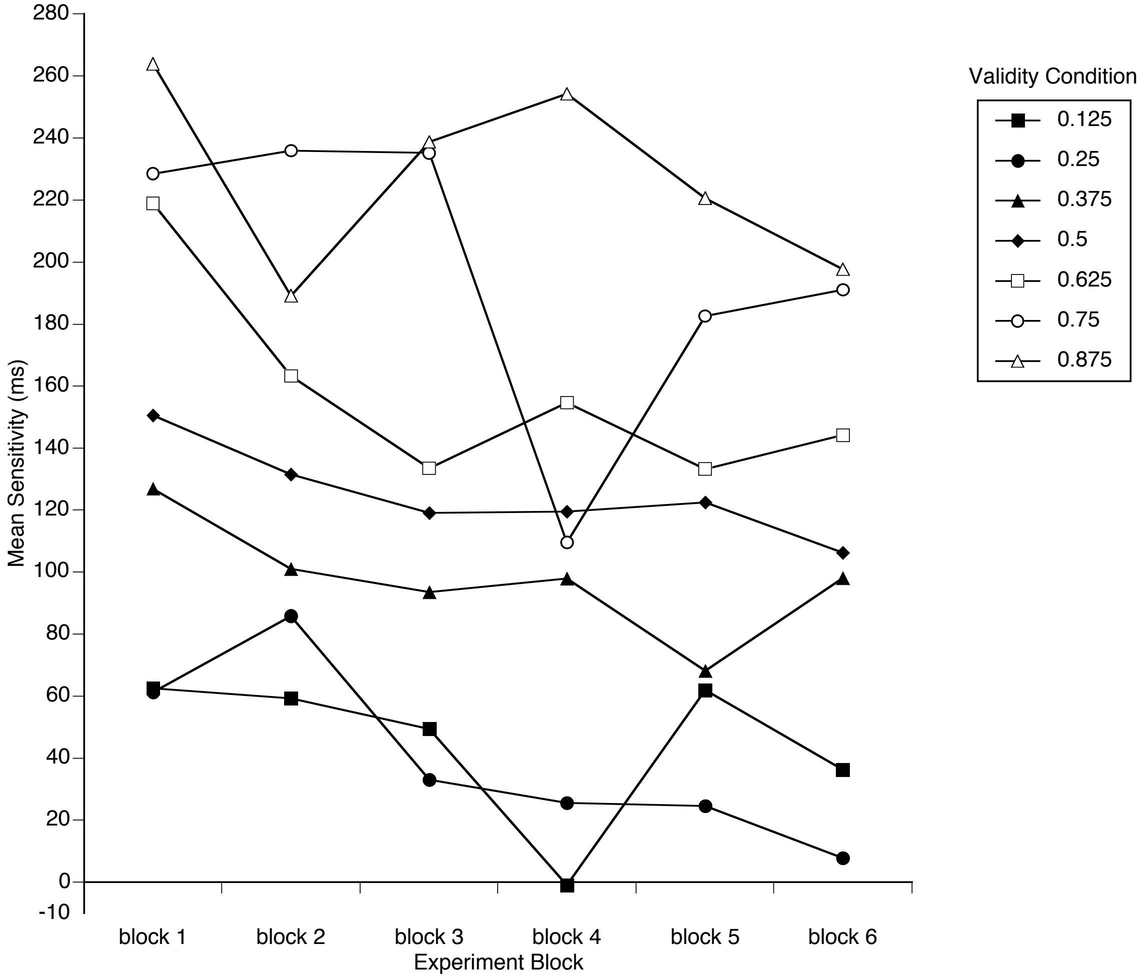


Figure 2. Sensitivity as a function of block and validity condition.

Subjects became faster overall with time. The mean response time for block 6 (692 ms) was 65 ms faster than the mean response time for block 1 (757 ms). ANOVA with contrast yielded a linear effect of block, *F* (1, 129) = 107.64, *p* < 0.001. But the block by validity condition interaction was not reliable, *F*(30, 645) = 1.46, *p* = 0.056. Thus there is insufficient evidence to support the notion that subjects improved their performance at different rates in different validity conditions.

Collapsing across validity conditions, response times for the valid trial type showed very little practice effect (Figure 3). The control trial type showed some practice effect, while the invalid trial type showed the most substantial practice effect of all three trial types. Note that the slope of the function of block for invalid trials drops more steeply than the slope for valid trials in Figure 3. For trials with valid highlighting, the mean response time dropped by 47 ms from block 1 (698 ms) to block 6 (651 ms). Mean response times for standard trials fell by 63 ms from block 1 (758 ms) to block 6 (695 ms). And for trials with invalid highlighting, that drop was 94 ms from block 1 (856 ms) to block 6 (762 ms), linear by linear interaction effect *F*(1, 129) = 19.70, *p* < 0.001. This linear by linear interaction effect of block with trial type demonstrates that the change in sensitivity over time referred to in Figure 3 is driven mostly by a drop in response time for invalid trials.

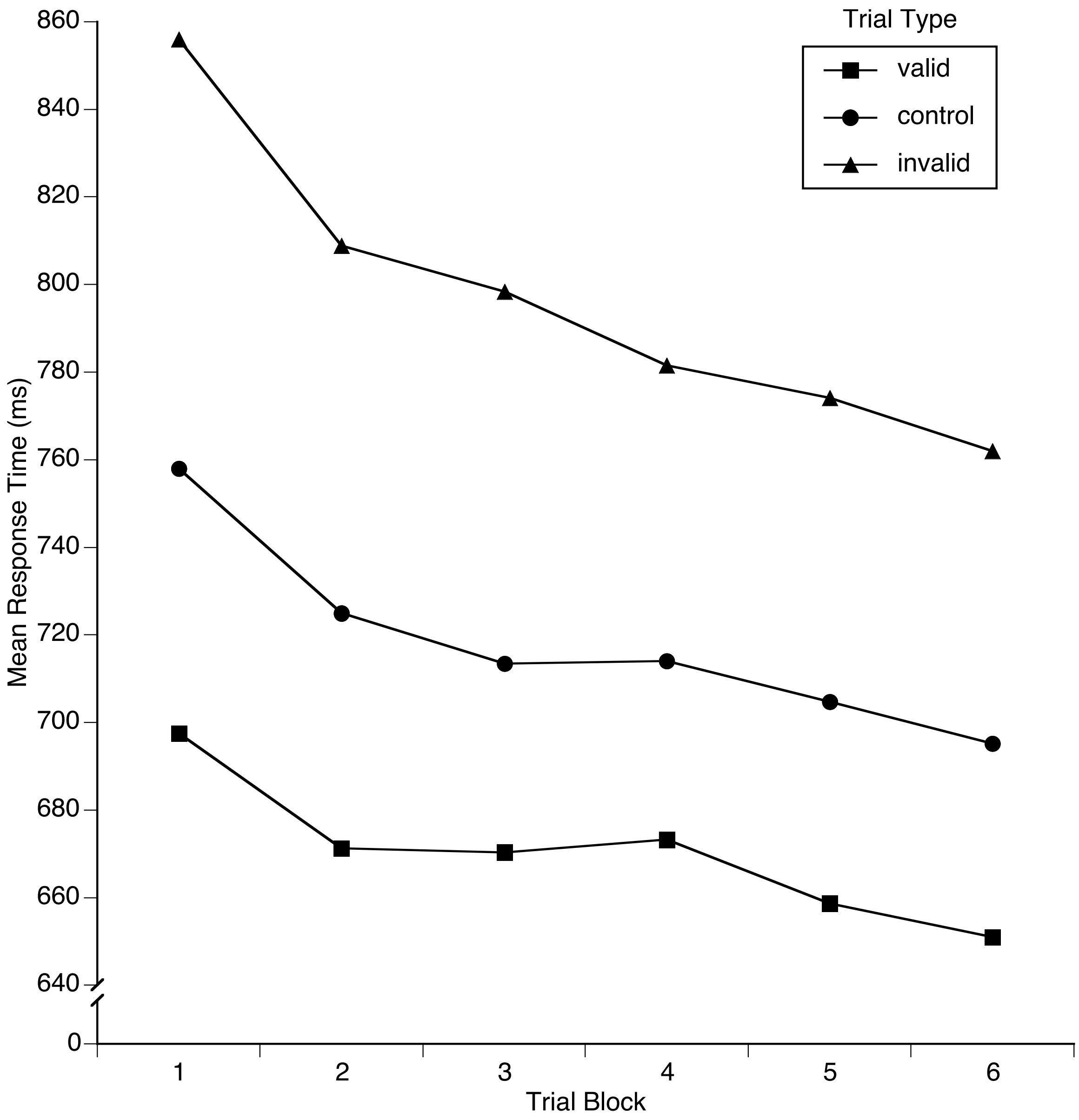


Figure 3. Block by trial type interaction.

It is interesting that the 100% validity condition was so much faster than the 87.5% validity condition. This may speak to the sensitivity issue in that if subjects were sensitive to the highlighting always being valid in 100% validity condition, then they probably used it exclusively and never engaged in a serial search for the target, instead letting their preattentive processing of color do all the work. Similarly, if performance for the valid trial types was subject to a floor effect because subjects initially relied upon preattentive visual processing for their search for the target, then there would be little room for a practice effect. A floor effect for valid trial types would explain why the practice effect was bigger in the invalid and standard conditions.

Analysis of response time for invalid trials as a function of target distance from the highlighted item shows that subjects took longer to respond as the distance between target and highlighted item increased (see Appendix A). This indicates that subjects likely processed the display serially, attending one item at a time.

The analysis of subjects' sensitivity to highlighting indicates that subjects learned fairly rapidly whether or not they could take advantage of highlighting in their visual search. The more valid the highlighting was, the more subjects used it and the faster their response times for valid trials became–and this was true even in the first block. However, as validity increased, so did the cost of using highlighting on the occasions that it was invalid. The results just presented indicate that people rapidly sensed different levels of highlighting validity, but what if that level changes? Therefore experiment 2 will be designed to assess how sensitive people may be to changes in highlighting validity.

EXPERIMENT 2

Method

With the following exceptions, Experiment 2 was identical to Experiments 1A and 1B. The main change in Experiment 2 was its focus on examining the adaptation to changing levels of highlighting validity. The validity levels therefore change during the course of the experiment, rather than being static as in Experiments 1A and 1B.

Participants

One hundred nine Rice University undergraduates participated to fulfill experiment participation requirements for their psychology classes.

Design

Experiment 2 incorporated a mixed design utilizing two within-subjects variables, block and trial type (standard, valid, and invalid), and three between-subjects variables: magnitude of validity change, direction of validity change, and change onset timing. Magnitude of validity change refers to by how many percentage points the highlighting validity proportion changed, either 34% or 68%. Direction of validity change refers to whether the highlighting became more valid or less valid. Finally, the experiment was divided into six blocks, affording short rest periods for the subjects between each block. The change in validity occurred with the onset of blocks three or five. Thus the total number of between-subjects conditions was eight.

Procedure

Depending upon which condition a subject was assigned to, the initial validity they encountered was 16%, 34%, 68%, or 84%. Blocks consisted of 60 trials, and with the onset of the third or fifth block the validity level changed. Subjects assigned to the 16% initial validity condition then received 84% validity, and vice versa. Similarly, subjects assigned to the 34% initial validity condition then received 68% validity, and vice versa. The experiment required a little less than 30 minutes for subjects to complete.

Subjects were not told in the instructions that the highlighting validity rate would change during the course of the experiment. It was thought this would be more ecologically valid than telling them that the validity would change. Given a real-world task of searching through some visual display, such as a web page, the user is not informed a priori about how useful visual cues will be. Upon completion of the experiment, subjects were asked explicitly whether or not they noticed any change in highlighting validity and if so, to characterize it.

Since subjects' sensitivity to changes in validity was assessed by changes in response times, we attempted to avoid confounding with practice effects. Therefore, subjects had a full block of 60 practice trials before beginning the actual experiment. The important task components to practice were searching the field to find the target and pressing the appropriate button in response. Allowing subjects to acclimatize to whatever level of highlighting validity they are assigned to before response times are actually recorded might prove detrimental to the attempt to assess their sensitivity to the highlighting validity. Therefore, practice trials were all of the standard type.

Results and Discussion

As in Experiments 1A and 1B, outliers were removed prior to statistical analysis, and this was done both for single trials and entire subjects. The same procedures were followed for data from Experiment 2 as from Experiments 1A and 1B. Two such subjects were found, and their data were removed from further analysis. Figure 4 depicts the mean sensitivity for each of the four conditions with large validity proportion changes while Figure 5 depicts those means for the four conditions with small validity proportion changes.

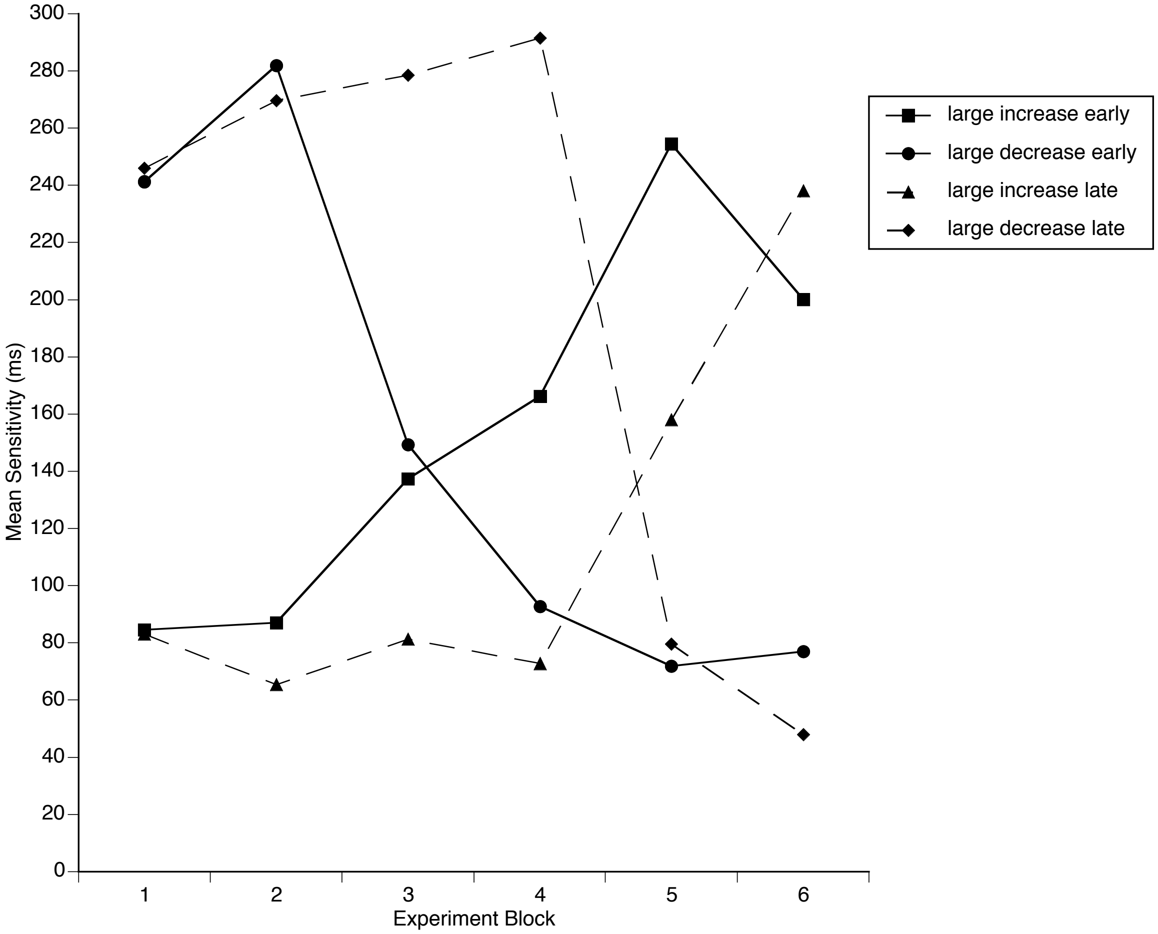


Figure 4. Mean sensitivities for large change conditions.

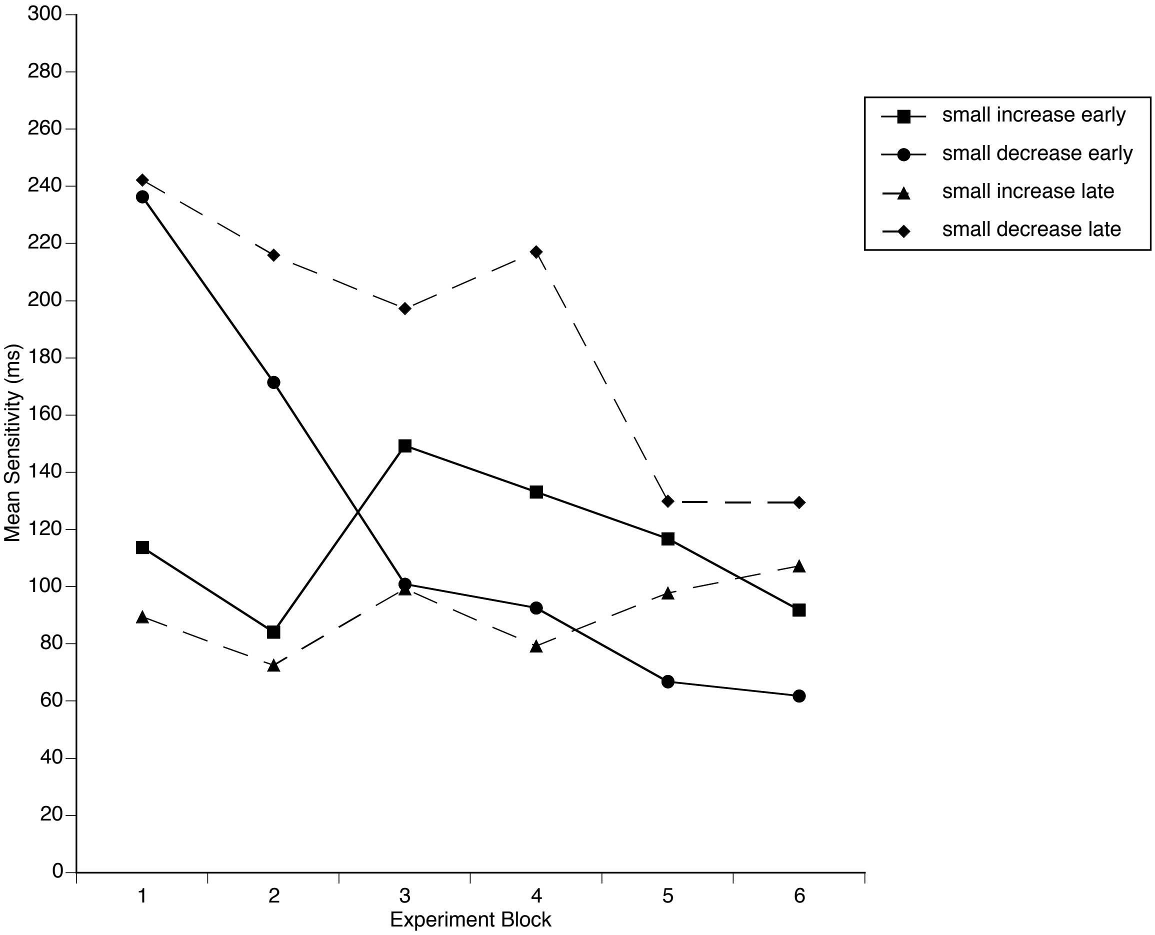


Figure 5. Mean sensitivities for small change conditions.

The slopes of the change in sensitivity for the block preceding change onset, the block of the change onset, and the block after the change onset were examined. These three blocks represent prior sensitivity, sensitivity under adjustment, and posterior sensitivity, respectively, and were therefore of most interest for analyzing changing sensitivity in subjects as they adjusted to new highlighting validity proportions. Breaking the experiment factors down into size of change, direction of change, and onset of change, only direction had any significant effect on slope, ANOVA $F(1,98) = 56.125$, $p < 0.001$. Figure 6 depicts slopes of change in sensitivity for each level of each factor from Experiment 2.

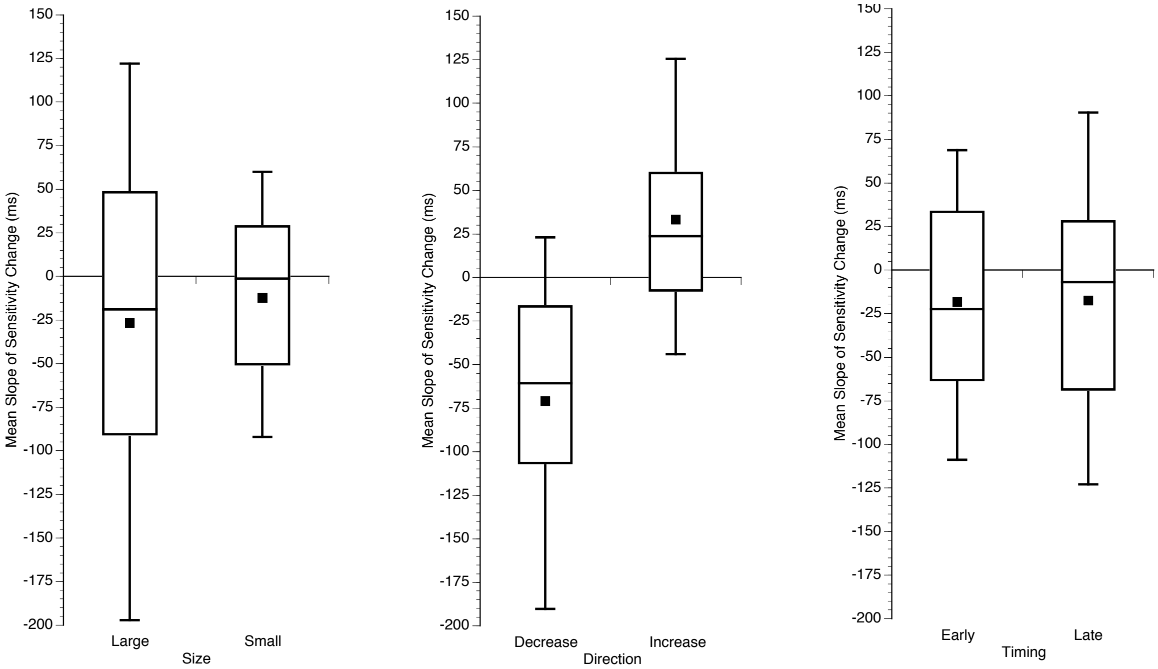


Figure 6. Box plots of slope of sensitivity change for Experiment 2 factor levels.

There was a reliable interaction of size with direction, *F*(1,98) = 13.617, *p* < 0.001. All other *F*'s < 2, *p*'s ≤ 0.170. The direction by size interaction coupled with the main effect of direction means that change direction matters, but it matters more when the change is large than small. Figure 7 plots the direction by size interaction for slope of sensitivity change.

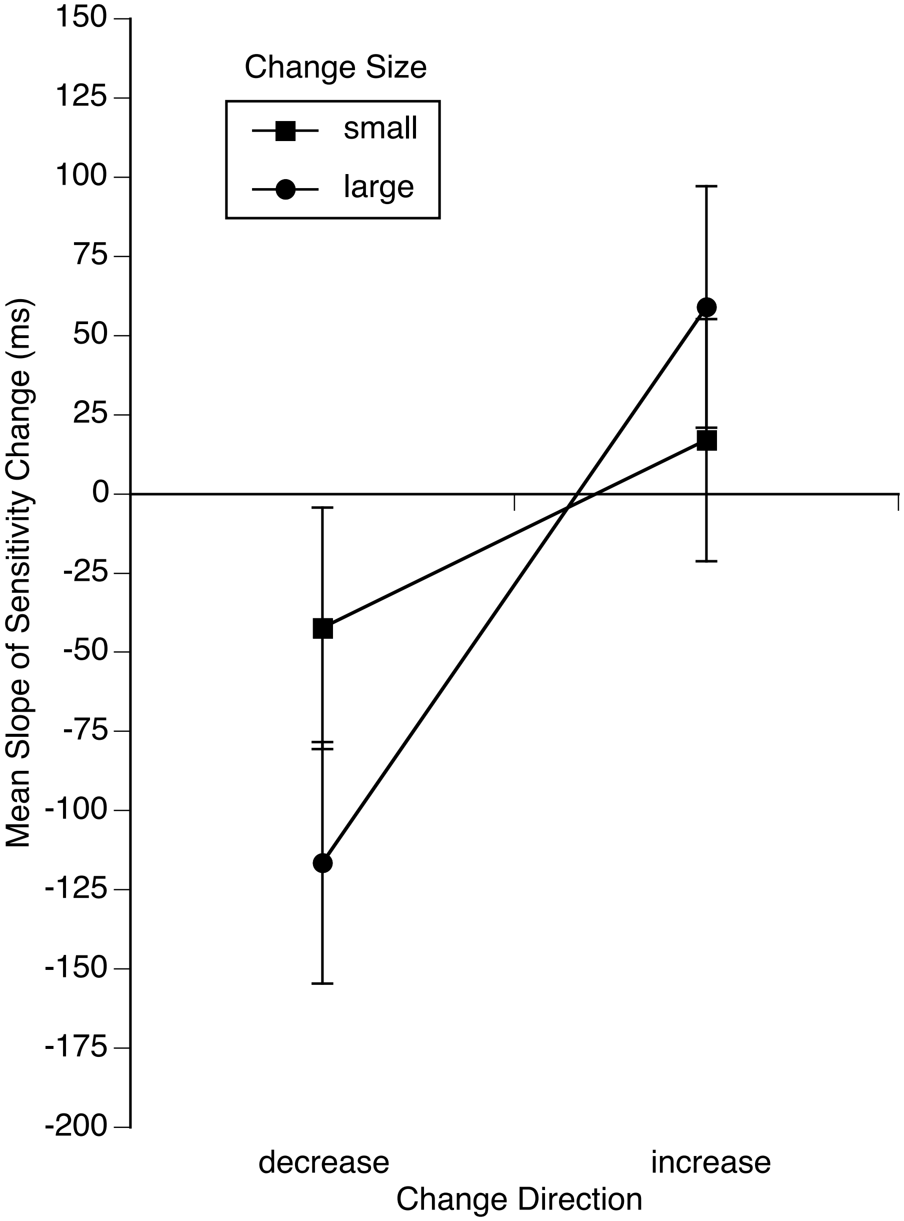


Figure 7. Interaction of change direction and change size for slope of sensitivity change. Error bars indicate standard error of the means.

An interesting question is whether subjects in the decreasing validity conditions changed their sensitivity faster in response to the changing validity than did subjects in the increasing validity conditions. Presumably subjects would notice a drop in highlighting validity faster than they would notice an increase in validity because of their greater attention paid to a higher prior level of validity, as suggested by the results of Experiment 1. Table 1 contains means and standard deviations for absolute slope of sensitivity change for the two groups of subjects. Since the standard deviations for the two groups appeared to be fairly dissimilar, an independent samples *t* test was performed on the absolute deviations of absolute slope, $t(104) = 1.631$, $p = 0.106$. Therefore there is no reason to believe the variances of the two groups are different. Equal variances being assumed, *t* test of the absolute slope for the two groups failed to yield a reliable difference in degree of sensitivity response to the changing validity in the two groups, $t(104) = 1.918$, $p = 0.58$. The observed effect size in this analysis was medium-small, Cohen's $d = 0.377$, and the power to detect an effect of that size was 0.48. The current evidence is therefore inconclusive as to whether the absolute rate of change in sensitivity was different depending upon whether subjects experienced increasing or decreasing validity.

Table 1. Means and standard deviations for absolute slope of increasing and decreasing validity groups.

Direction	Mean	Standard Deviation
Decrease	82.596	78.734
Increase	56.785	58.704

It is perhaps somewhat surprising that no evidence for an effect of onset timing was found, that it does not matter how much prior experience with a certain proportion of highlighting validity a subject has. What is less intuitive is the lack of a main effect of size, and that size only matters in combination with direction. This finding implies that people adjust equally quickly to changes in validity rates no matter the size of the change in validity rate. This result could be a consequence of the small change condition being confounded with relatively moderate levels of highlighting validity, and likewise for large validity changes. Future experimentation might seek to determine whether a lack of main effect for size holds when prior and posterior validity levels are disjoined.

THE ALGEBRAIC MODEL OF OPTIMAL VISUAL SEARCH IN A HIGHLIGHTED DISPLAY

Fisher and colleagues (1989) constructed an algebraic model of visual search in an attempt to determine average search time for a target in displays with or without highlighted options. The model computes expected search times given total set size, highlighted subset size, probability of the target being in the highlighted or unhighlighted subsets, and probability that the subject searches first in the highlighted or unhighlighted subset (Equation 2). The authors concluded that only a model that captures validity probability can account for observed variance of RT as a function of highlighting validity probability. When the display consists of a set of discrete options, as in the Fisher and Tan (1989) paradigm, then the average display search time is a probability mixture of the time to find the target when the highlighted options are searched first and the time to find the target when the unhighlighted options are searched first. Their model assumes users exhaustively search through one subset before searching the other.

$$E_H[T] = e\left\{[(h+1)/2]\left[\sum_{i=1}^{h} P(O_i)\right] + [(n+h+1)/2]\left[\sum_{i=h+1}^{n} P(O_i)\right] + nP(O_{n+1})\right\} + r \tag{2}$$

Equation 2 states that the expected time to find the target, $E_H[T]$, given that the subject searches first the highlighted set, is the sum of four quantities. The first quantity is the product of the conditional expected search time given that the target is present and highlighted times the probability that the target is present and highlighted. Note that in both Experiment 1 and Experiment 2 this last term is equal to the probability that the target is highlighted, since the probability that the target is present in any trial is 1. The second quantity is the product of the conditional expected search time given that the target is present but not highlighted times the probability that the target is present but not highlighted. The third quantity is the product of the conditional expected search time given that the target is absent times the probability that the target is absent. The fourth quantity is the expected response time, in other words, the time required for the subject to decide to press a certain button and then to execute that physical movement. The expected response time when the unhighlighted options are searched first is similarly computed, as shown in equation 3.

$$E_{H'}[T] = e\left\{[(2n-h+1)/2]\left[\sum_{i=1}^{h} P(O_i)\right] + [(n-h+1)/2]\left[\sum_{i=h+1}^{n} P(O_i)\right] + nP(O_{n+1})\right\} + r \tag{3}$$

Assuming the display consists of a set of discrete options, as it does in both experiments, then the average display search time can be derived. It is the probability of first searching the highlighted subset times the expected time to find the target if it is in the highlighted subset plus the probability that the target is in the unhighlighted subset times the expected time to find the target if it is in the unhighlighted subset (Equation 4).

$$E[T] = p_H E_H[T] + (1 - p_H)E_{H'}[T] \tag{4}$$

Fisher et al. Assume that p_H (probability that subjects search first through the highlighted subset) should increase as the average time it takes to locate the target by attending first to the unhighlighted options grows proportionately larger than the average time it takes to locate the target by attending first to the highlighted options (i.e., as the ratio $E_{H'}[T]/E_H[T]$ increases). Equation 5 is the function they offer that satisfies that constraint.

$$p_H = 1 - EXP\{-2.0(E_{H'}[T]/E_H[T])\} \tag{5}$$

The present instantiation of the Fisher et al. model assumes 50 ms to decide to attend highlighting at the beginning of a trial, 85 ms to shift visual attention, 100 ms to decide to attend a new location, and 150 ms to issue a response. These latencies are based on ACT-

R’s (Anderson et al., 2004) predictions of how long these actions should take. Figure 8 plots the predicted mean RTs of Fisher et al.’s optimal model of visual search in the 5 x 1 array paradigm for trials with valid and invalid highlighting. Also shown is data collected from human subjects. Note that the optimal model predicts a highlighting sensitivity in excess of 300 ms for even the low validity percentage conditions, whereas the humans showed sensitivity in these conditions more on the order of 40 or 50 ms. Note also how the predicted RTs for valid trials are approximately 350 ms less than those obtained. Predicted RTs for invalid trials also were reliably less than those obtained from humans, $t(14) = 2.409$, $p = 0.030$, and the human RTs increased at a faster rate as a function of increasing validity level, ANOVA with linear contrast on highlighting validity $F(1,14) = 5.803$, $p = 0.030$.

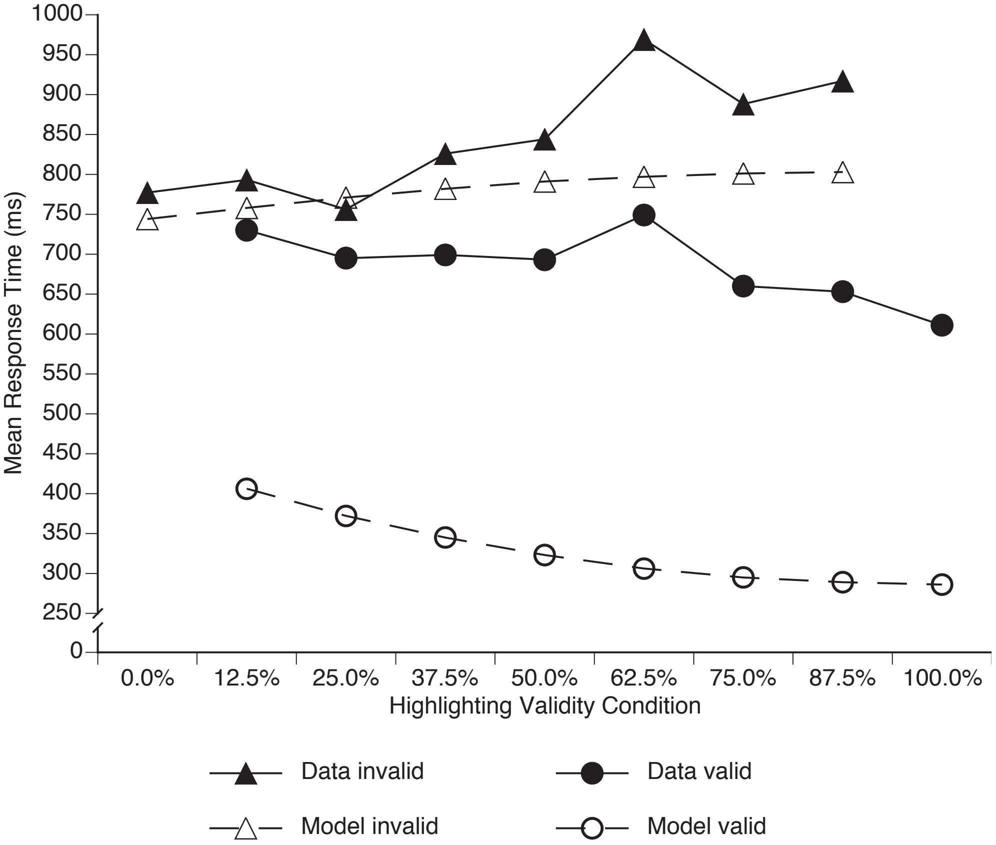
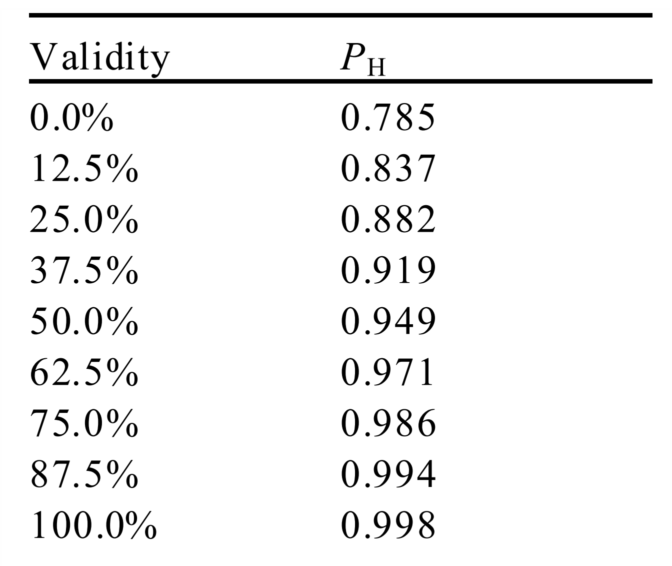


Figure 8. Predicted RTs of Fisher et al.'s optimal model and RTs from Experiment 1.

The optimal model computes P_H, the probability that the subject first examines the highlighted subset. Table 2 lists P_H for each level of highlighting validity. When both the total number of items and number of highlighted items is low, as it is in this experiment, P_H will always be high because the random chance of finding the target in any given position is high (20%). The high P_H is what drives the model to predict such fast times for valid trials, and it may be that subjects are not as fast in valid trials because they are not so likely to search the highlighted item first.

Table 2. Values for P_H given highlighting validity.



Validity	P_H
0.0%	0.785
12.5%	0.837
25.0%	0.882
37.5%	0.919
50.0%	0.949
62.5%	0.971
75.0%	0.986
87.5%	0.994
100.0%	0.998

Finally, Fisher et al.'s algebraic model does not take into consideration the possibility that subjects may attend to some subset of the unhighlighted options, then attend to the highlighted option, then if necessary, finish searching the unhighlighted options. As shall be shown with the ACT-R models, the ability to attend to the highlighting at any time may be important for replicating the human data.

What contributes to the adaptability of human visual search behavior in an environment where cues have some probability of being helpful or hindering? Given a fairly simple paradigm like the digit search in a 5 x 1 array used by Fisher and Tan, there are a couple of decisions that could be made by the user at a variety of different times throughout the task. Users can potentially elect to attend or avoid the highlighting each time they shift their visual attention, though they might not actually make that decision that often. Then there are issues of self-evaluation of performance: do users evaluate their own performance at the level of whole trials or some smaller level, such as individual shifts of attention? Fisher et al.'s algebraic model has no answer to offer for these questions.

An ACT-R model can shed some light on the above questions, since it must be explicitly programmed to attend or avoid highlighting either with every shift of attention or only the first one. It could also evaluate its own performance at either a macro level (i.e., a full trial) or micro level (i.e., each shift of attention). Two ACT-R models were constructed to compare to Fisher et al.'s algebraic model and to make inferences about people's search behavior in reference to the micro versus macro issues discussed above.

THE ACT-R MODEL

A useful framework for considering the information search task outlined above is the embodied cognition-task-artifact (ETA) triad (Byrne, 2001). The central notion is that the behavior and performance of a user interacting with some device is a function of the properties of three things: the cognitive, perceptual and motor capabilities of the user, termed embodied cognition, the task the user is engaged in, and the artifact the user is employing in order to perform the task. Cognitive modeling forces the analyst to consider all three pieces of the triad at once. For a simulation model to run that provides a description of the capabilities and limitations of the user, it must be given both a task to perform and a complete and detailed description of the artifact being used. Byrne constructed two models of visual search using computer menu artifacts. He discovered that the model that fit user data better used local, rather than global, search strategies. Furthermore, learning played a role in the development of those strategies as the model weeded out ineffective strategies.

The ACT-R (Anderson et al., 2004) cognitive architecture was used to construct a detailed model of the data from Experiments 1A, 1B, and 2. ACT-R is a hybrid cognitive architecture using a combination of a symbolic production system and subsymbolic computation with perceptual and motor modules that allow the cognitive system to interact with the simulation's outside world. The symbolic production system uses a set of production rules, each of which specify under what circumstances that production may be executed (known as the left-hand side of the production), and then what the model will do when that production fires (right-hand side of the production). The subsymbolic component of ACT-R computes thresholds and latencies for certain operations, such as the retrieval of information from long-term memory. ACT-R is an ideal tool to explore the effect of highlighting validity on performance in a visual search task because of its inclusion of perceptual-motor modules that can accurately simulate human response times in a wide variety of tasks, as well as its subsymbolic evaluation of the utility of production rules.

ACT-R's visual system separates vision into two modules, "visual-location" and "visual-object", each with an associated buffer (Anderson et al., 2004). When a production makes a request of the visual-location module, the production specifies a series of constraints, and the visual-location module returns information representing a location meeting those constraints. Constraints are attribute-value pairs that can restrict the search based on visual properties of the object (such as "color: red") or the spatial location of the object (such as "vertical: top"). This is akin to preattentive visual processing (Treisman & Gelade, 1980) and supports visual pop-out effects. But to identify an object, ACT-R must make a request of the visual-object module. A request to the visual-object module entails providing information representing a visual location, which will cause the visual-object module to shift visual attention to that location, process the object located there, and return information representing the object. Thus ACT-R can conduct feature-guided visual search.

One of ACT-R's learning mechanisms uses a competitive evaluation of production rule utility. In any given cycle of computation, the left-hand side of multiple production rules can match the state of the buffers. ACT-R fires the competing production with the highest utility. These utilities are learned through experience–each time a production is fired, ACT-R "remembers" the time cost of using that production, and whether the production helps the model to succeed or fail in meeting its goal.

Specifically, ACT-R evaluates a production rule's utility according to Equation 6, where P_i is the expected probability that firing production *i* will lead to successful completion of the objective. The objective is completed when a later production marked as a success or failure is executed. C_i is the expected cost of achieving the objective, in seconds, and G is the value of the objective, also in seconds (the default value is 20 seconds). ACT-R uses Equation 6 to compute the parameter, P and Equation 7 to compute the parameter, C. Successes and Failures are the number of successes and failures experienced when using production *i*, and Efforts is the accumulated time over all the successful and failed applications of production rule *i*. Successes, Failures, and Efforts can all be set to simulate prior experience.

$$U_i = P_i G - C_i \tag{6}$$

$$\mathrm{P} = \frac{\mathrm{Successes}}{\mathrm{Successes} + \mathrm{Failures}} \tag{7}$$

$$\mathrm{C} = \frac{\mathrm{Efforts}}{\mathrm{Successes} + \mathrm{Failures}} \tag{8}$$

ACT-R uses the utility of a production to decide which production to fire out of a set of competing productions (the conflict set) that match the conditions of the internal state of the model and the state of the outside world. The probability that production *i* will be selected on the basis of its utility, *U*, out of conflict set *j* is calculated with Equation 9, where *t* controls the noise in the utilities. Summation is over all competing productions.

$$P_i = \frac{e^{U_i/t}}{\sum_{j}^{n} e^{U_j/t}} \tag{9}$$

On any given trial with highlighting present, the model selected and fired one of two productions which caused it to attend to (”attend-red”) or avoid the red item (”avoid-red”). At the start of each highlighted trial, both productions match the contents of the buffers, which include a buffer-stuffed red item. “Attend-red” made the model move visual attention to the red item. “Avoid-red” made it seek an unattended black item. If the currently attended item was a target, the models output a press of the appropriate key. If the item was a distractor, a production, “highlighted-distractor”, fired. This production was marked as a failure and it made the critical difference for getting the model to duplicate the effect of validity condition on sensitivity. If the red item was still unattended, then the model still had the decision to make as to whether to attend to or avoid the highlighting (“avoided-red-distractor-find-red” and “avoided-red-distractor-find-black”, respectively). In the case of control trials, the model simply fixated unattended items until it found the target.

Production priors were set as in Table 3. The priors for "avoid-red" were set equal to the random chance that the target was in either the highlighted or unhighlighted subset, 20% chance for the highlighted group since one item would be highlighted out of a total of five items in the display. Priors for "avoided-red-distractor-find-red" and "avoided-red-distractor-find-black" were set equal to the random chance that the target would be in either highlighted or unhighlighted subset, given that one unhighlighted item was already examined and found to be a distractor. The probabilities for successes and failures for "attend-red" were originally set following the same rule, thus 20 successes and 80 failures. But it was found that the model did not attend the highlighted item often enough to generate a similar RT pattern to the human data, so they were changed to 50 successes and 50 failures. This could imply that people may be inclined to check highlighted subsets at rates greater than chance.

Table 3. Production priors for the ACT-R model.

Production	Successes	Failures	Initial Utility
attend-red	50	50	9.95
avoid-red	80	20	15.90
avoided-red-distractor-find-red	25	75	4.90
avoided-red-distractor-find-black	75	25	14.90

One consequence of this set of priors is that the model will be slightly biased toward examining an unhighlighted item first, but not as biased as if it were simply ignoring highlighting altogether and checking each item randomly. ACT-R computes the utility of every production *i* using Equation 6, where P_i is an estimate of the probability that if production *i* is chosen the current goal will be achieved, *G* is the value of that current goal, and C_i is an estimate of the cost (typically measured in time) to achieve that goal (Anderson et al., 2004). Equation 7 computes *P* and C_i is the simulated time elapsed between firing production *i* and firing another production that is marked as a success. Thus "attend-red"'s utility started at .5 * 20 - 0.05 = 9.95 versus 15.90 for "avoid-red."

This seems valid as a set of starting assumptions about the expectations about highlighting subjects brought with them to the experiment given a lack of measurement of the match between subjects' information search goals over the course of their lives and the highlighted items they have encountered. The priors were scaled to values between 10 and 100 such that the model would start a run with some biases in the productions it chose, but not so biased as to be inflexible to learning new utilities over the course of the 384 trials.

Figure 9 illustrates the mean response times for human and model data, per trial type, per highlighting validity condition. The model provided a good fit to the data, $r = 0.982$, mean deviation = 67 ms. The one flaw in the model is its poor prediction of response times for valid trials in low validity conditions. Because it is too slow on such trials, the model's invalid trials fail to show the kind of sensitivity to validity condition that the human invalid trials do. While we have determined that the model actually does actually avoid red items all the way through invalid trials in the low validity conditions, it is not apparent why the model generates such slow response times for these trials. We hope that detailed examination of proportions of firings of "attend-red", "avoid-red", and their subsequent fixation counterparts will yield clues as to what the model is actually doing for these trials.

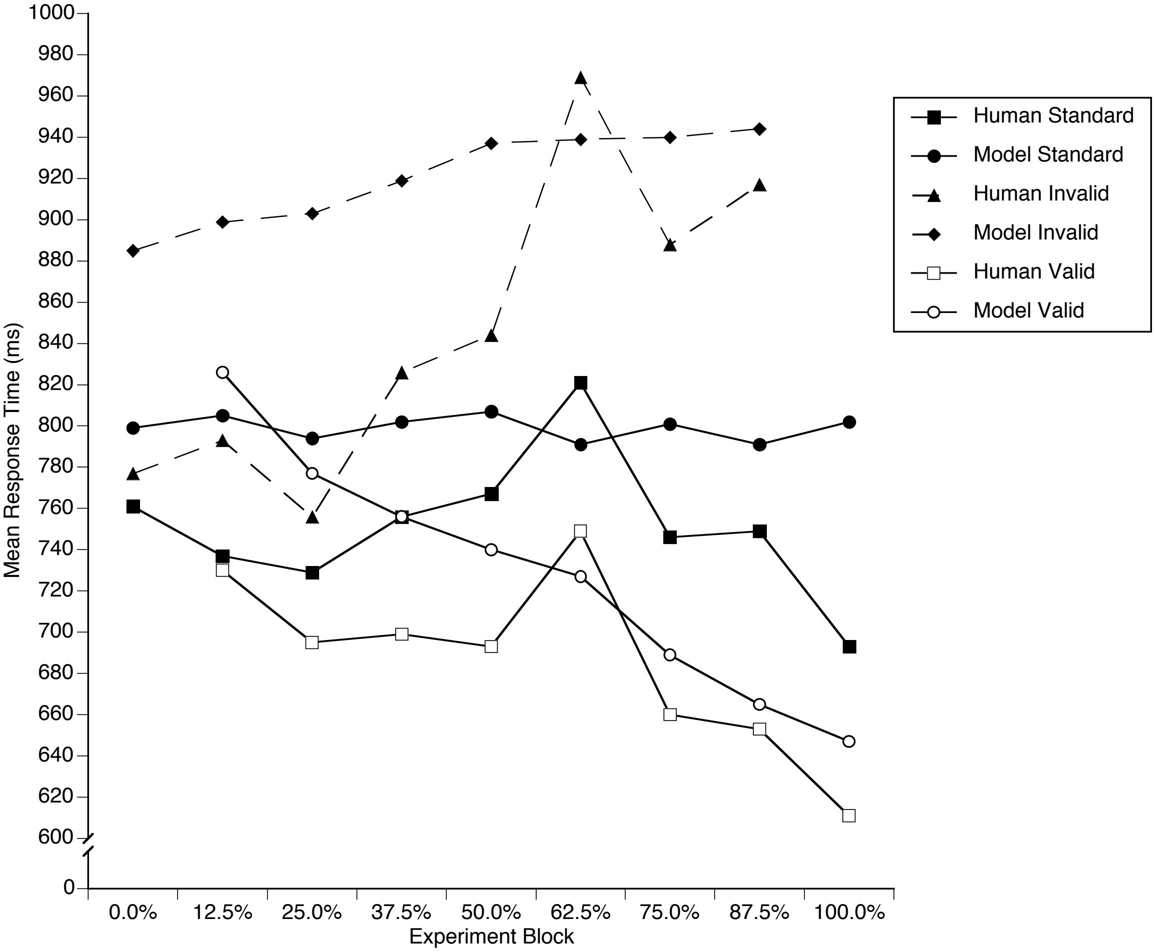


Figure 9. Human data from Experiments 1A and 1B compared to model data.

Despite the major flaw described above, the model's sensitivity effect of validity proportion does closely mimic that of the human data (Figure 10), $r = .957$, $p < 0.001$. With the model for Experiments 1A and 1B displaying such odd behavior, it seems unwise to proceed to attempt to model Experiment 2 until the Experiment 1 model can more closely simulate the human data for invalid trials. This is especially true given that we hope that an unaltered model of the first experiments will provide a good fit to Experiment 2's data. We hope one model will capture all the major effects in both experiments because we think that the same basic mechanisms are driving the sensitivity effect of validity proportion in Experiments 1A and 1B and the rate of change of that sensitivity in response to changing validity in Experiment 2: decision-making at the sub-trial level and evaluation of success and failure at the sub-trial level. We think these mechanisms are important for generating the discussed effects in this paradigm because the most successful model of the first experiment does these two things. Models that did not incorporate both of these features simply did not generate the sensitivity functions that the described model has.

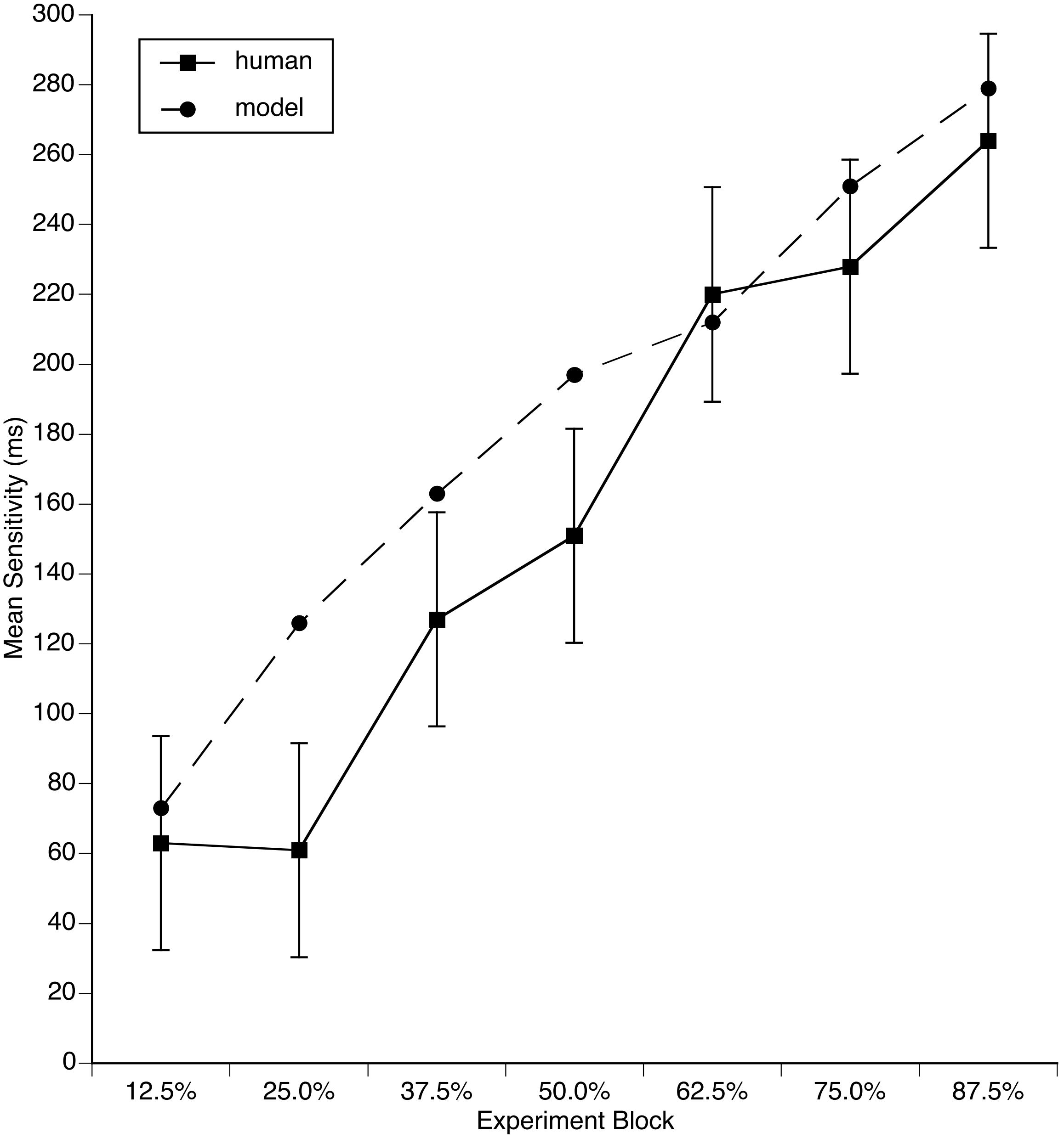


Figure 10. Mean sensitivity for model and human data.

The ACT-R model generally showed the same variances as humans related to highlighting validity condition and trial type. Appendix B contains histograms plotting number of trials against 100 ms RT bins for humans and the model by validity condition and by trial type. The model's data appears more discrete in its distributions compared to human data because it ran with most stochastic features turned off for the sake of reducing time required for computation. The lack of stochasticity in computations is also likely the reason for the model's sometimes lack of positive skew in its RT distributions relative to the human data.

GENERAL DISCUSSION

Fisher and Tan (1989) found that color highlighting at 50% validity led to response times that were no faster than no highlighting at all. Yet the results from Experiments 1A and 1B plainly contradict that: subjects were 74 ms faster on average with 50% valid highlighting than with no highlighting. Calculating the expected response time shows why. For standard trials, on average subjects are going to have to make 2.5 shifts of attention to find the target, since there is a total of five items in the display. At approximately 200 ms per attention shift, that means it will take 500 ms on average to find the target, plus another 250 ms on average to press the button. Indeed, the mean response time for standard trials in Experiments 1A and 1B was 751 ms. By comparison, highlighting is still worthwhile even at 50%, because at that proportion a highlighted trial will only require an average of two shifts of visual attention. If the highlighting is valid, then only one attention shift is necessary. If the highlighting is invalid, then three attention shifts are necessary on average. Since there is an equally likely chance of the highlighting being valid or invalid, the overall average number of required attention shifts will be two.

People have demonstrated themselves to be sensitive to this kind of optimality, or at least they behave that way. The model reveals that how they get there seems to be by the use of micro-level decisions and feedback. Furthermore, the model posits an explicitly endogenous locus of control for attending to the cue. The model only attends to the red item when it selects and fires the production that causes it to “voluntarily” attend to that stimuli. While it is possible that this completely explicit endogenous cuing might have something to do with the model’s current difficulties with invalid trials, it is difficult to see how that may be. If invalid trials hurt performance because people involuntarily attend to the highlighting, then in the extreme case that cuing is entirely exogenous there should be no sensitivity function of highlighting proportion at all because people should be attending to the highlighting equally often in all validity proportion conditions. If people operate on some kind of mixed endogenous and exogenous cuing mechanism in this paradigm, then there should still be some function of sensitivity because people are still going to be able to exercise some degree of control over whether or not they attend to the highlighting, and it is not clear how that could affect the valid and invalid trials differentially, as they are in the model. To some degree, people are learning the probabilities with which highlighting leads them to the target, they are using that knowledge to exercise control over where they look, and they are doing this at temporal levels on the order of 200 ms.

REFERENCES

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Quin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*, 1036-1060.

Byrne, M. D., (2001). ACT-R/PM and menu selection: Applying a cognitive architecture to HCI. *International Journal of Human-Computer Studies, 55*, 41-84.

Fisher, D. L., Coury, B. G., Tengs, T. O., & Duffy, S. A. (1989). Minimizing the time to search visual displays: Trole of highlighting. *Human Factors, 31(2)*. 167 – 182.

Fisher, D.L., and Tan, K.C. (1989). Visual displays: The highlighting paradox. *Human Factors*, *31*(1), 17 – 30.

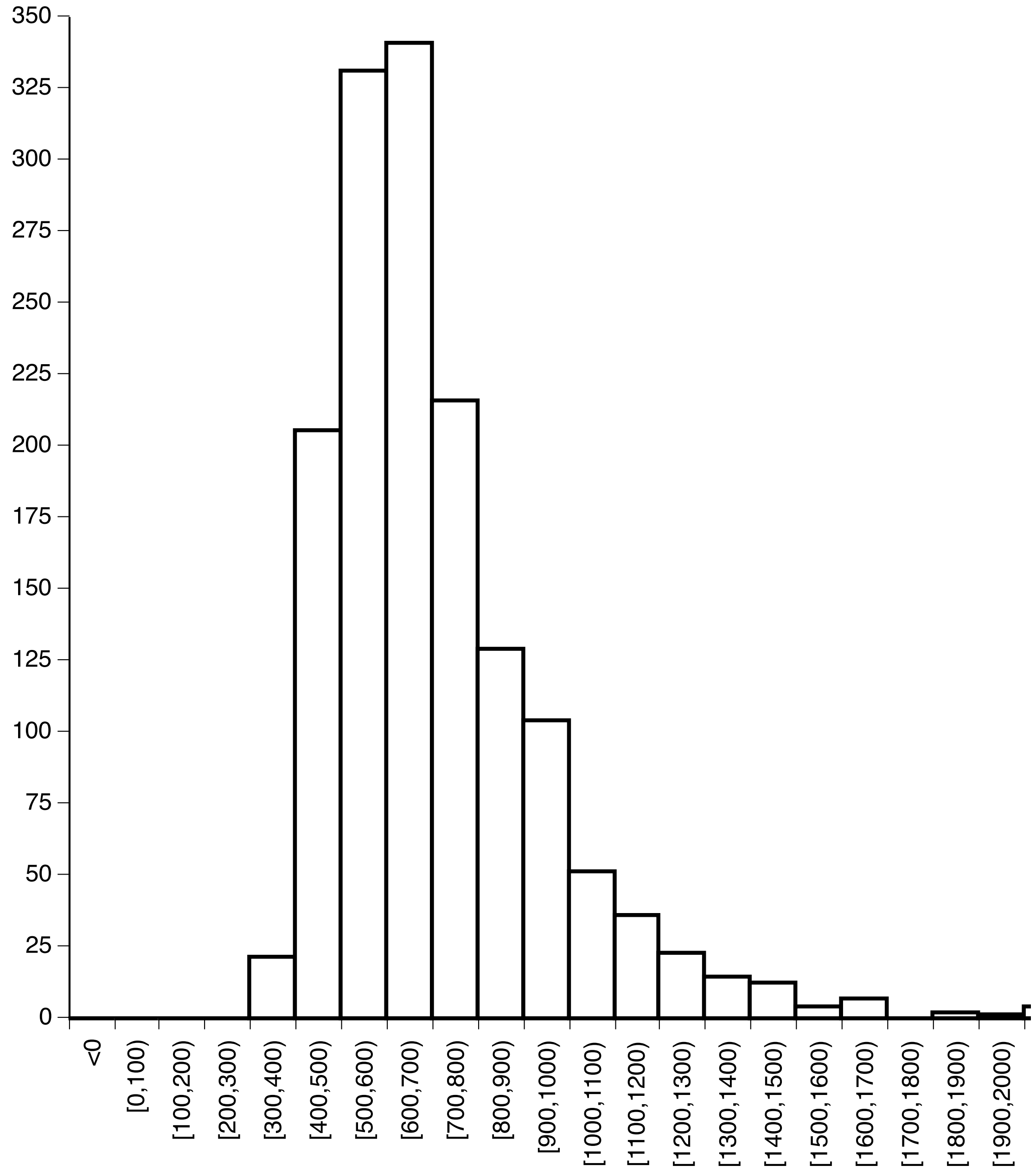
Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re- examination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15.* 233 – 250.

Treisman, A.M., and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*. 97 – 136.

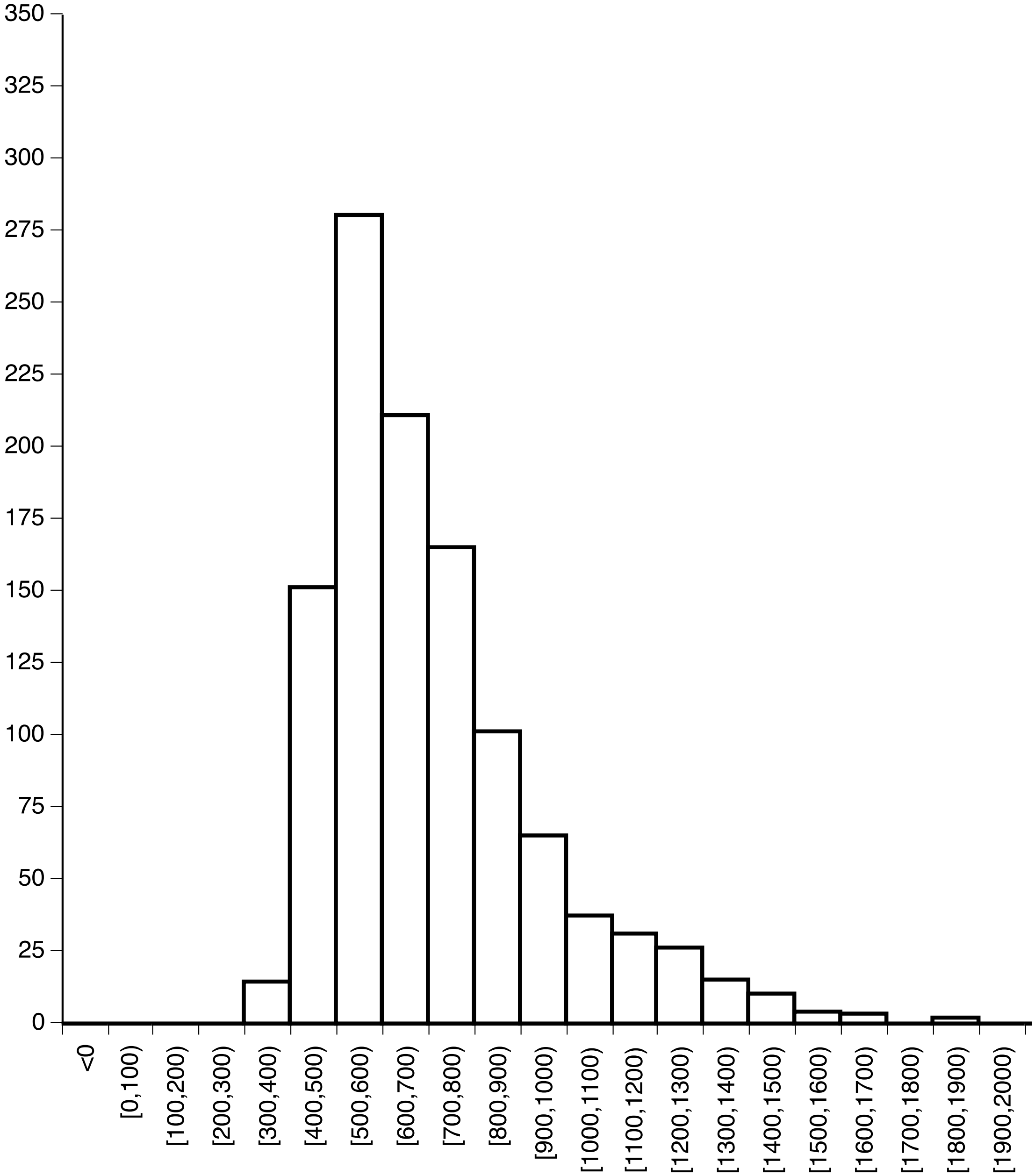
Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, 1*(2). 202 – 238.

APPENDIX A: HISTOGRAMS OF HUMAN RESPONSE TIME IN INVALID TRIALS AS A FUNCTION OF TARGET DISTANCE FROM HIGHLIGHTED ITEM

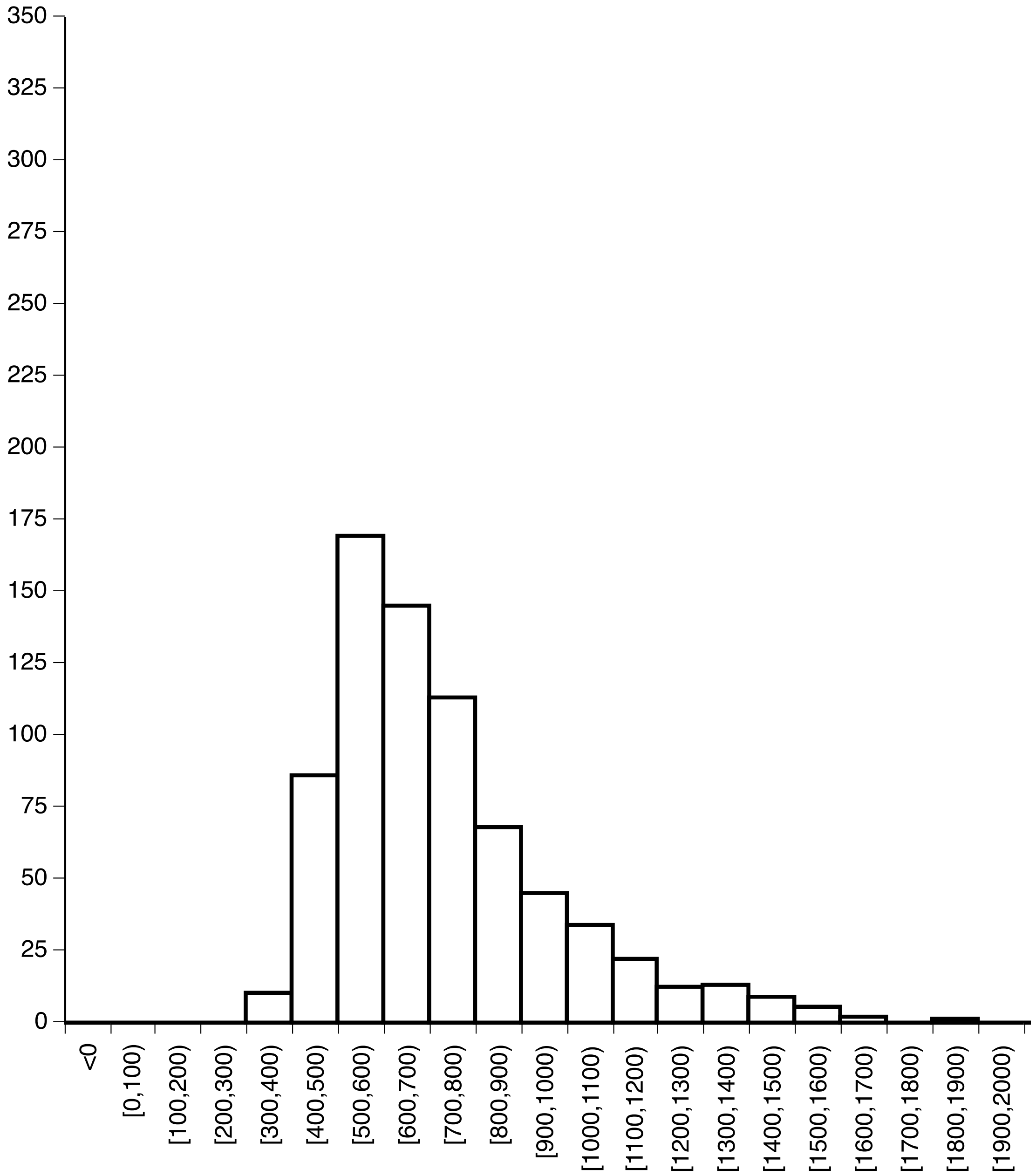
All histograms plot number of trials on the y axis and RT bins on the x axis. All distances are noted in terms of the number of item places in the display (e.g., “distance = 1” indicates that the items are adjacent).



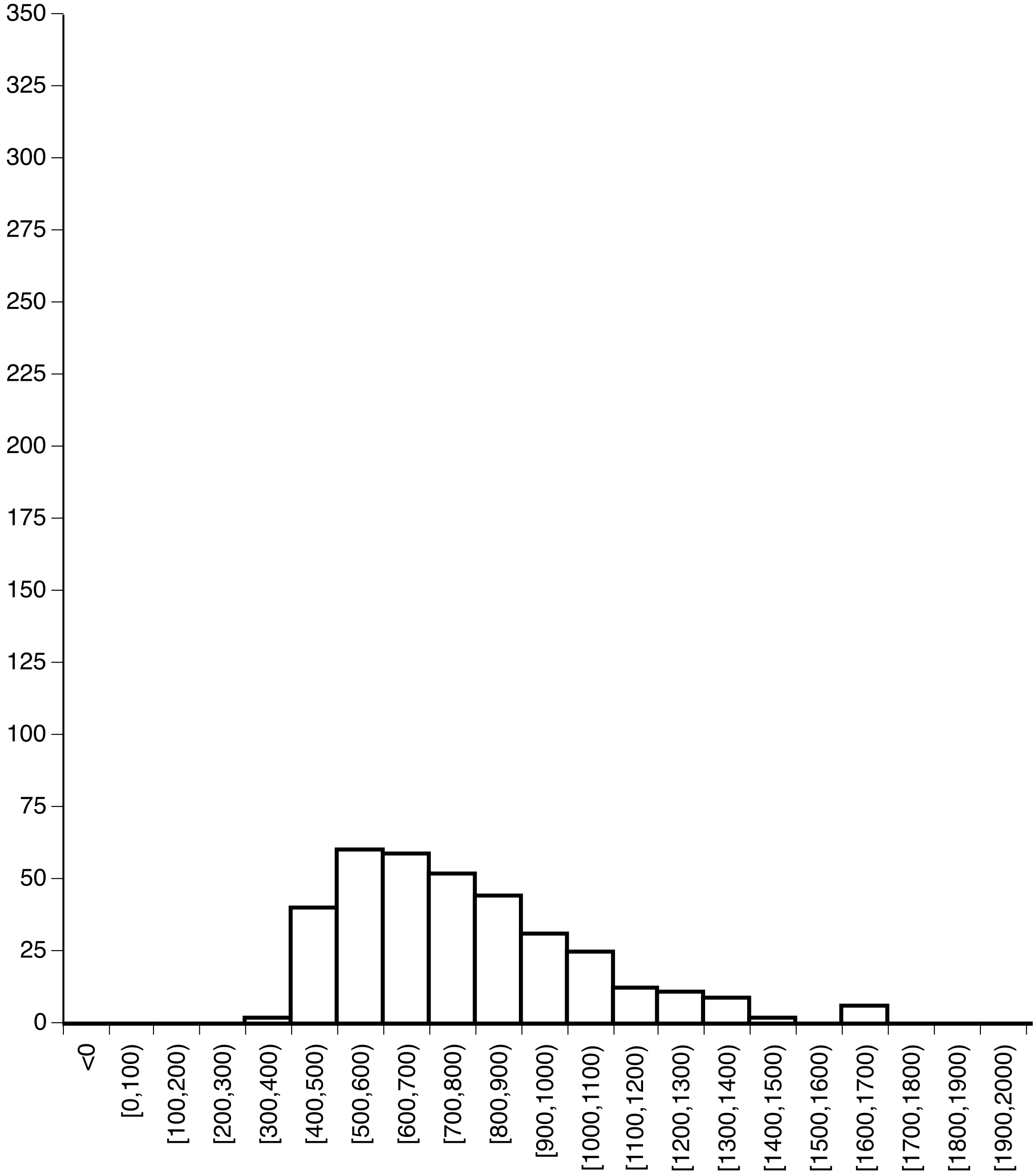
0% Validity, Distance = 1



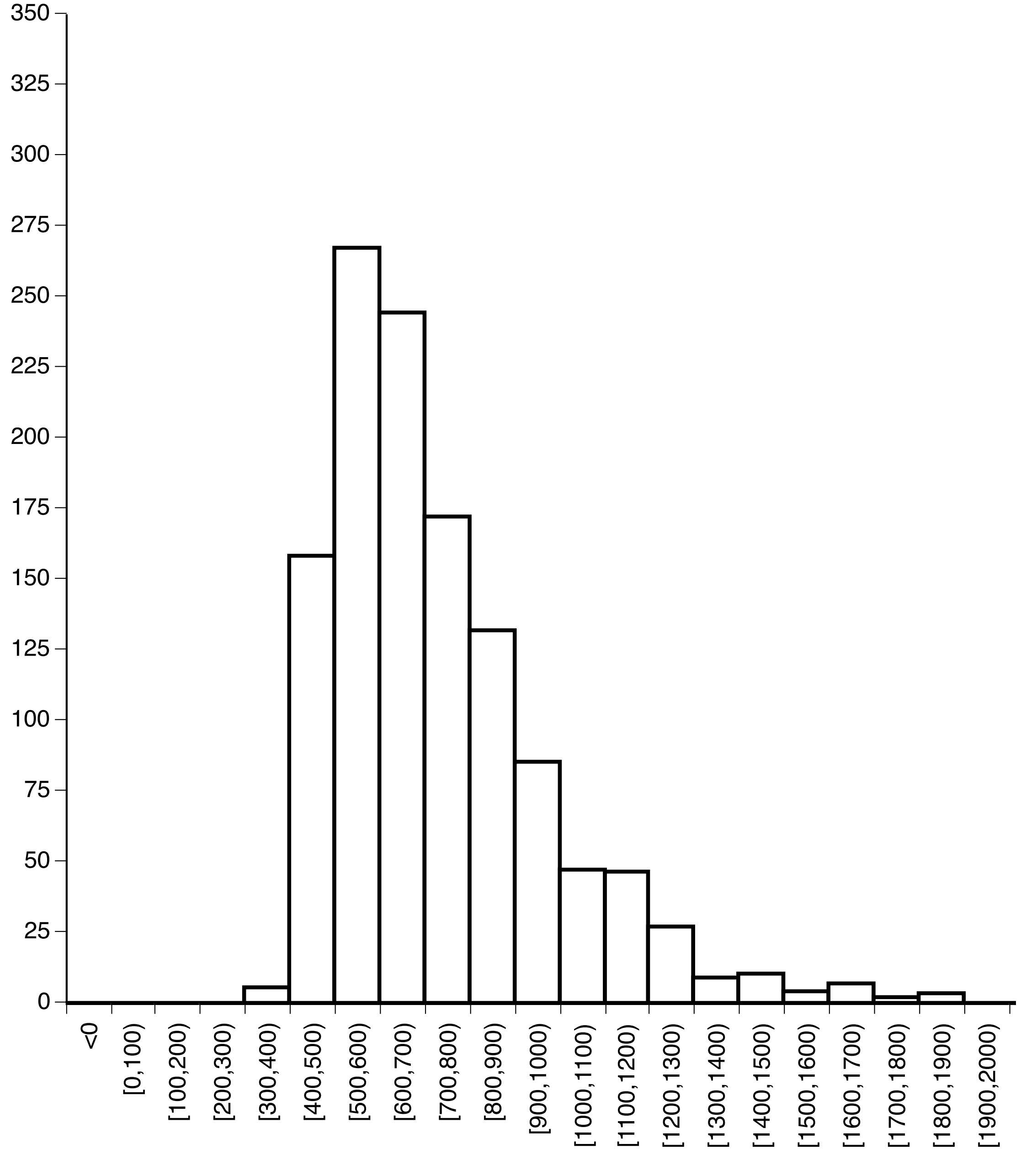
0% Validity, Distance = 2



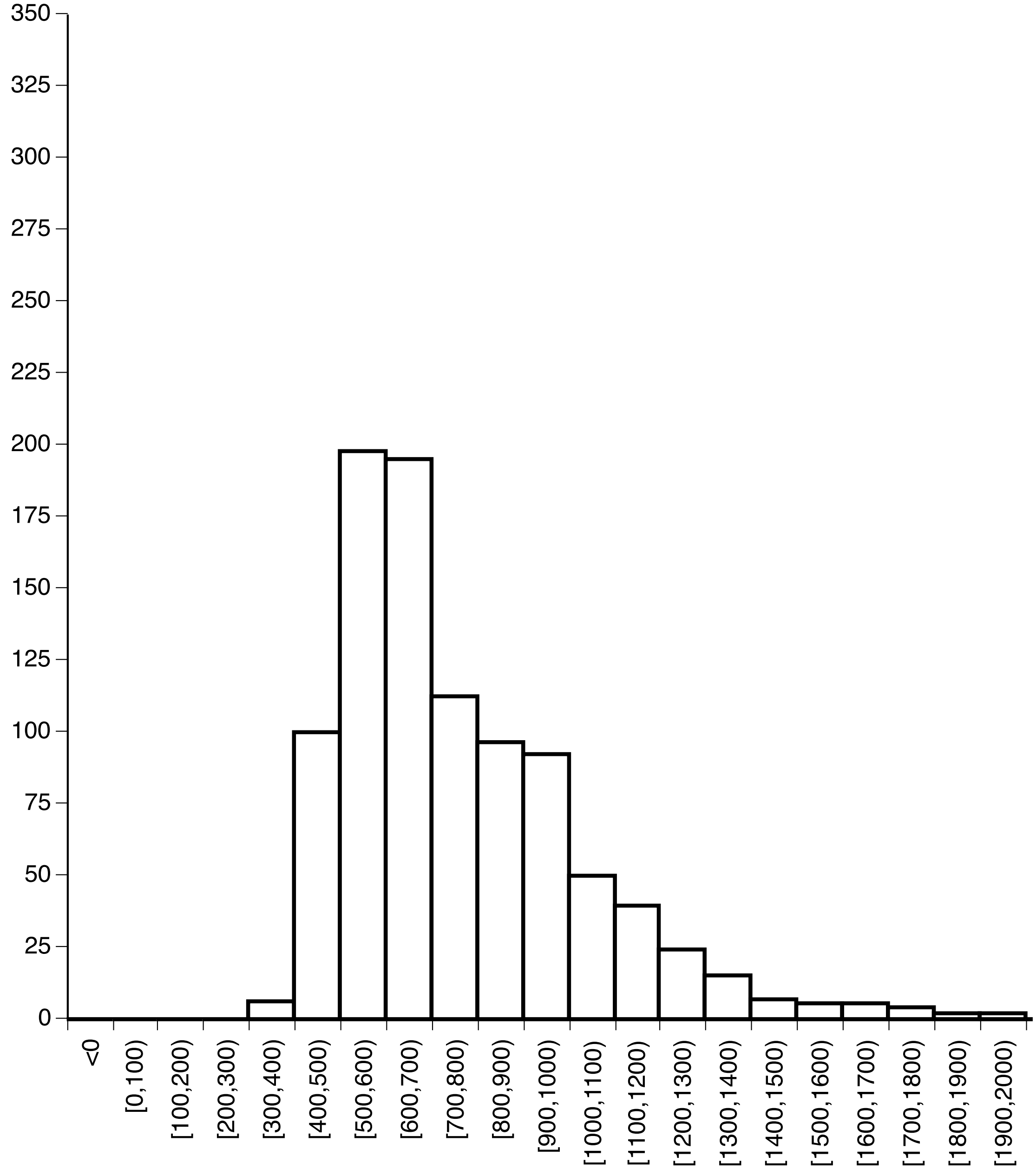
0% Validity, Distance = 3



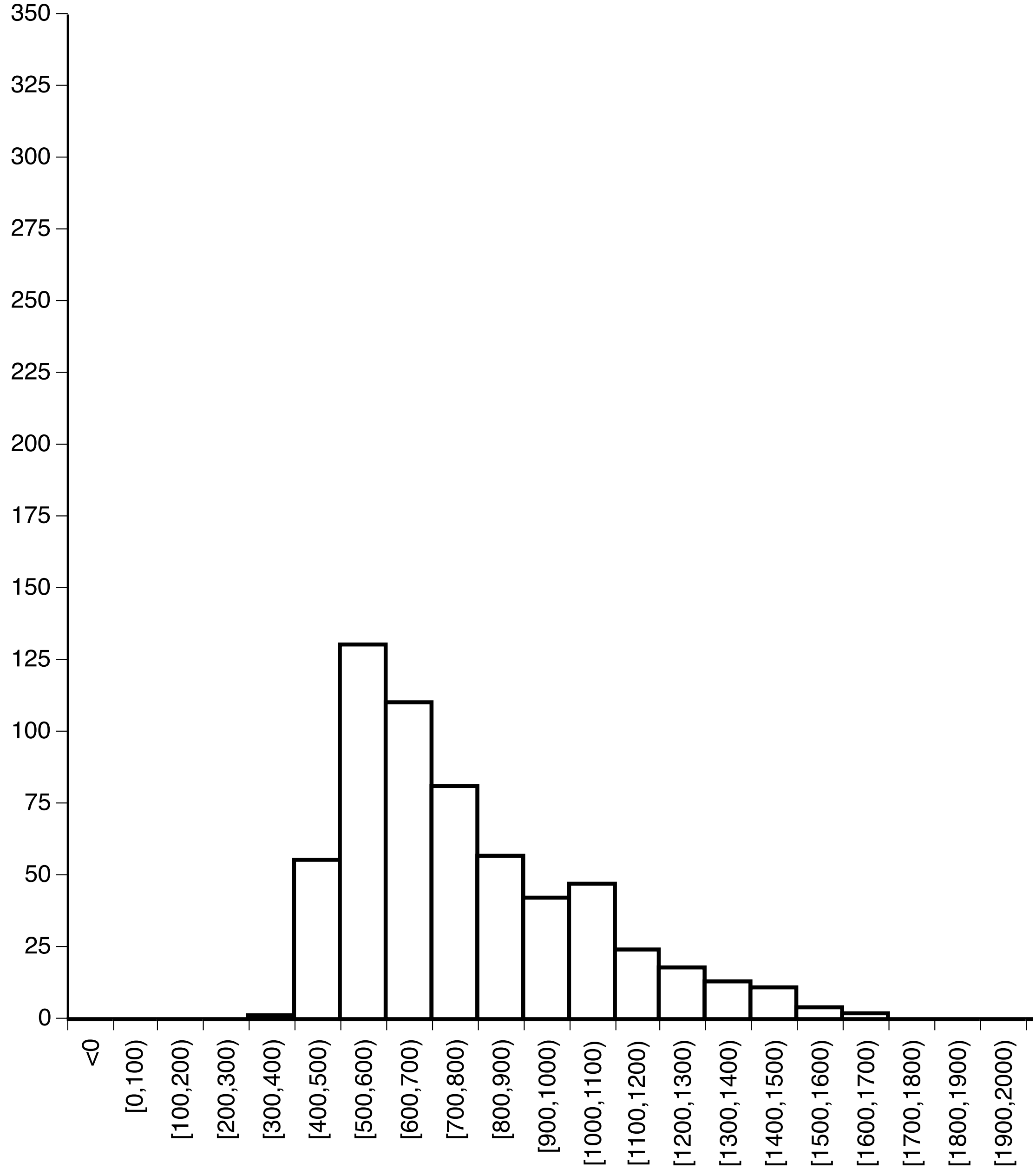
0% Validity, Distance = 4



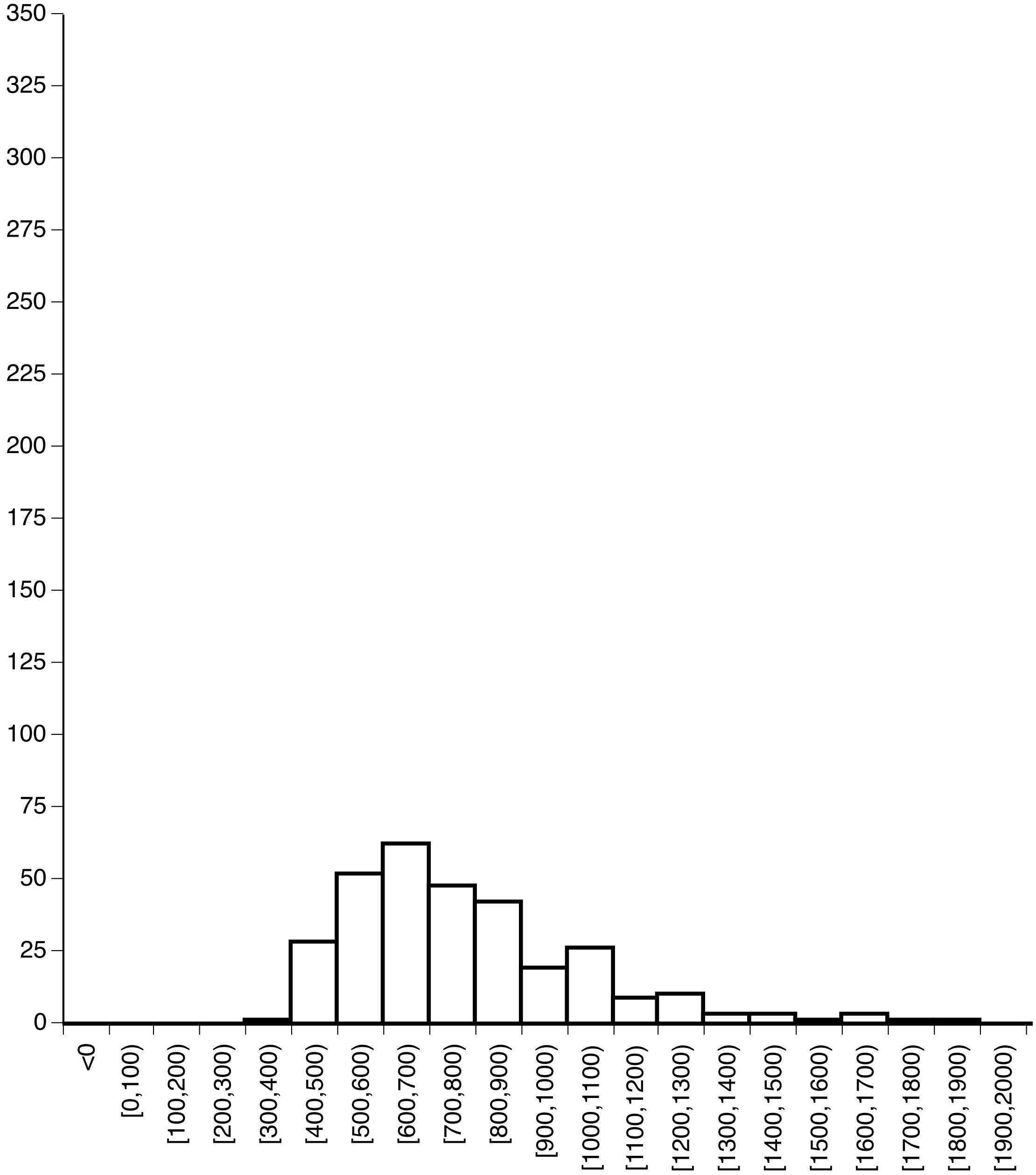
12.5% Validity, Distance = 1



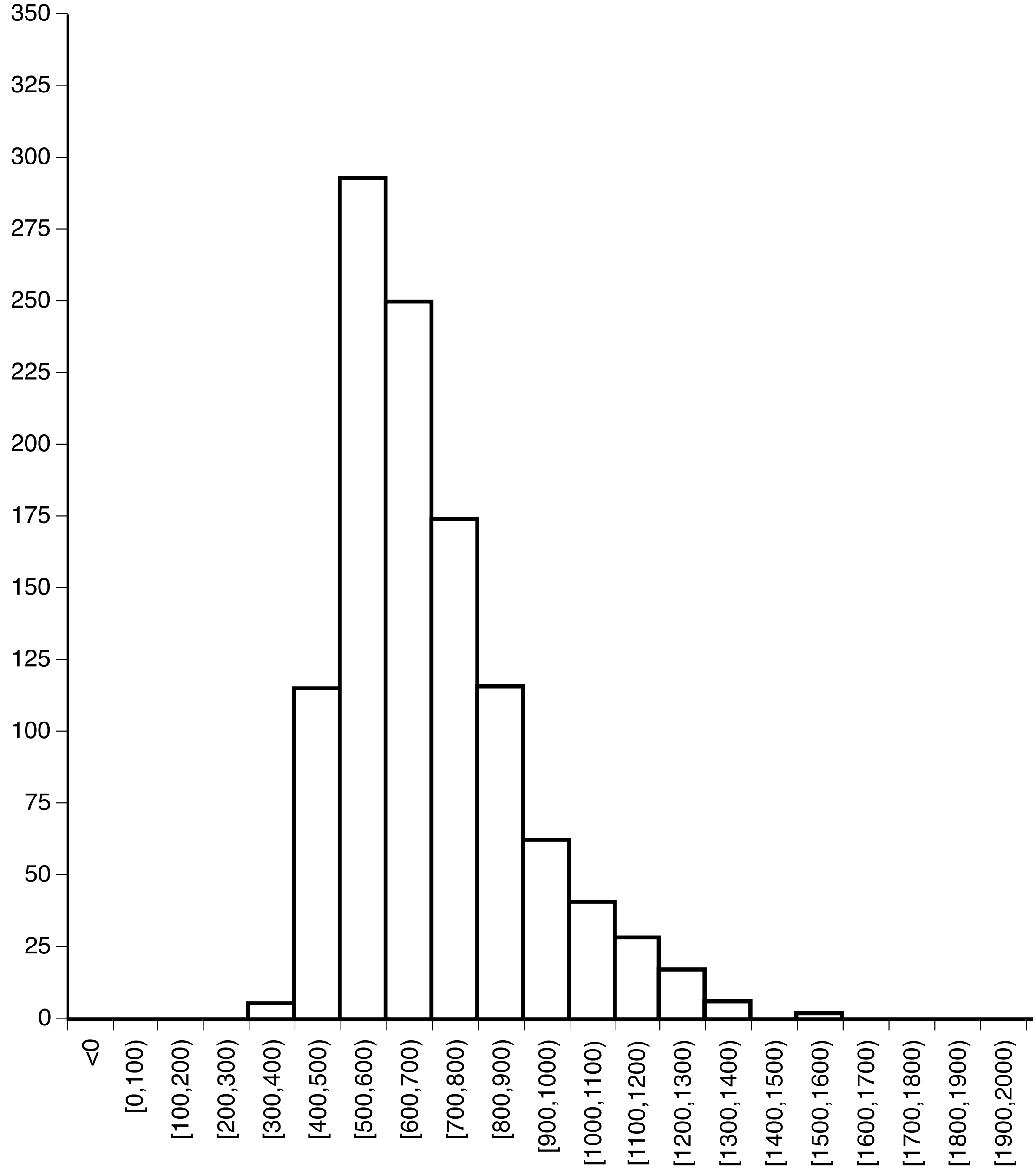
12.5% Validity, Distance = 2



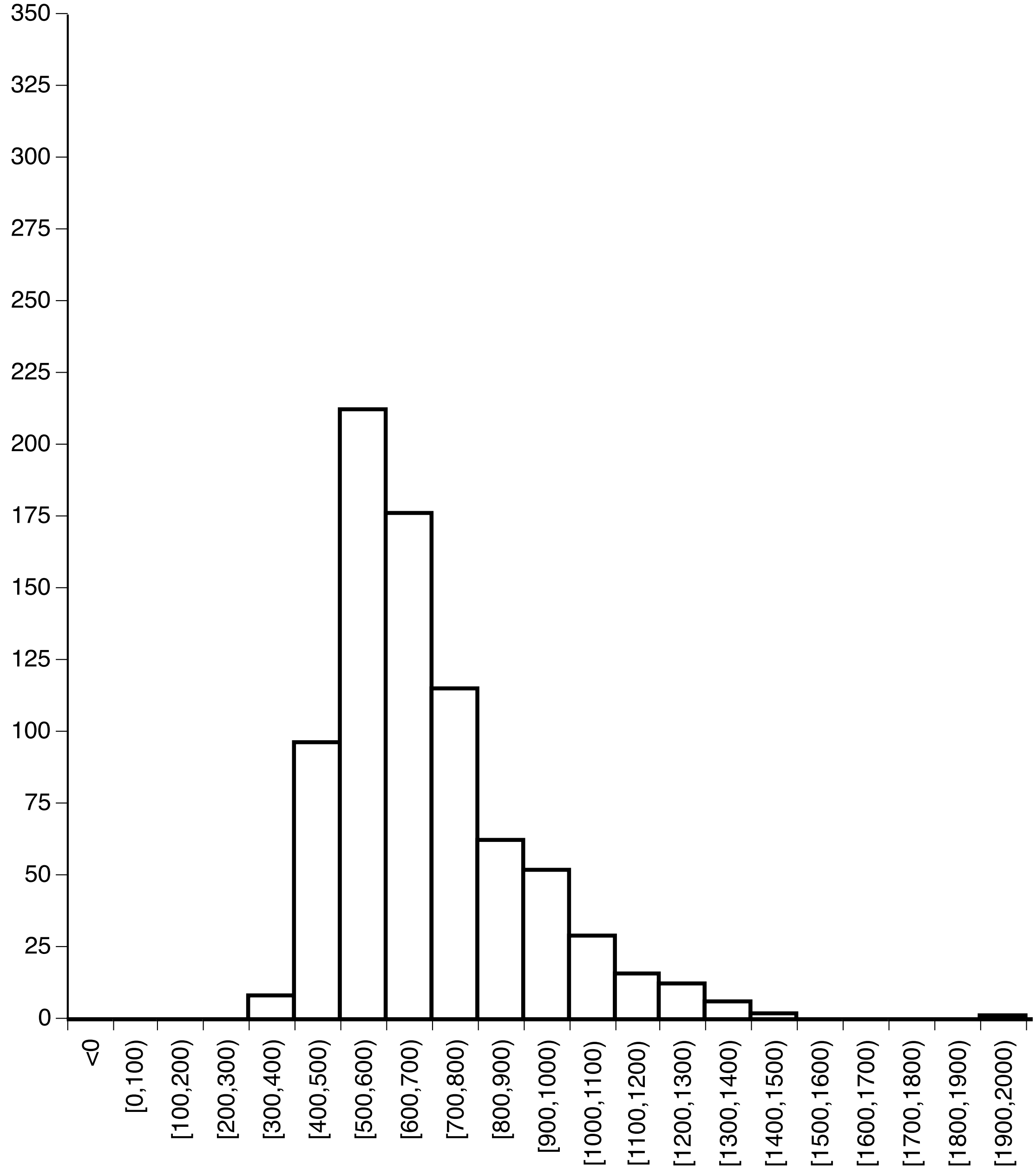
12.5% Validity, Distance = 3



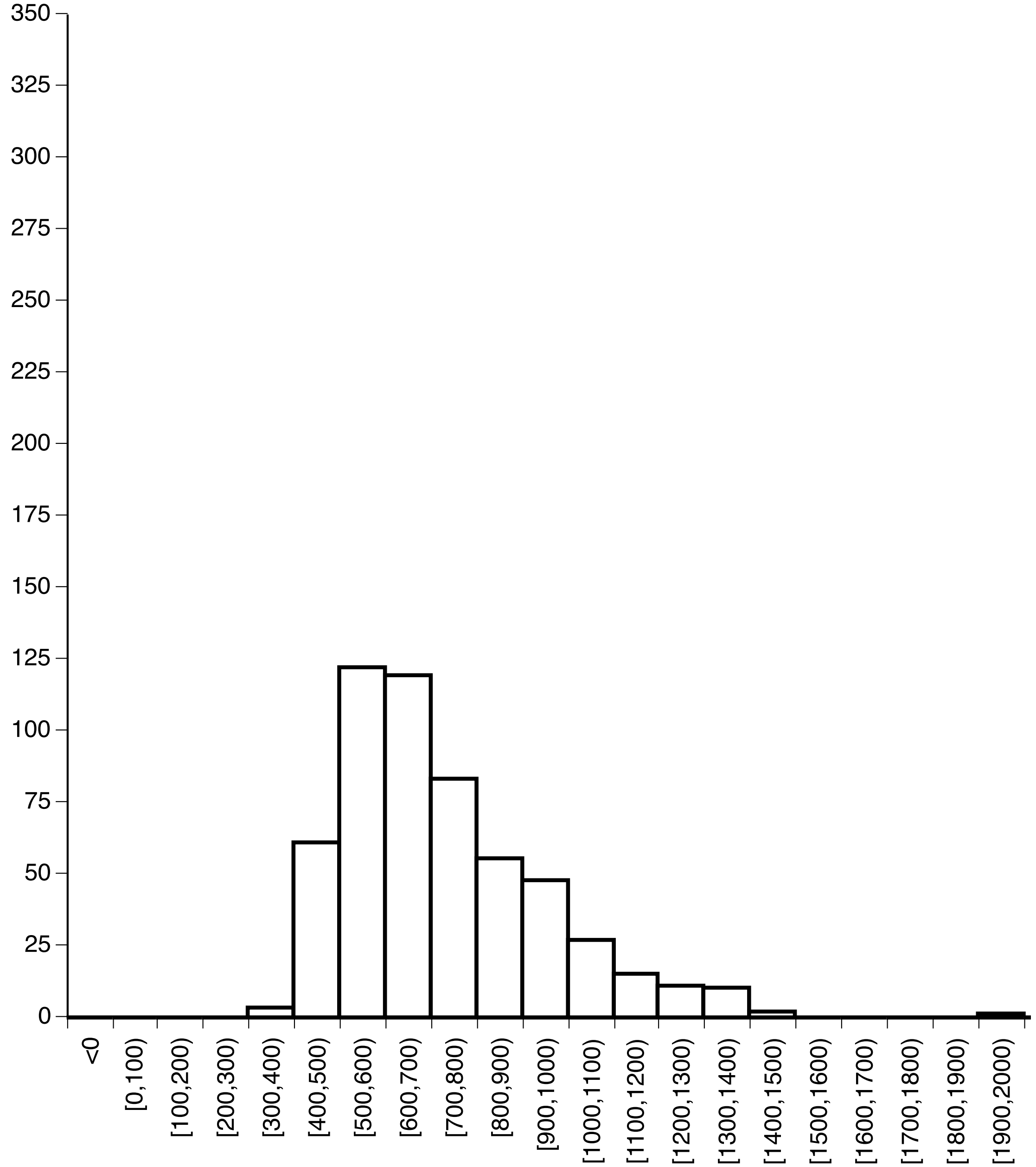
12.5% Validity, Distance = 4



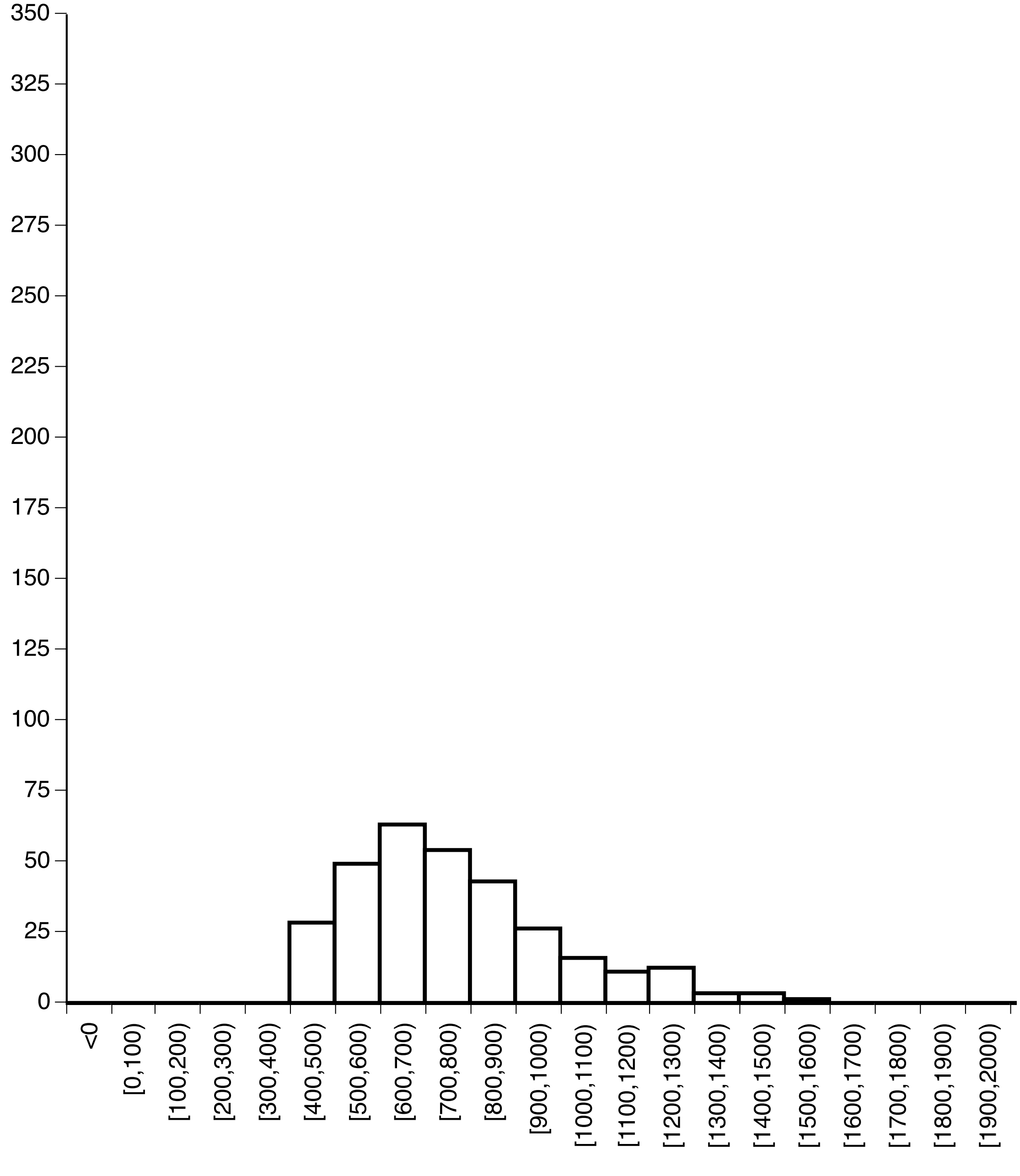
25% Validity, Distance = 1



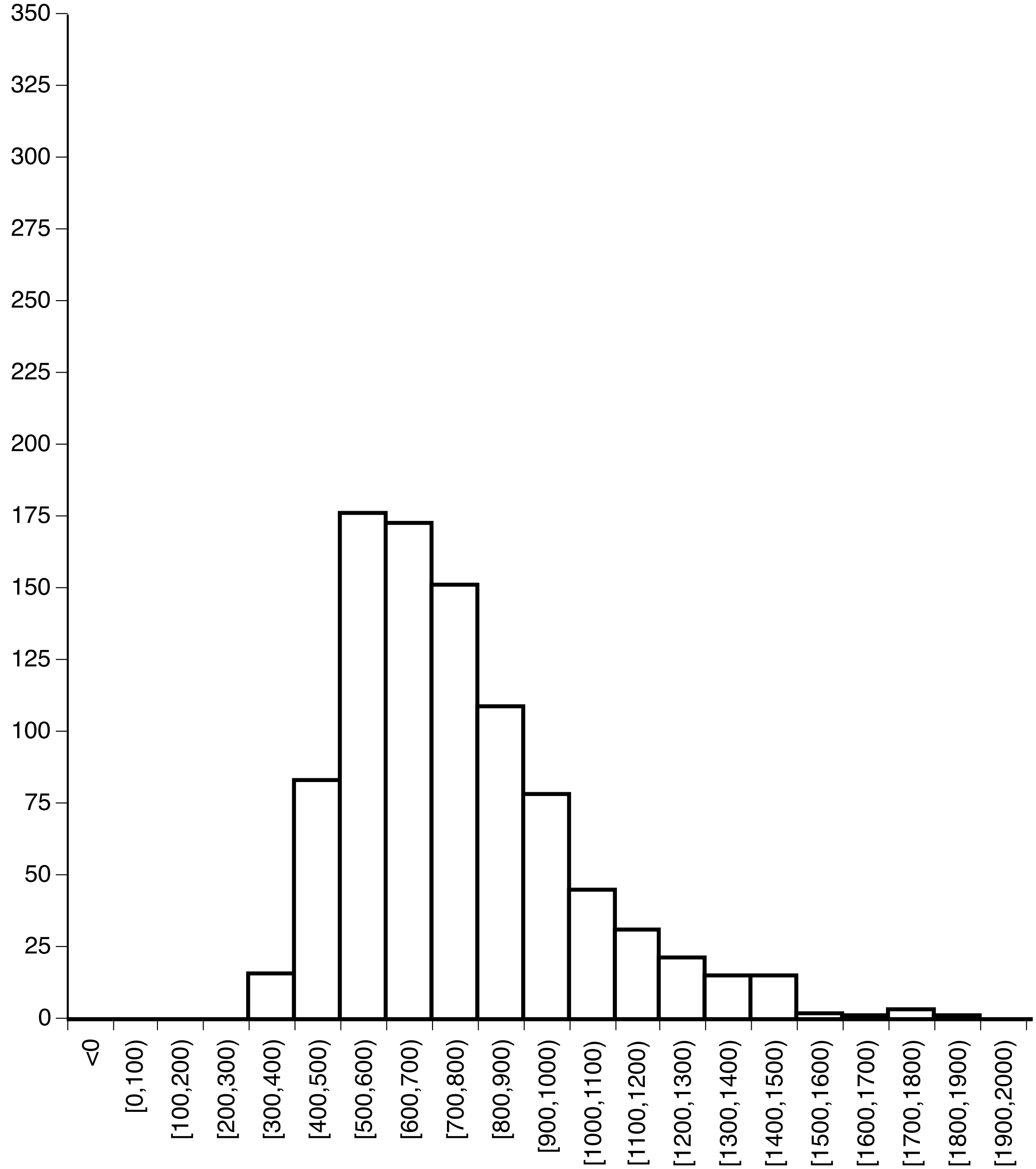
25% Validity, Distance = 2



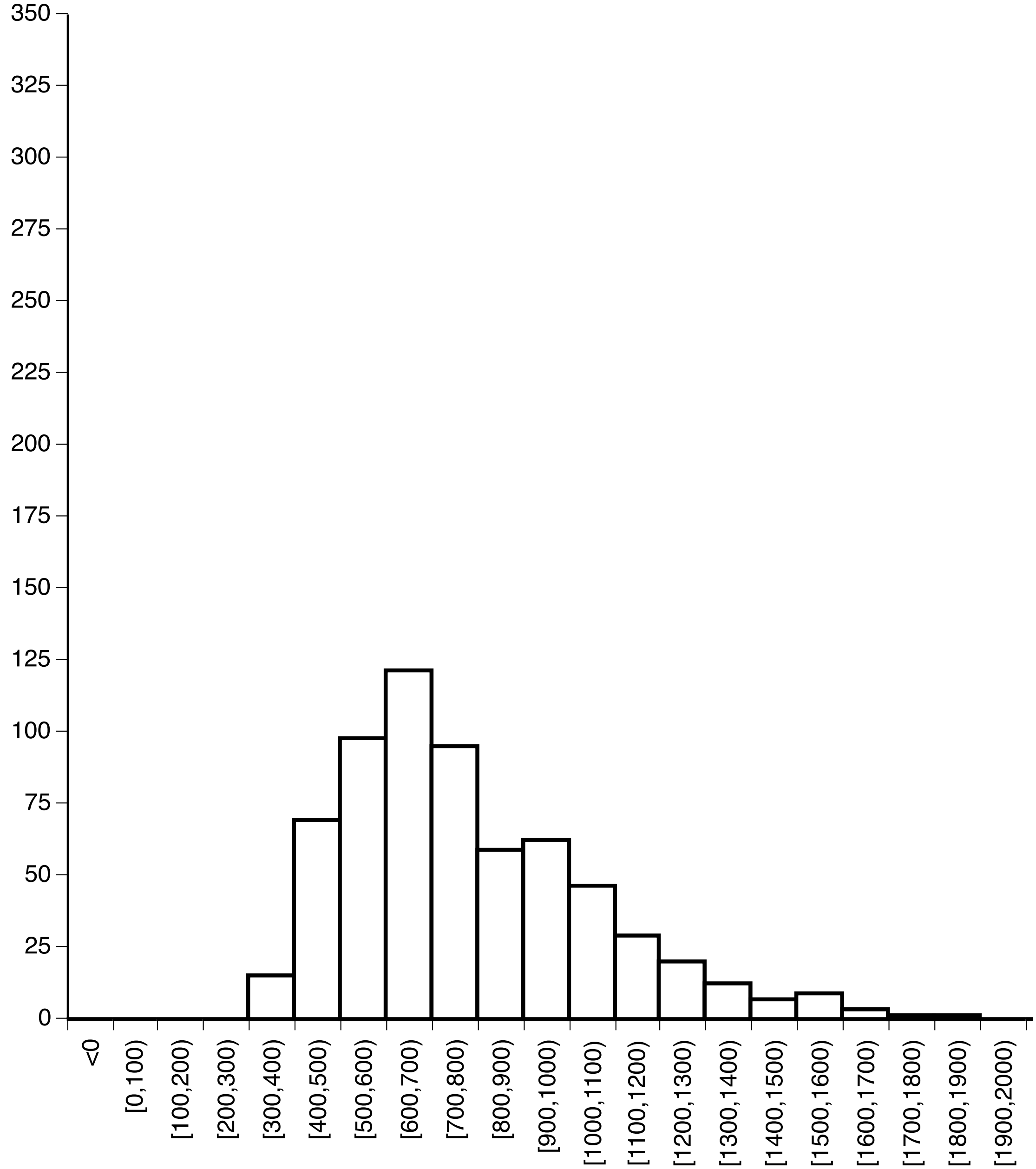
25% Validity, Distance = 3



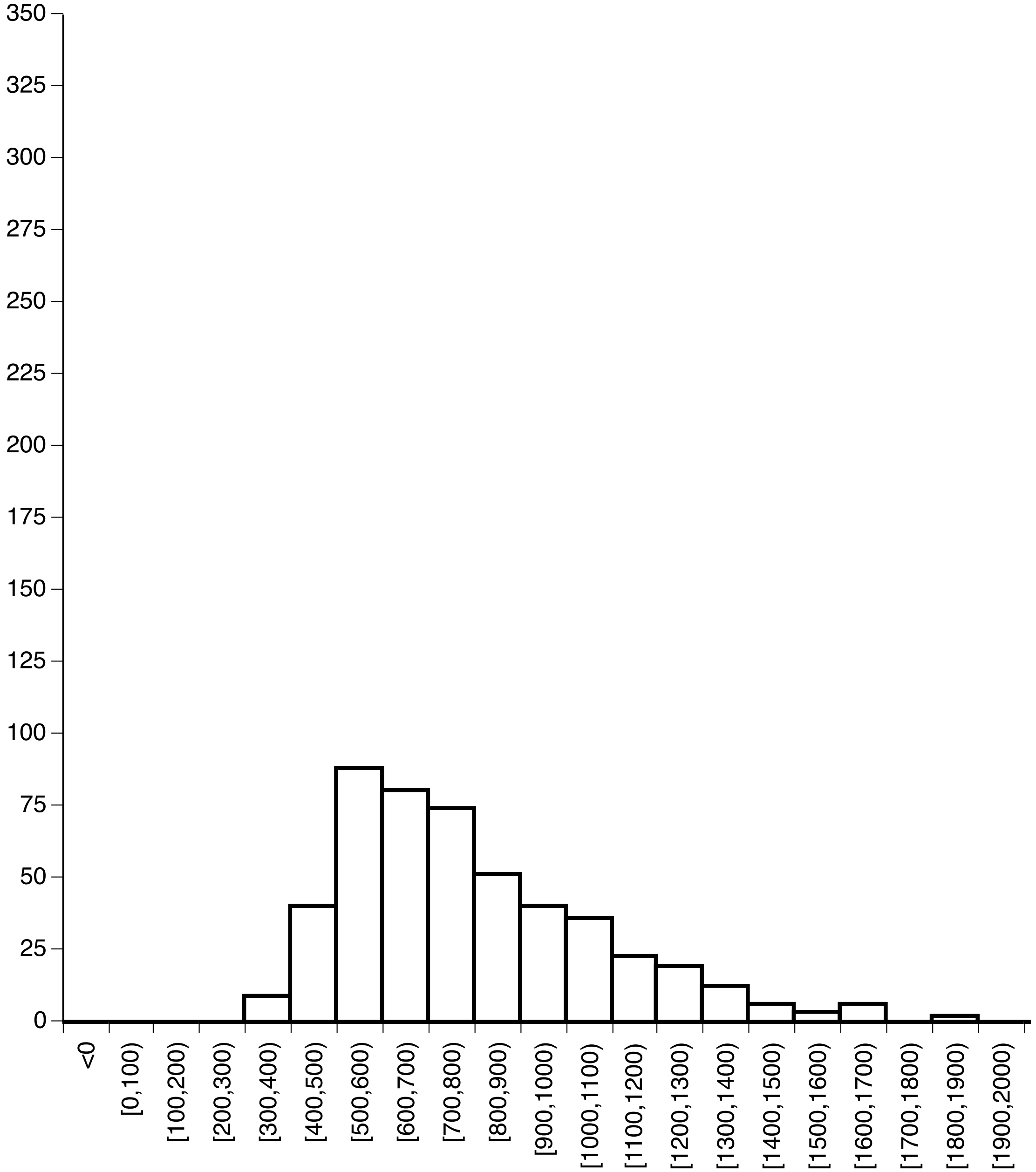
25% Validity, Distance = 4



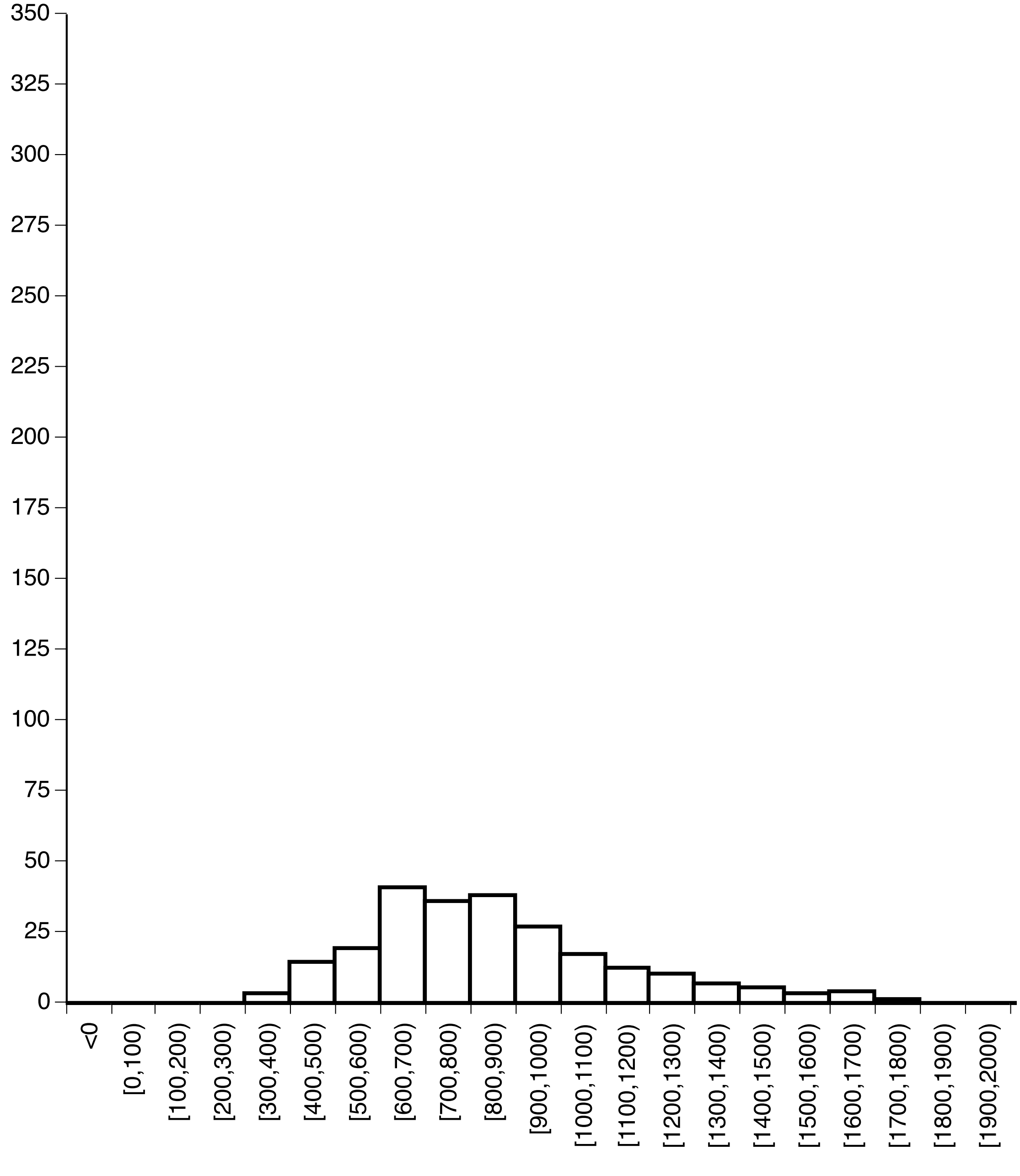
37.5% Validity, Distance = 1



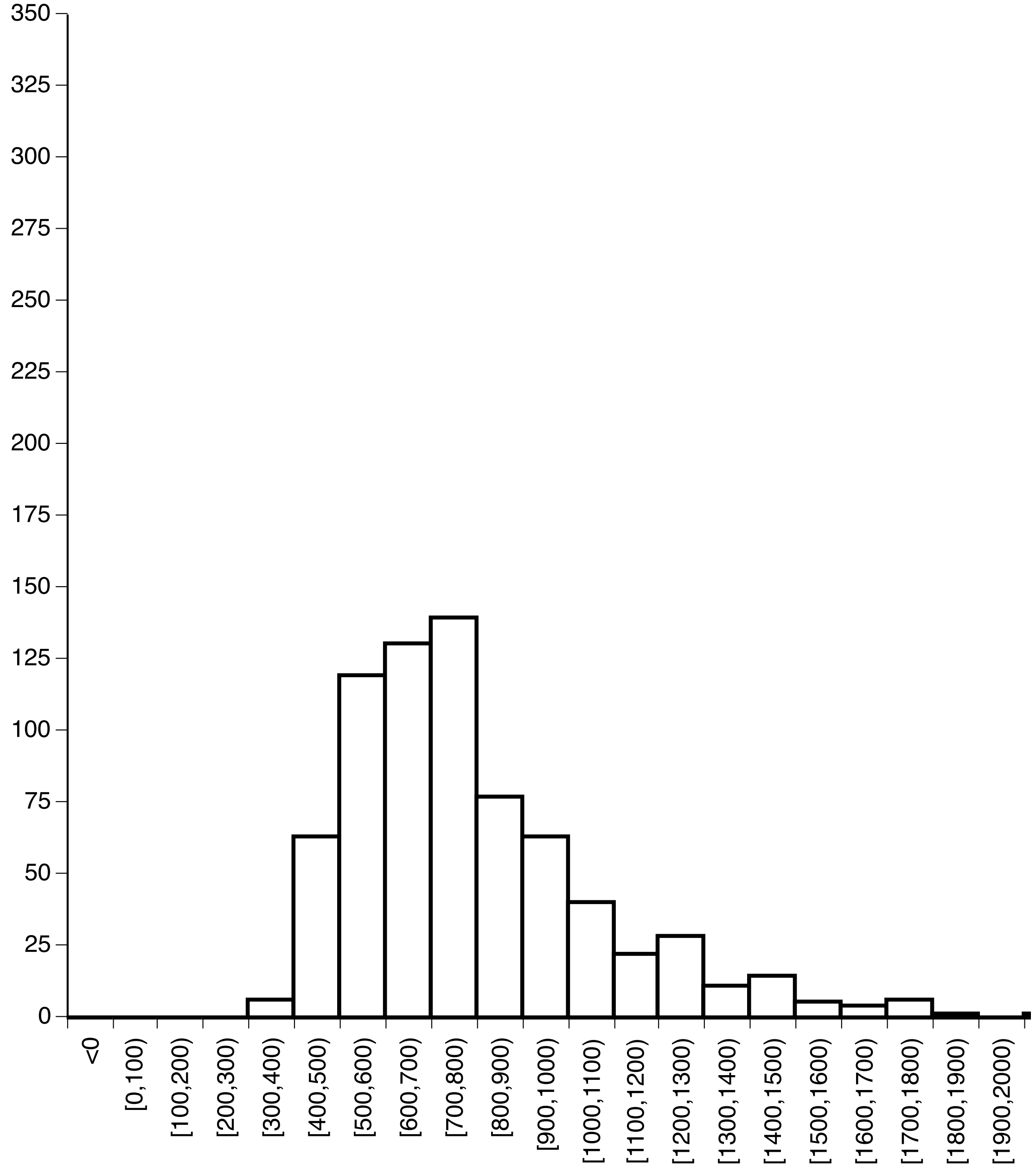
37.5% Validity, Distance = 2



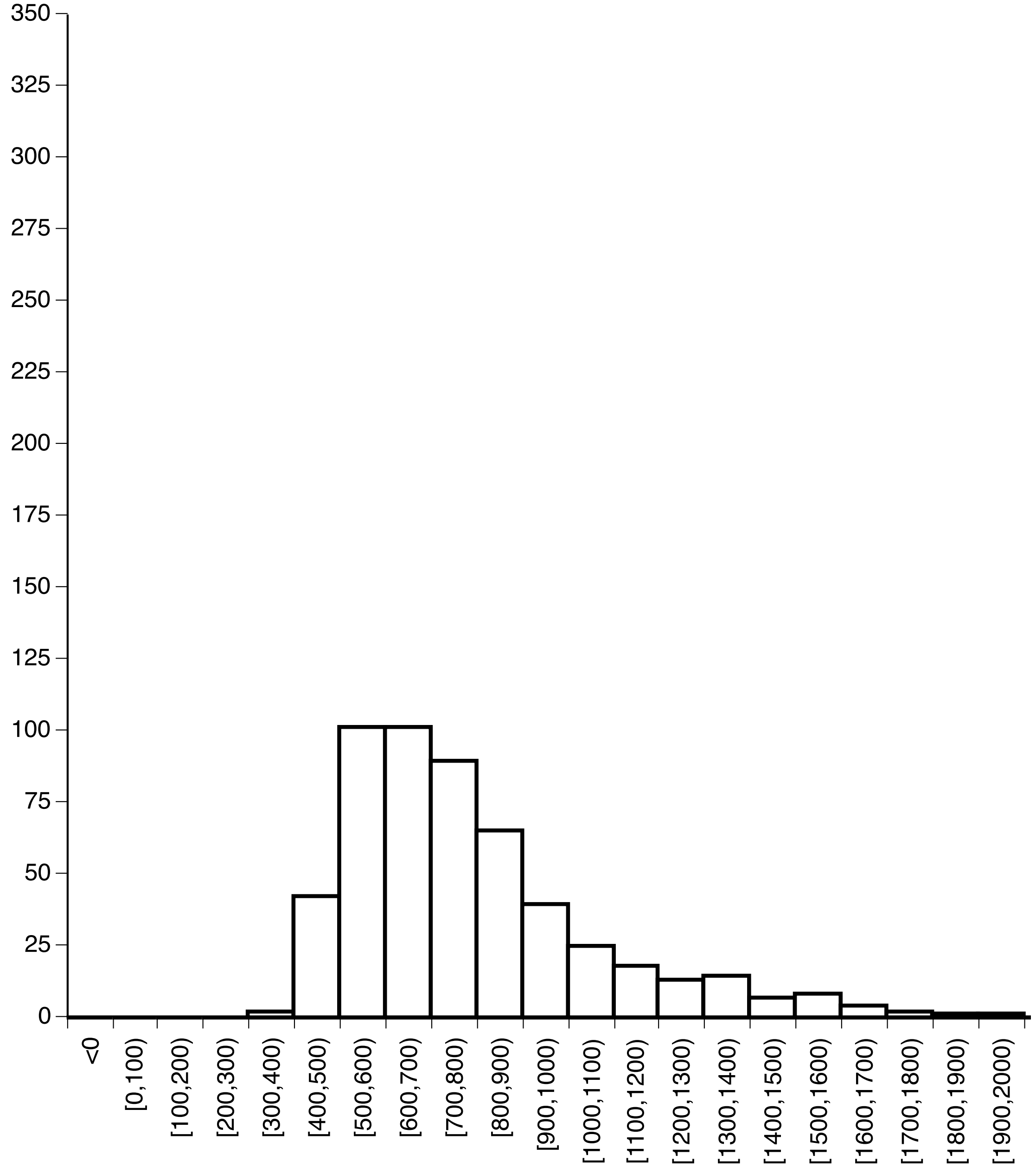
37.5% Validity, Distance = 3



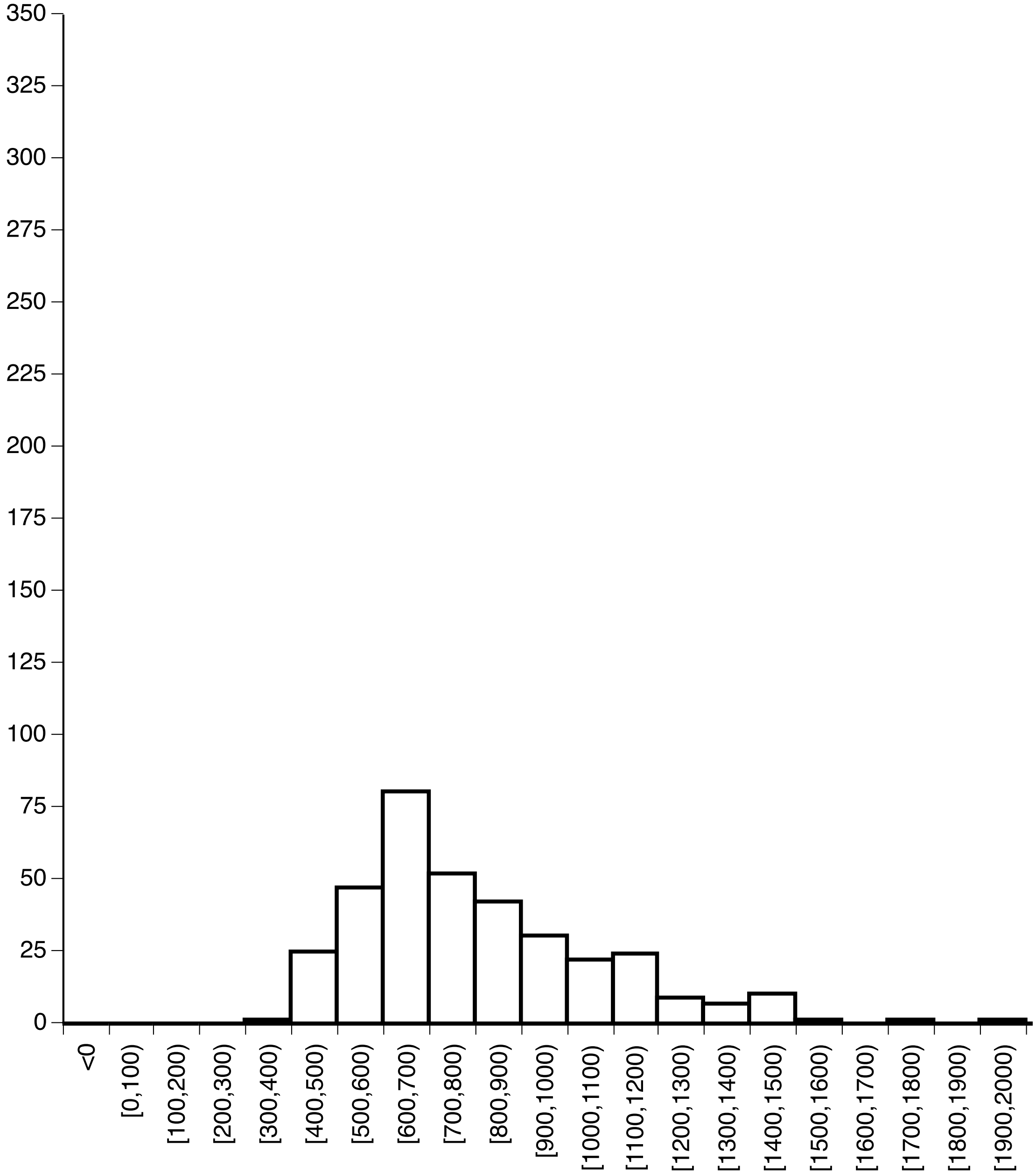
37.5% Validity, Distance = 4



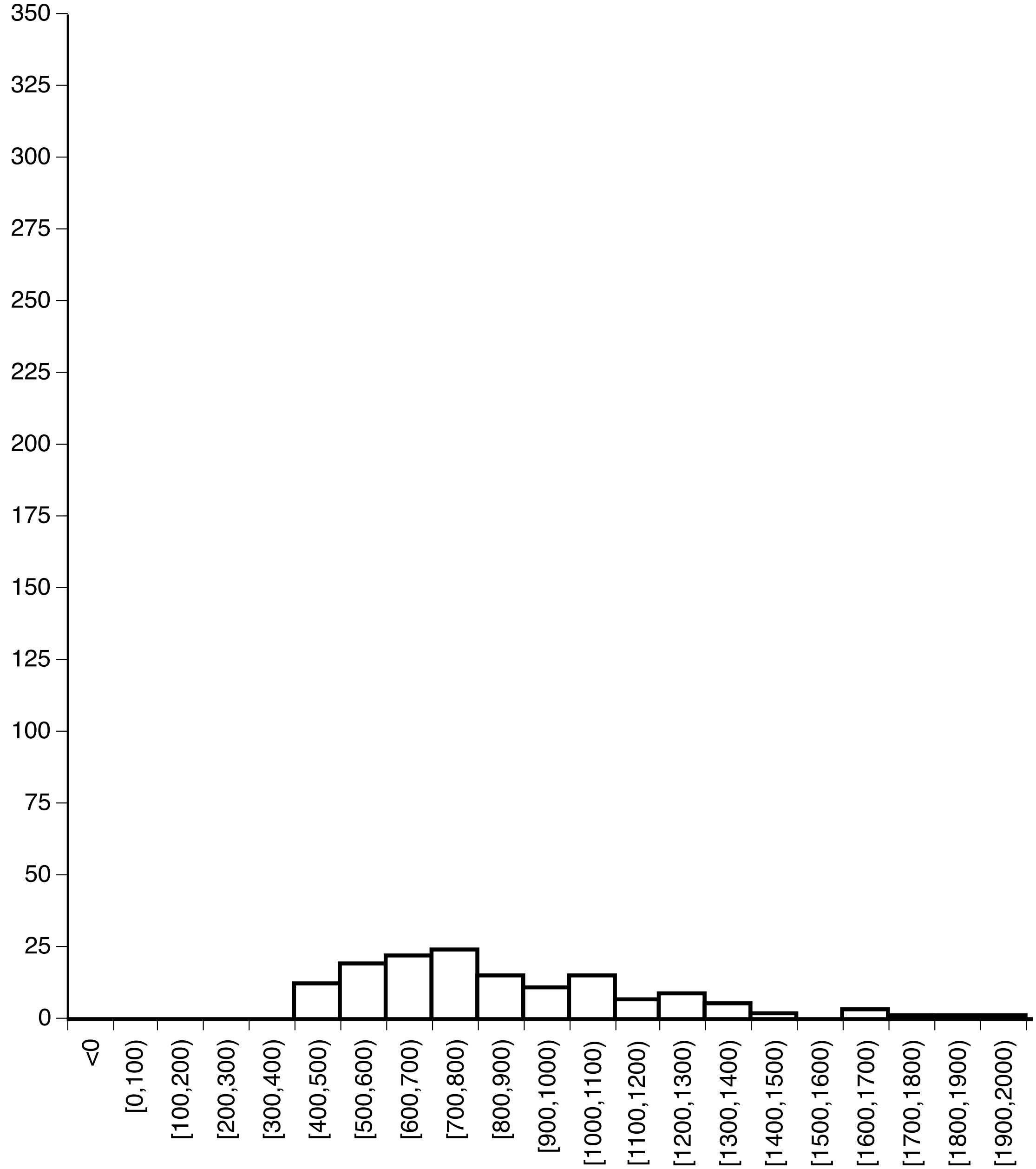
50% Validity, Distance = 1



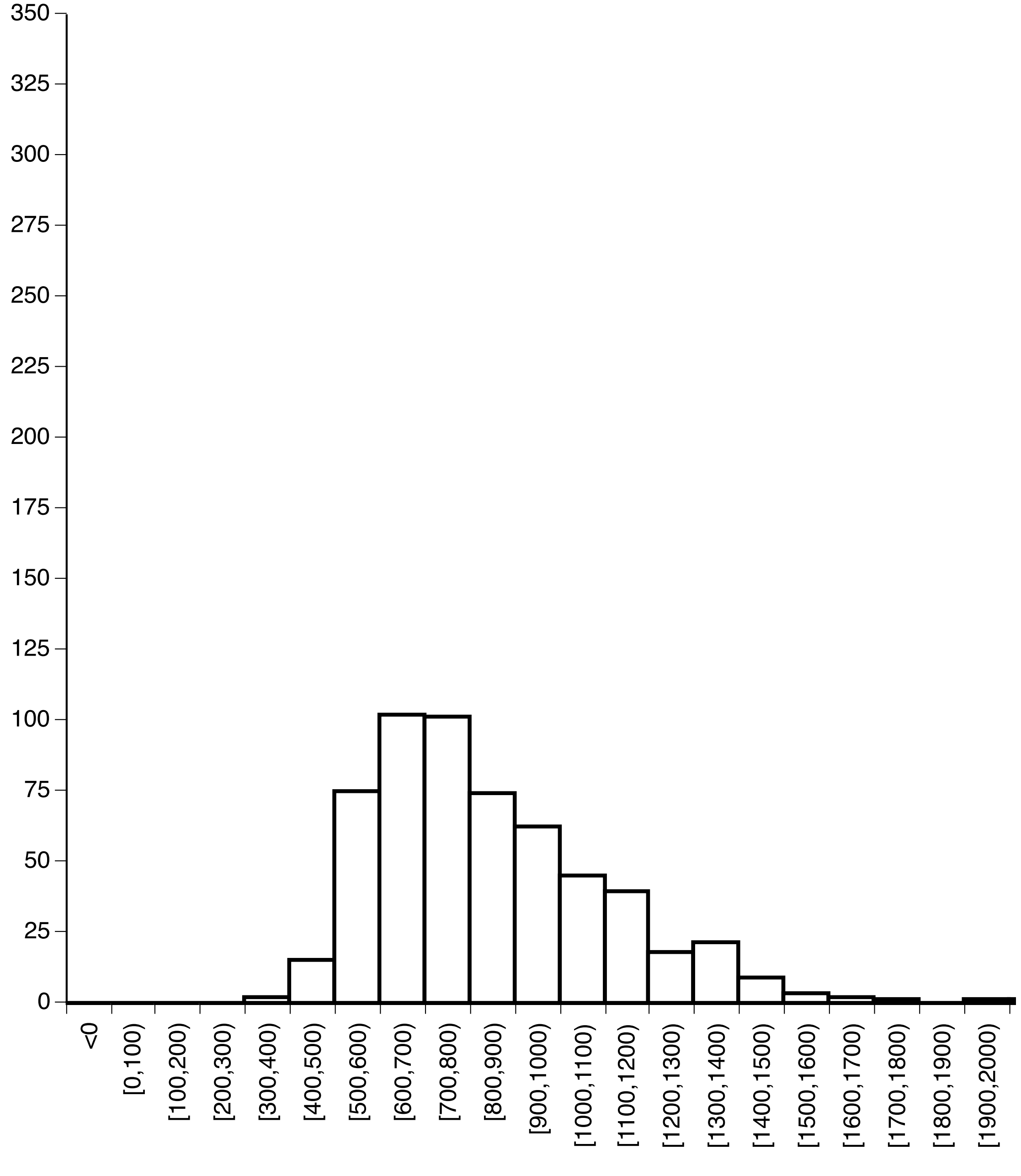
50% Validity, Distance = 2



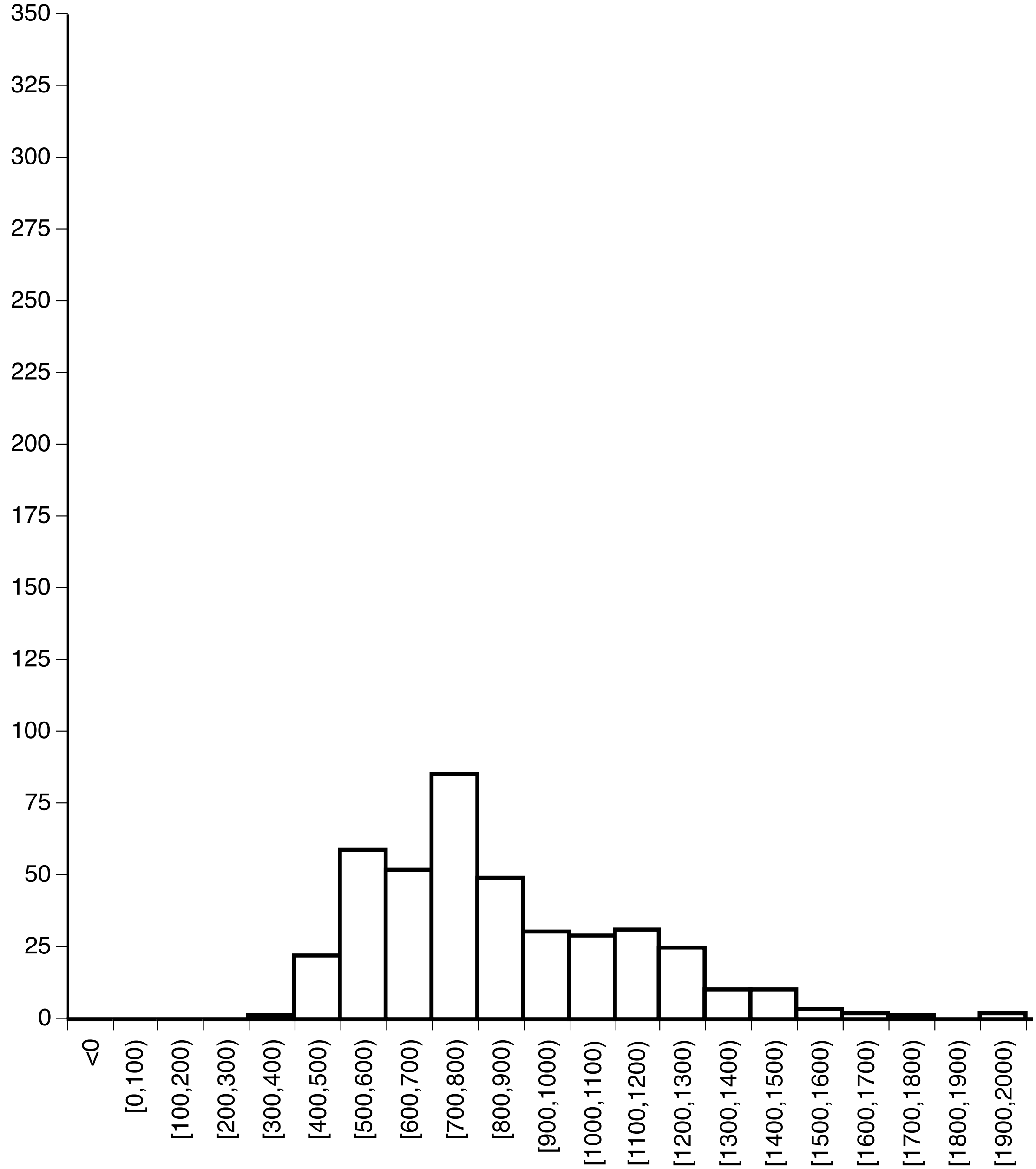
50% Validity, Distance = 3



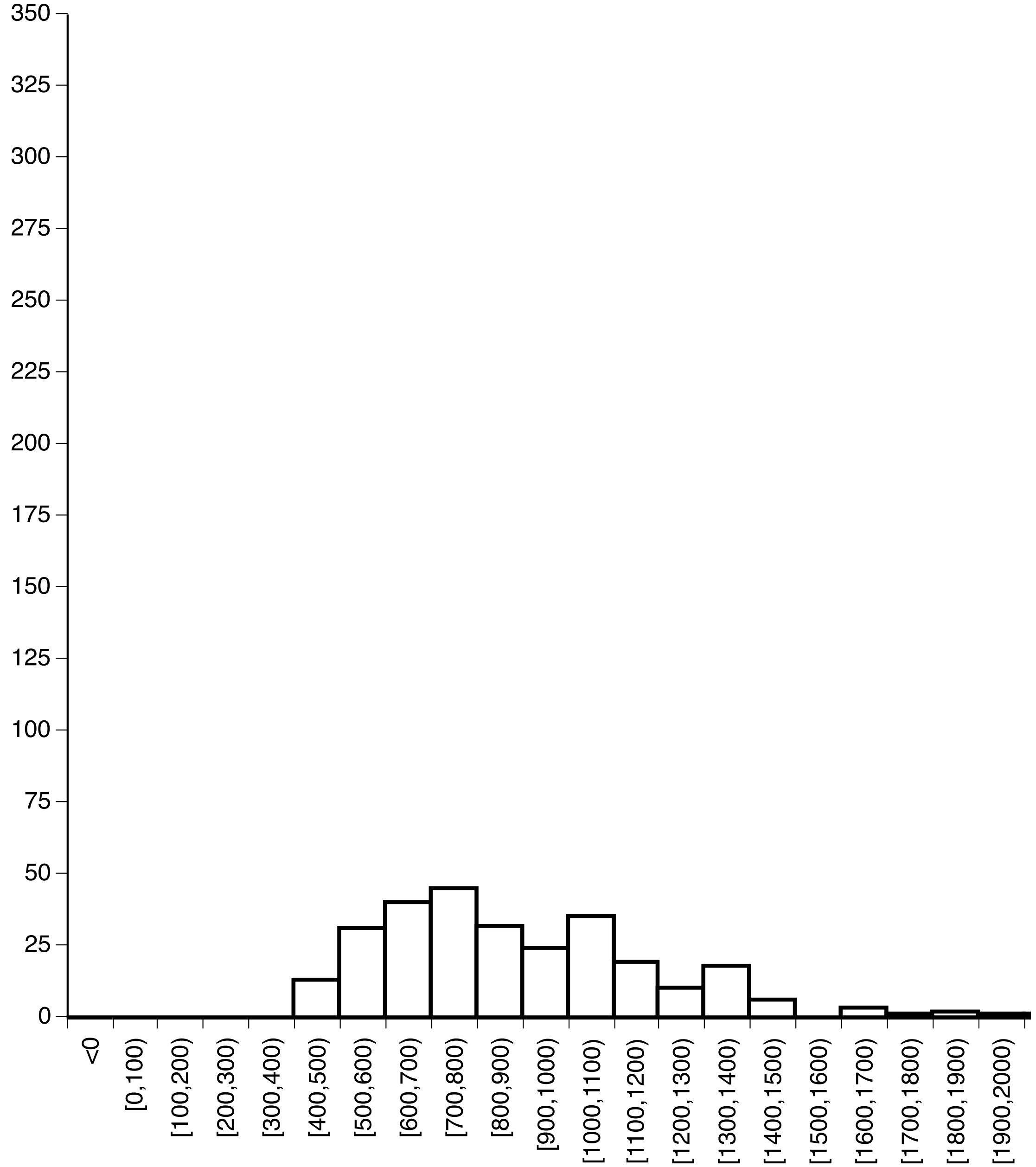
50% Validity, Distance = 4



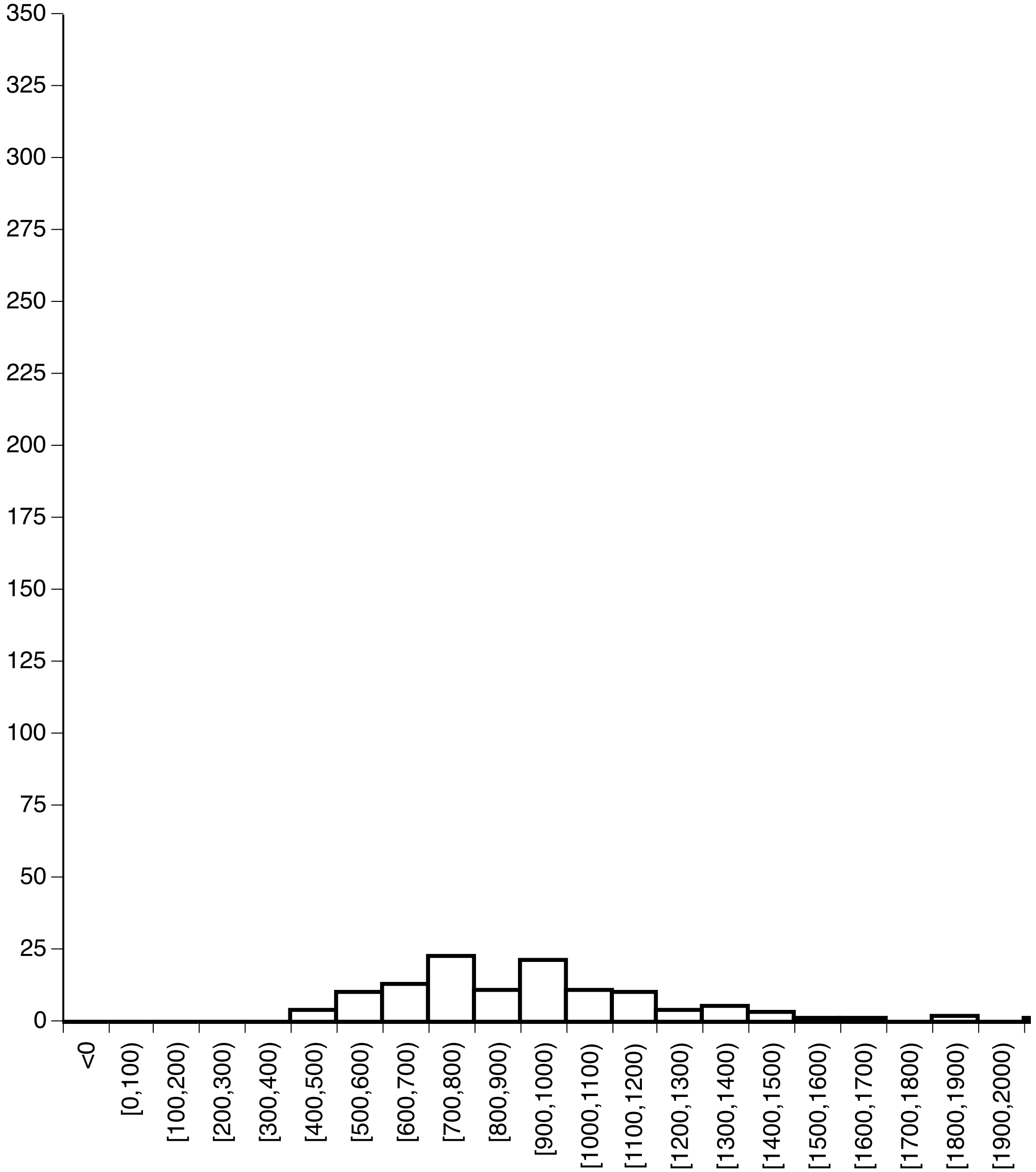
62.5% Validity, Distance = 1



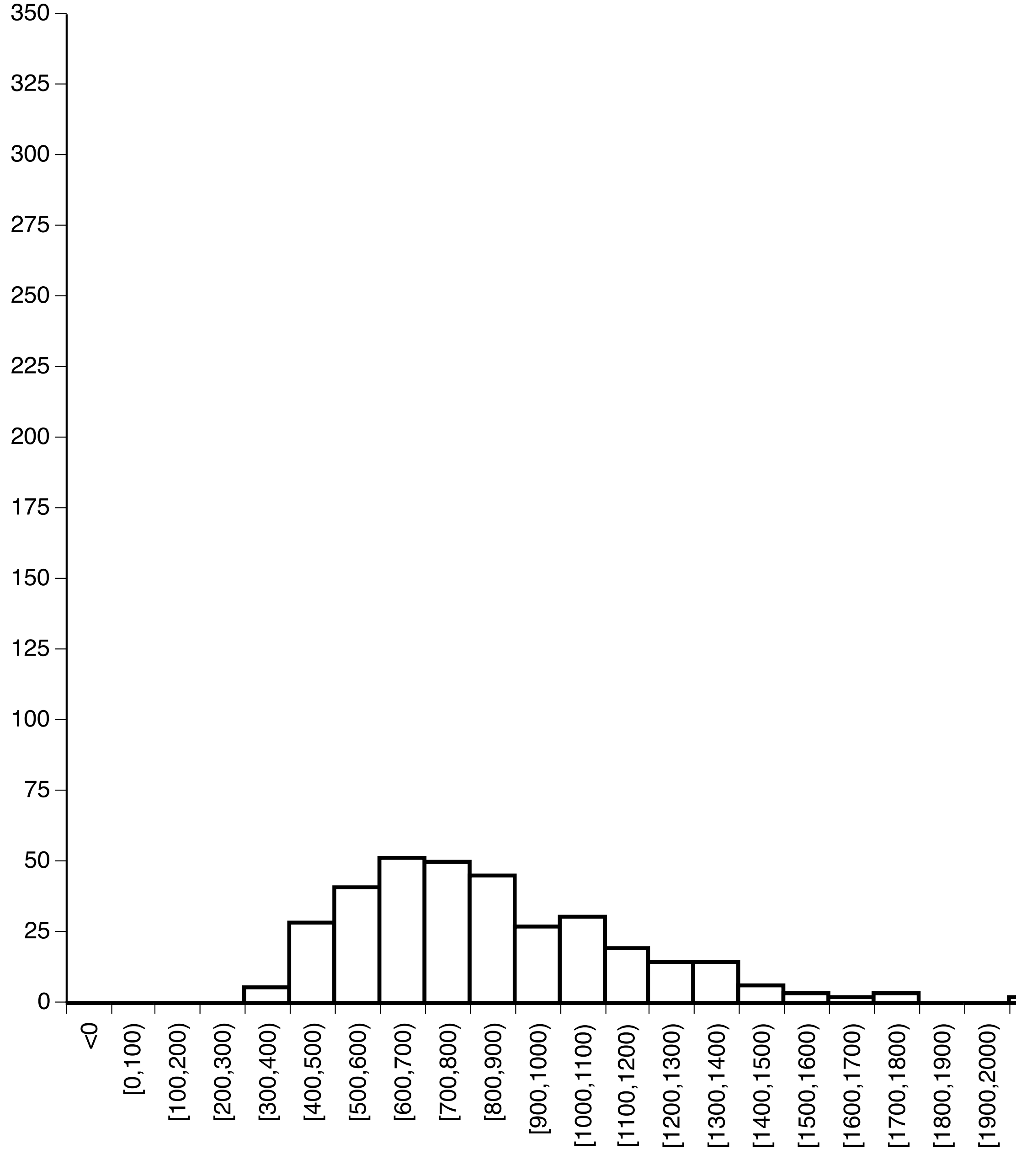
62.5% Validity, Distance = 2



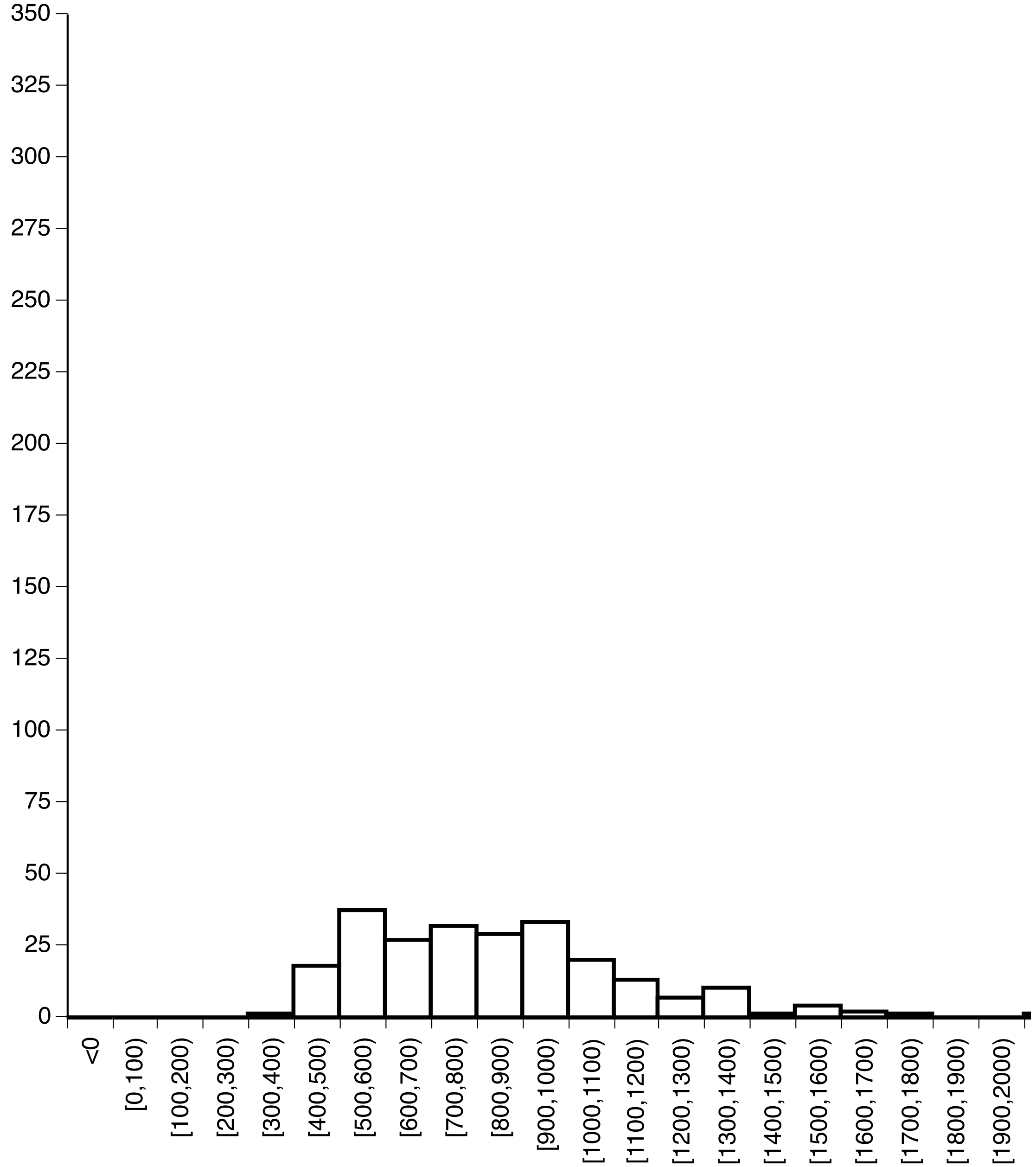
62.5% Validity, Distance = 3



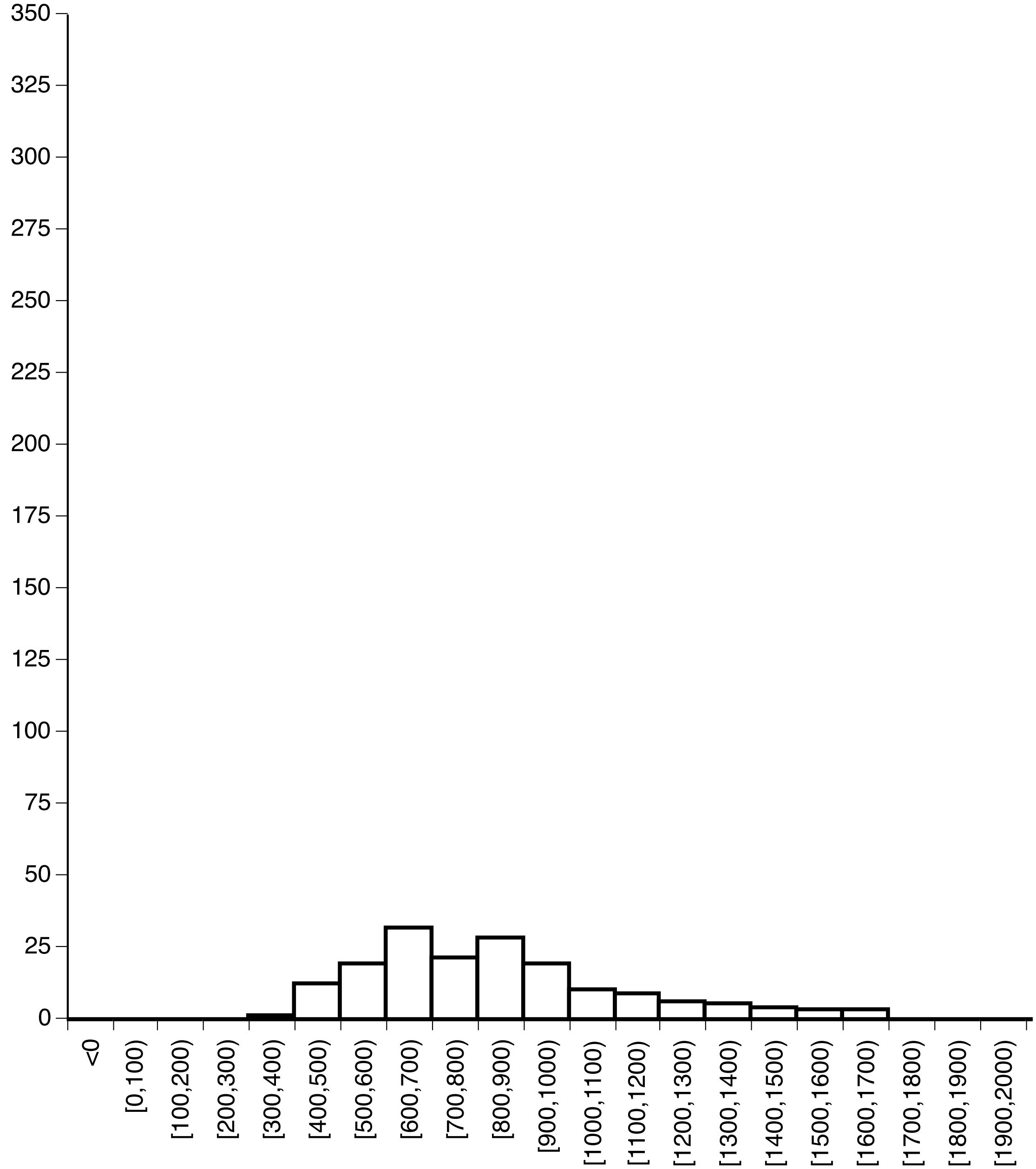
62.5% Validity, Distance = 4



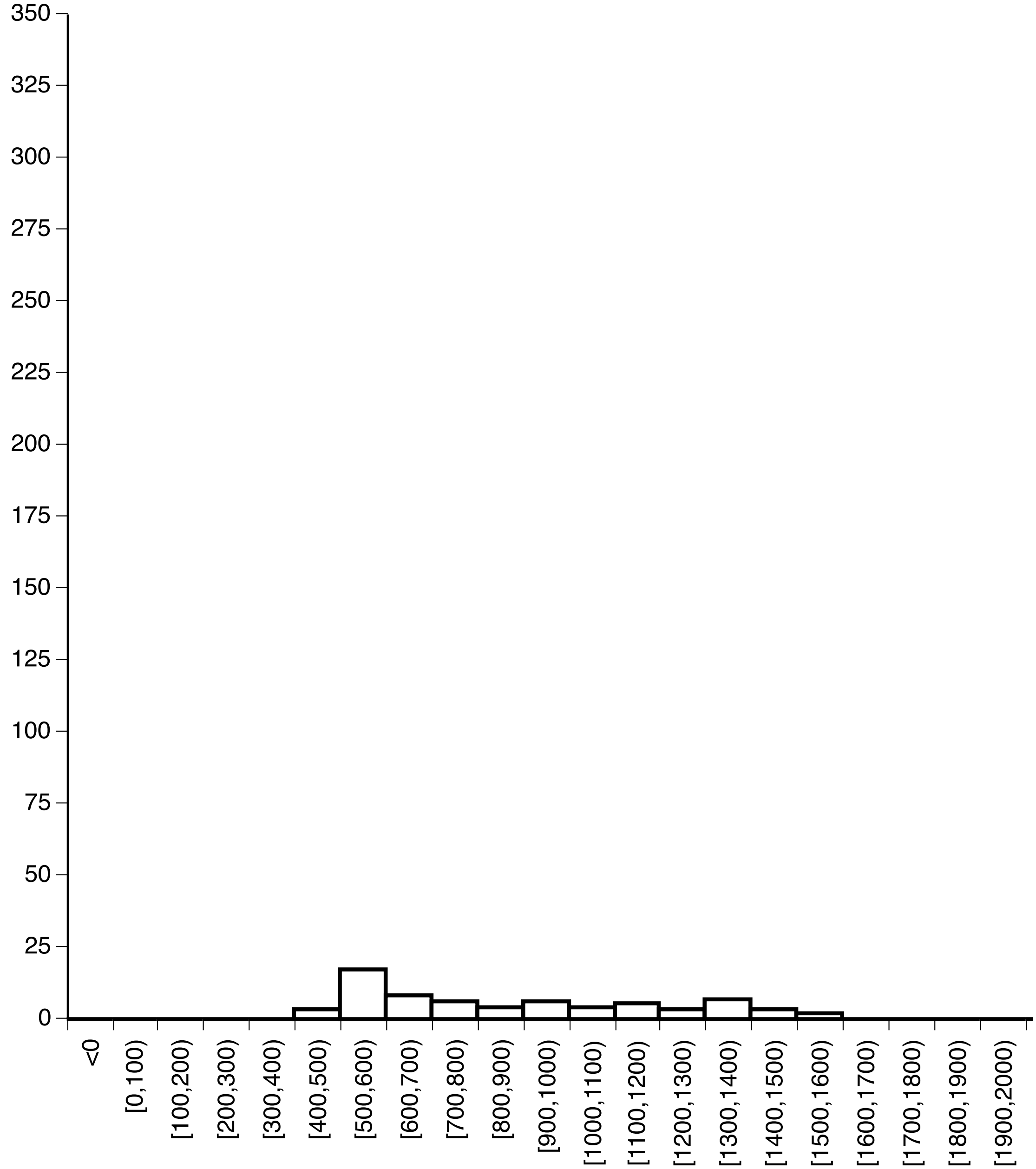
75% Validity, Distance = 1



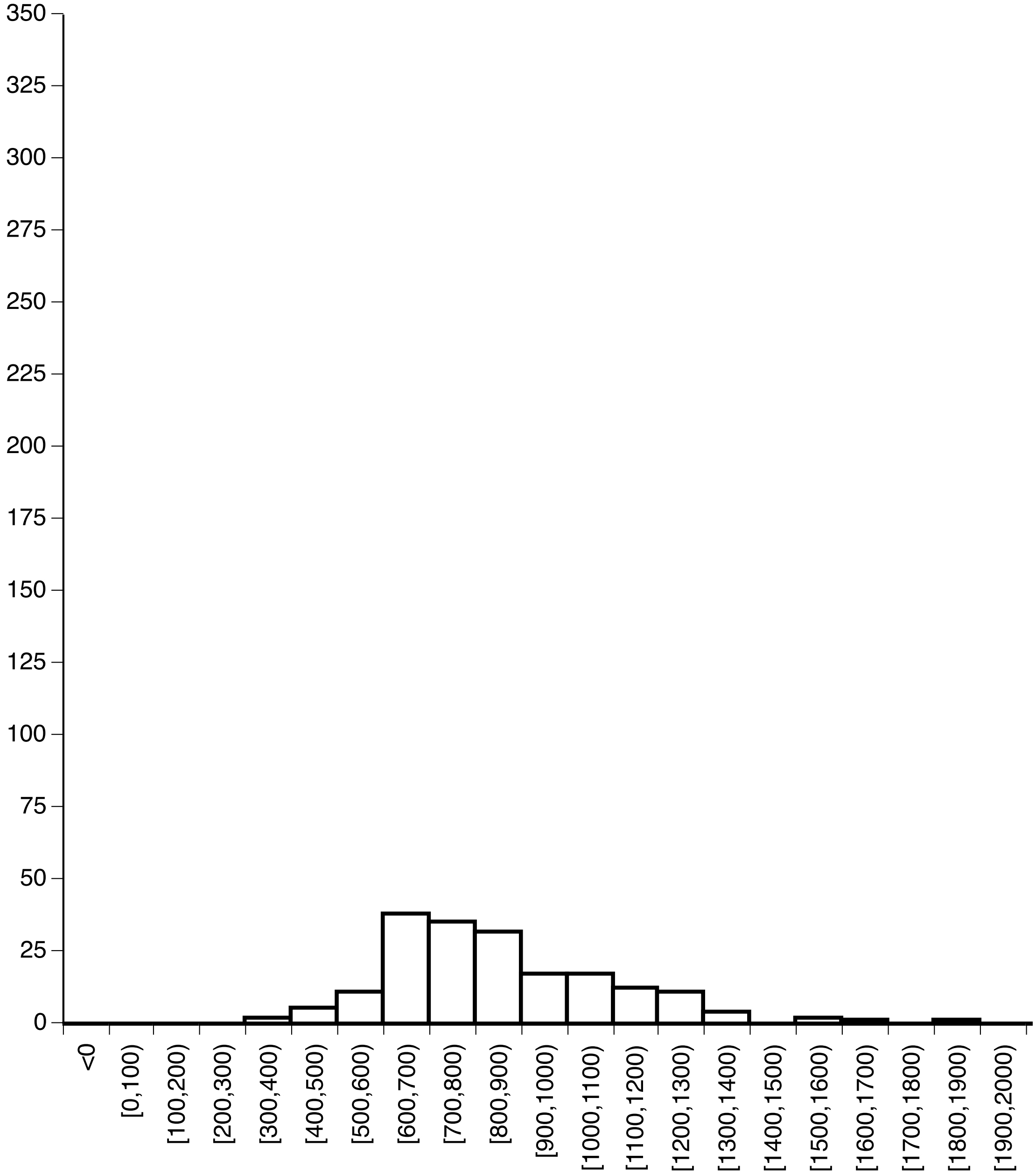
75% Validity, Distance = 2



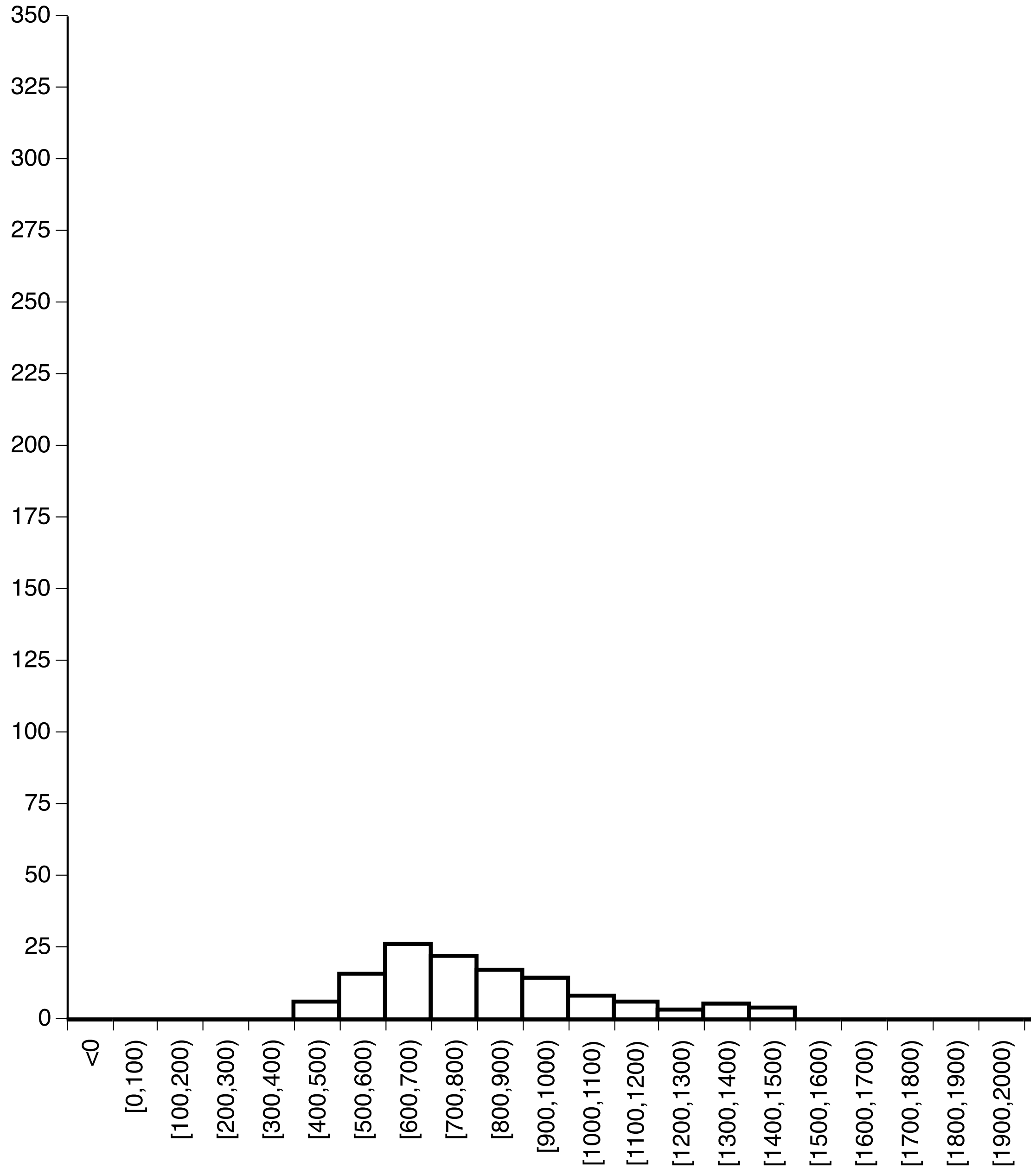
75% Validity, Distance = 3



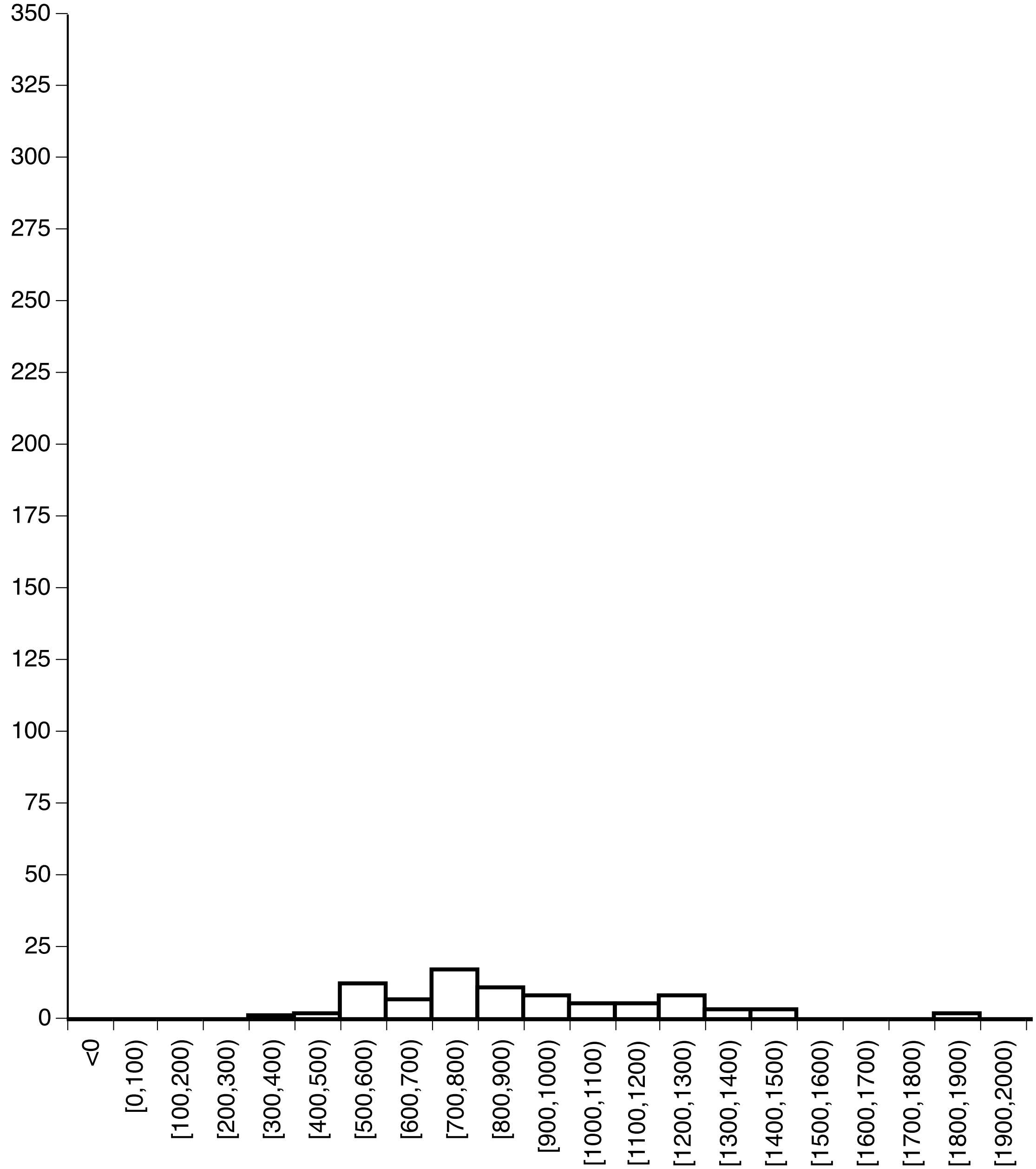
75% Validity, Distance = 4



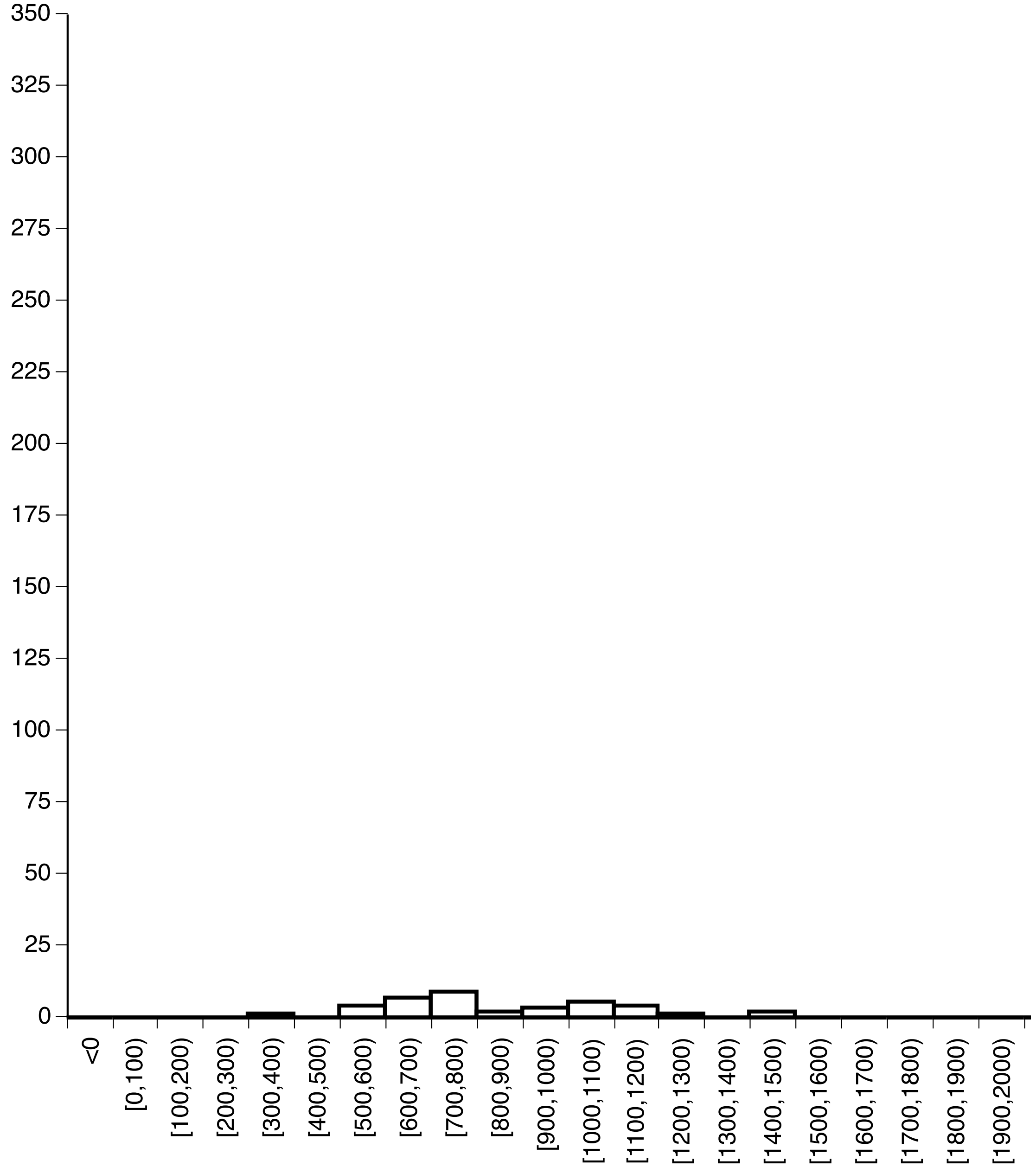
87.5% Validity, Distance = 1



87.5% Validity, Distance = 2



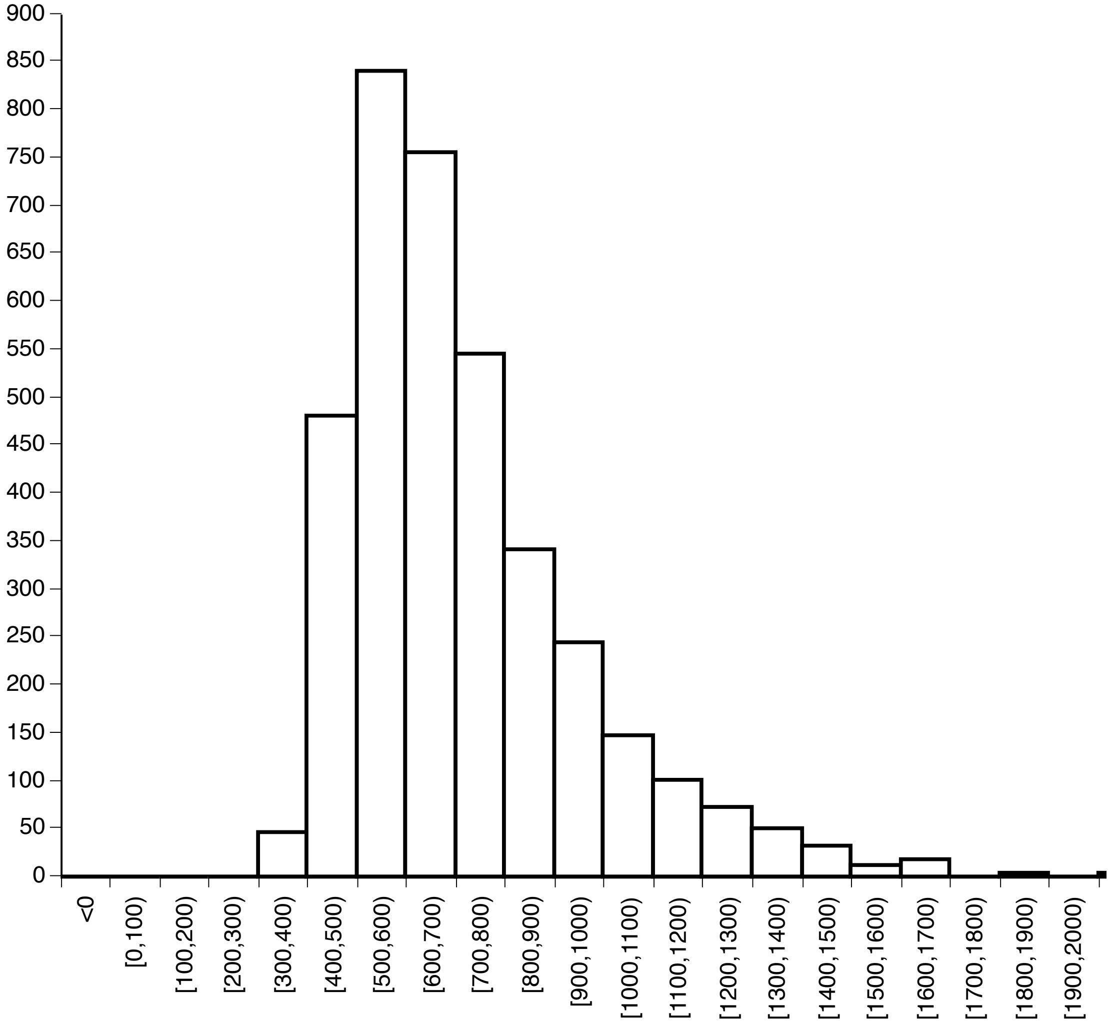
87.5% Validity, Distance = 3



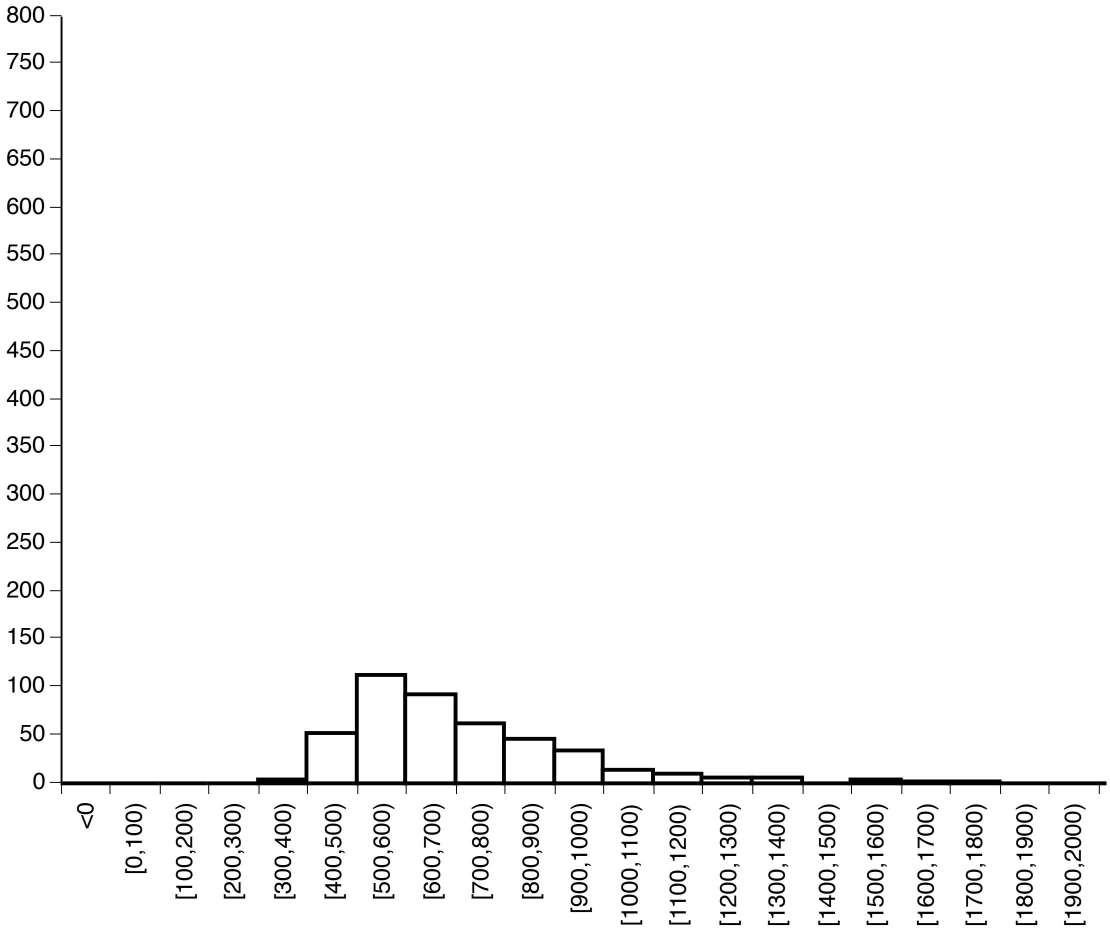
87.5% Validity, Distance = 4

APPENDIX B: HISTOGRAMS OF HUMAN AND ACT-R MODEL DATA

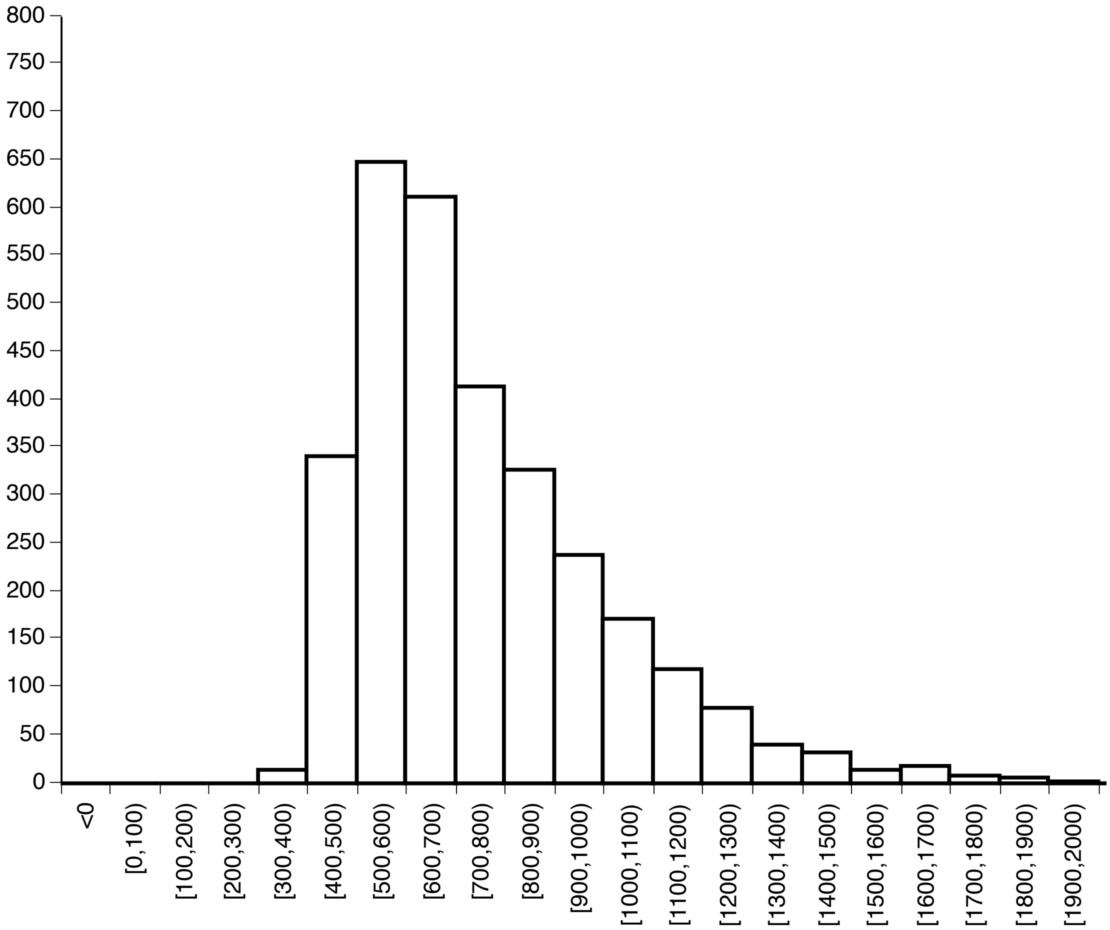
All histograms plot number of trials on the y axis and RT bins on the x axis.



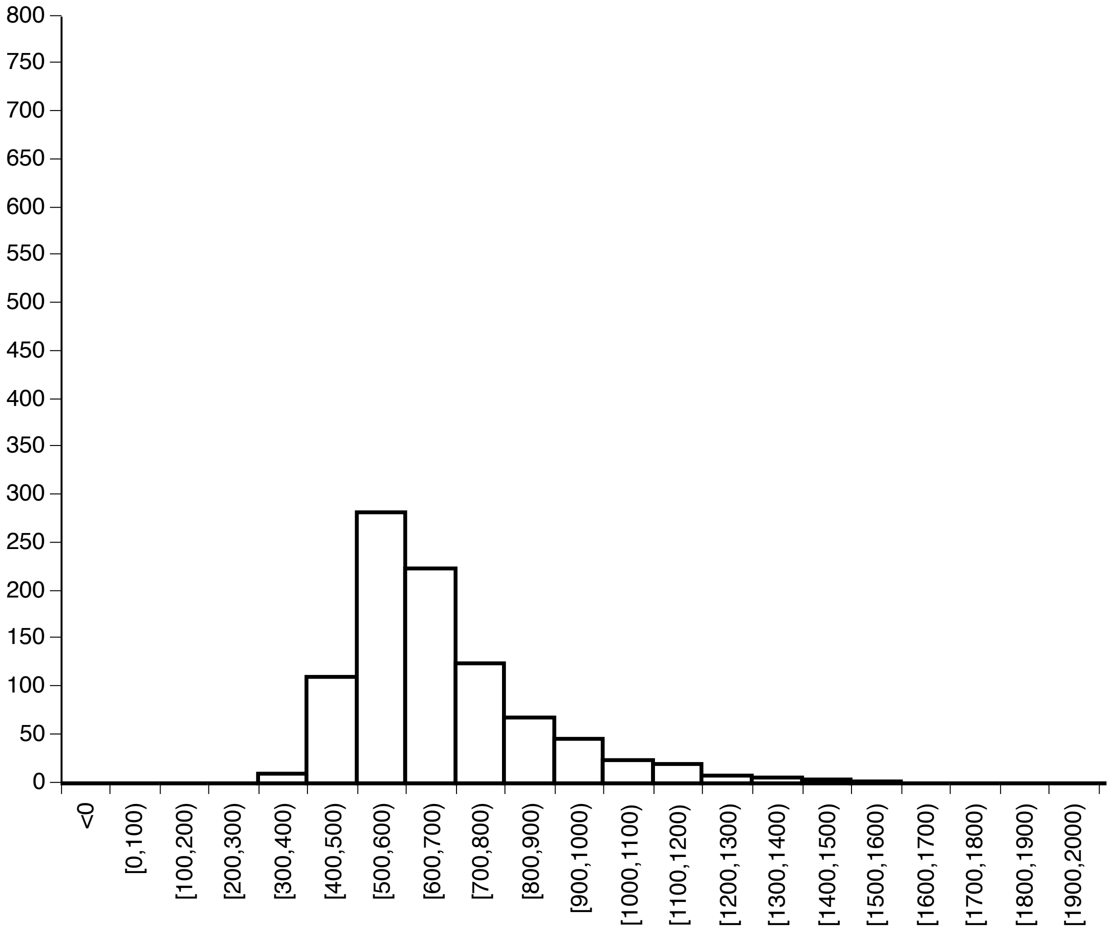
Human Data: 0% Validity, Invalid Trial Type



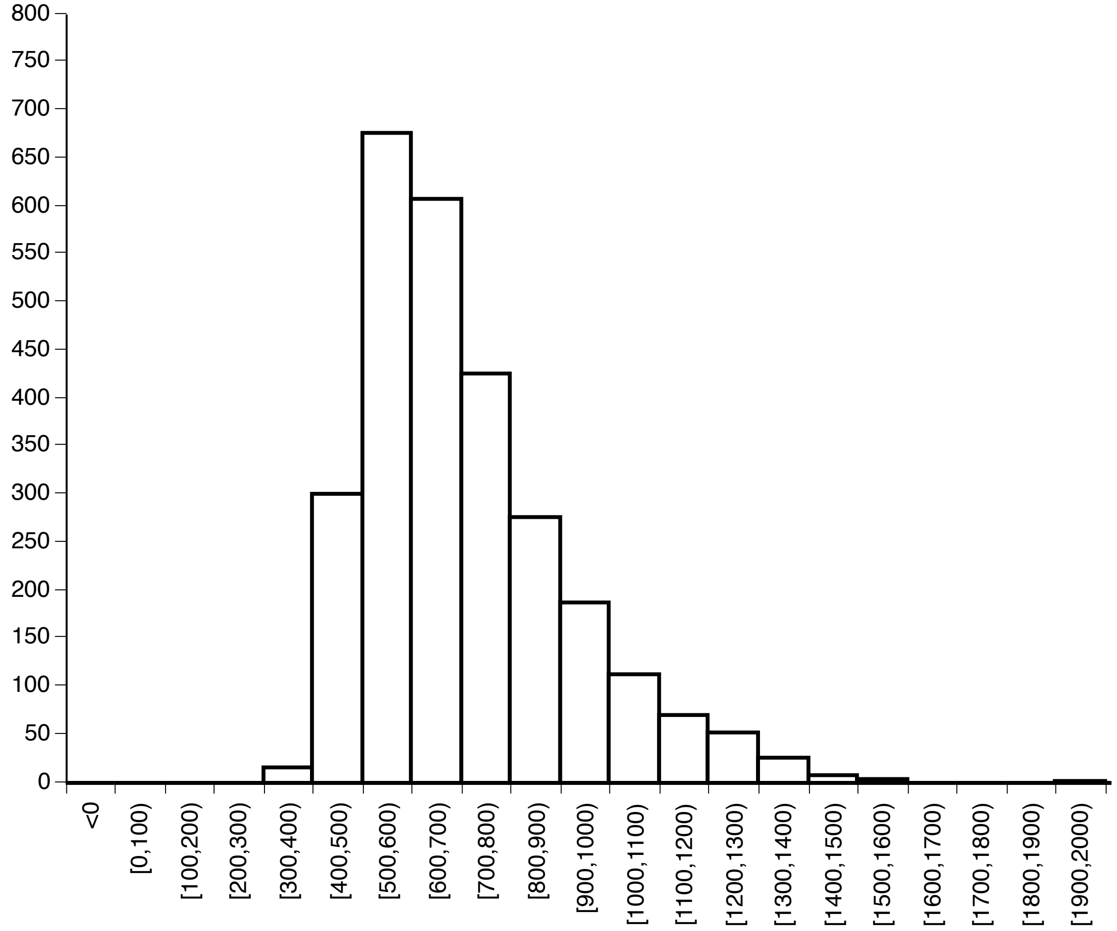
Human Data: 12.5% Validity, Valid Trial Type



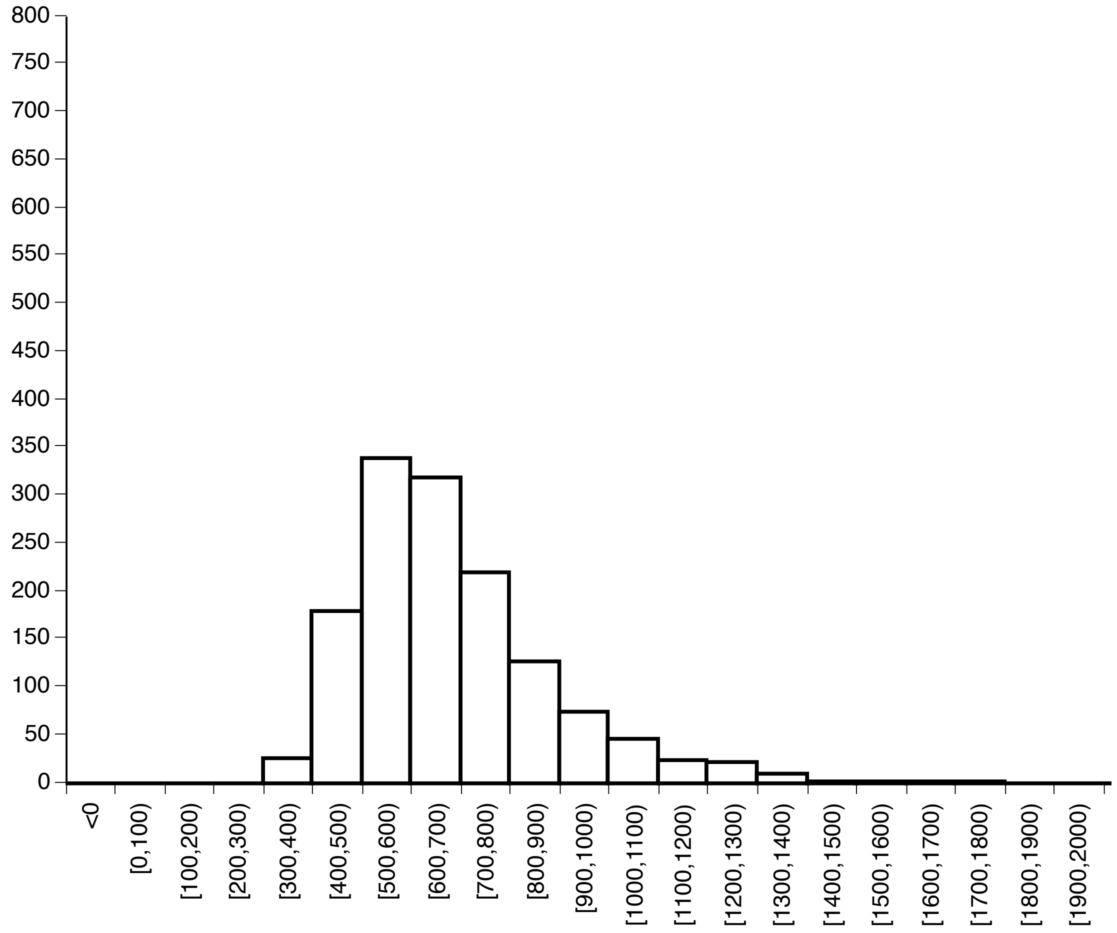
Human Data: 12.5% Validity, Invalid Trial Type



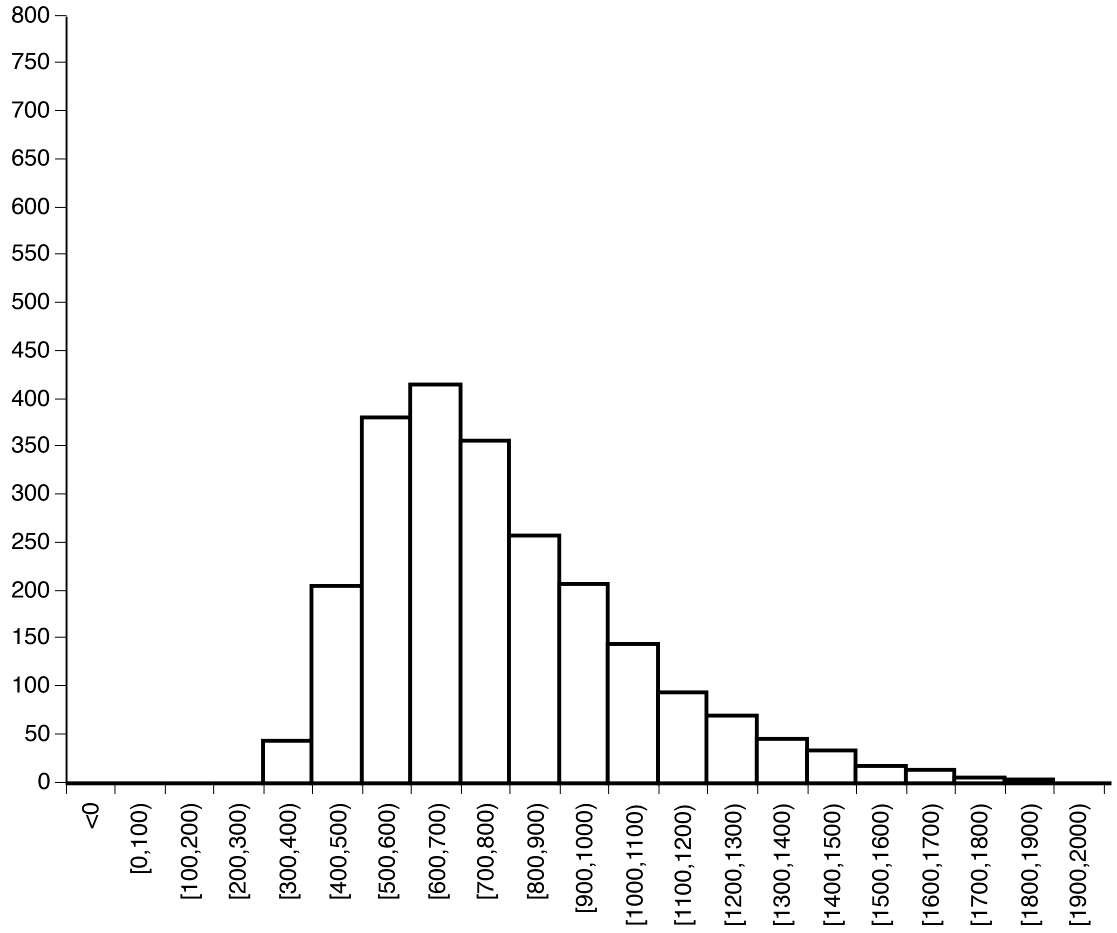
Human Data: 25% Validity, Valid Trial Types



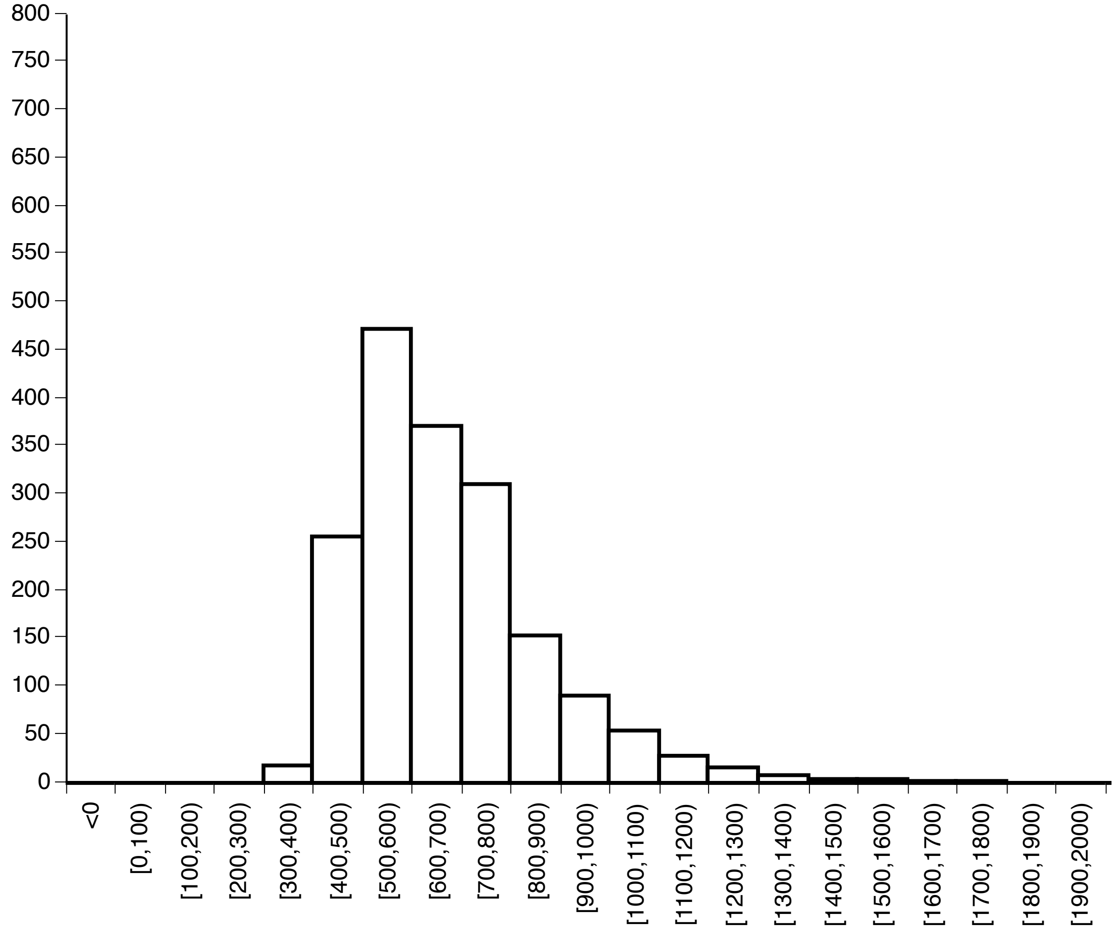
Human Data: 25% Validity, Invalid Trial Types



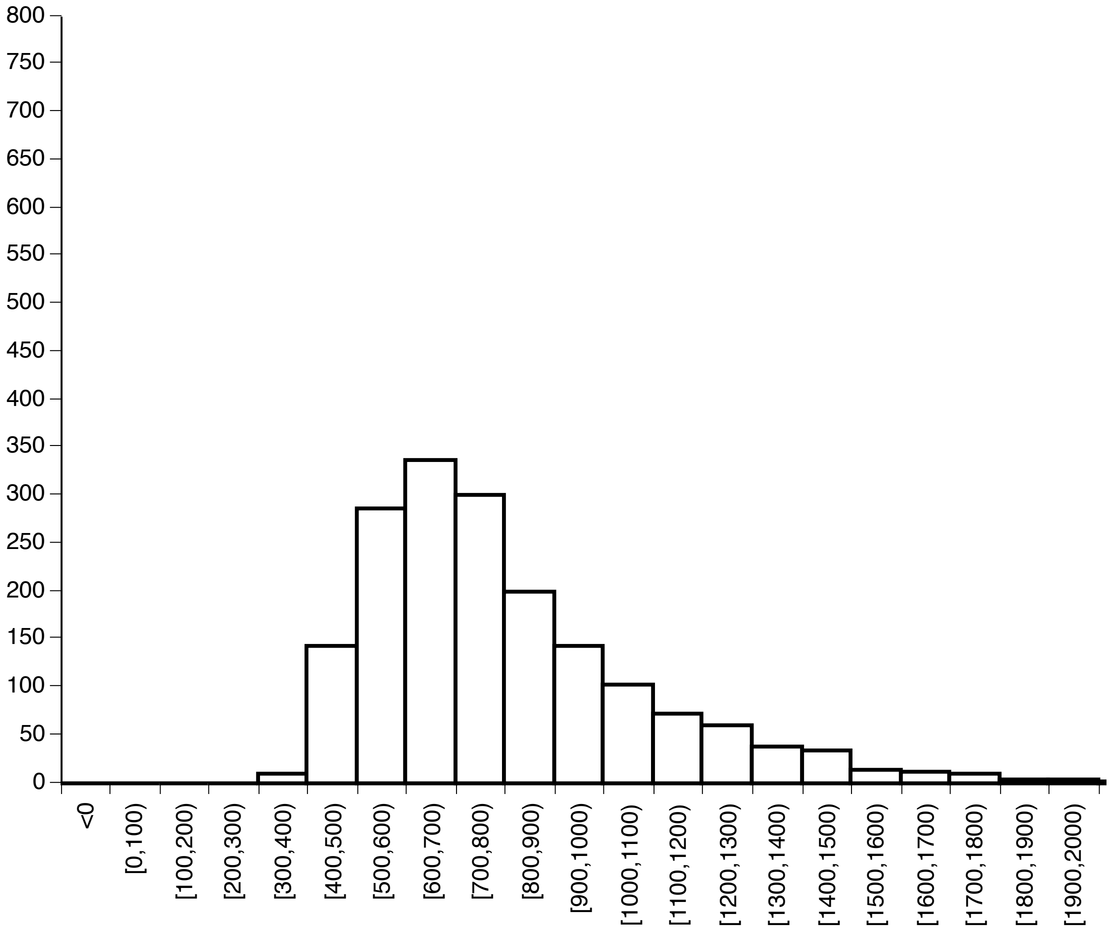
Human Data: 37.5% Validity, Valid Trial Types



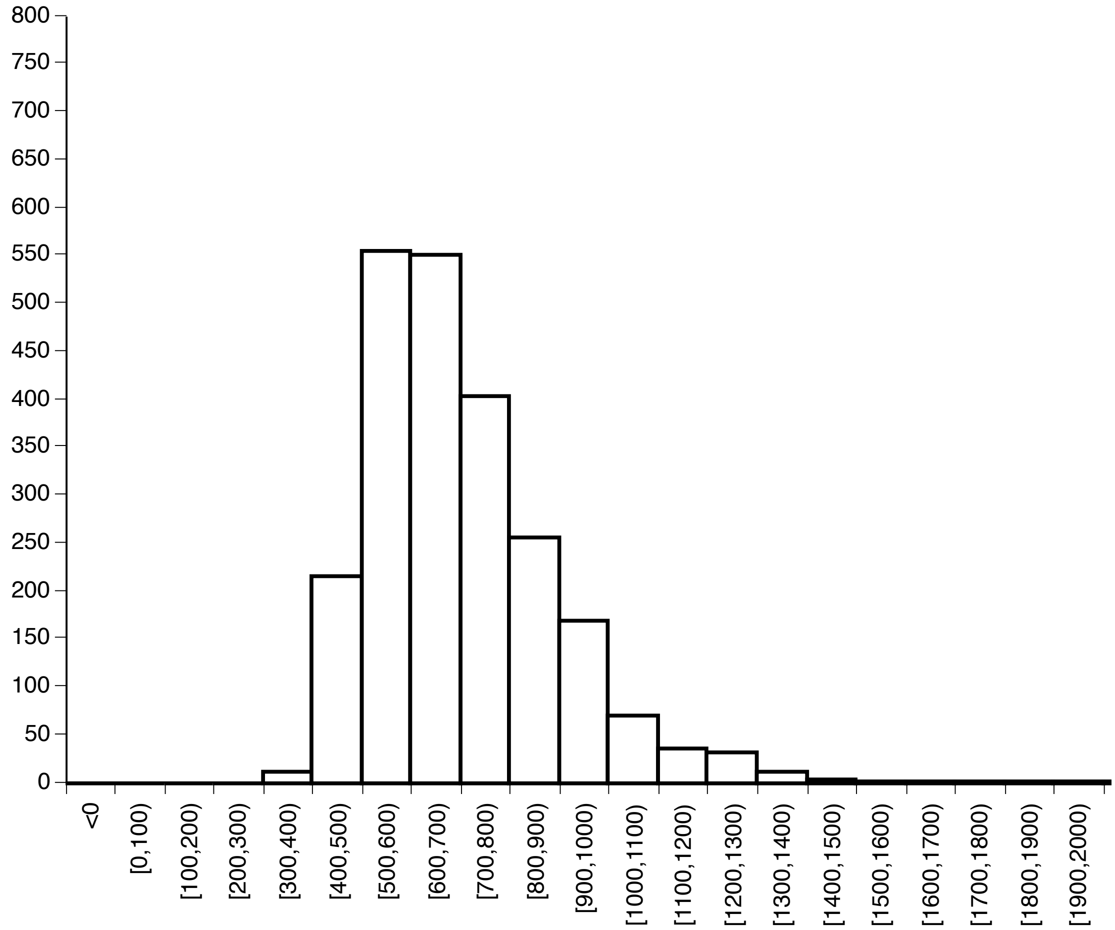
Human Data: 37.5% Validity, Invalid Trial Types



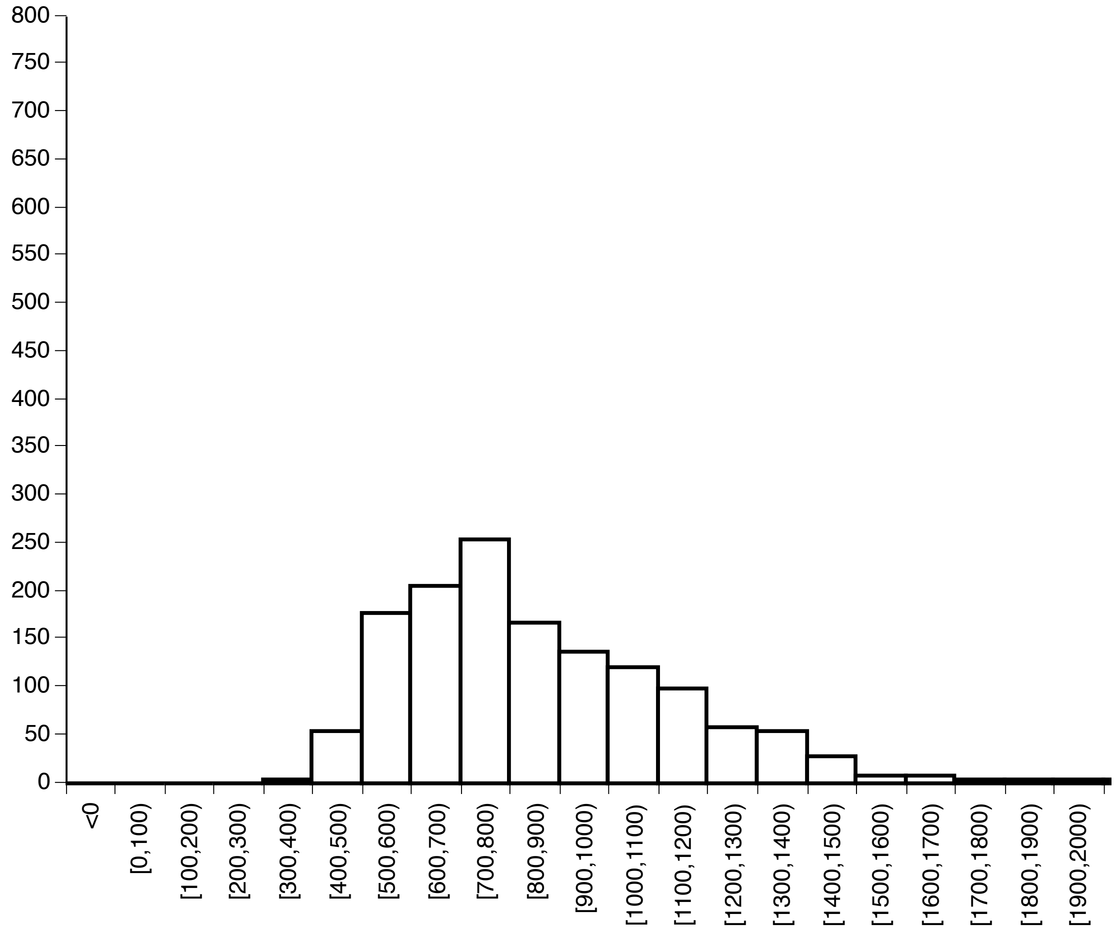
Human Data: 50% Validity, Valid Trial Types



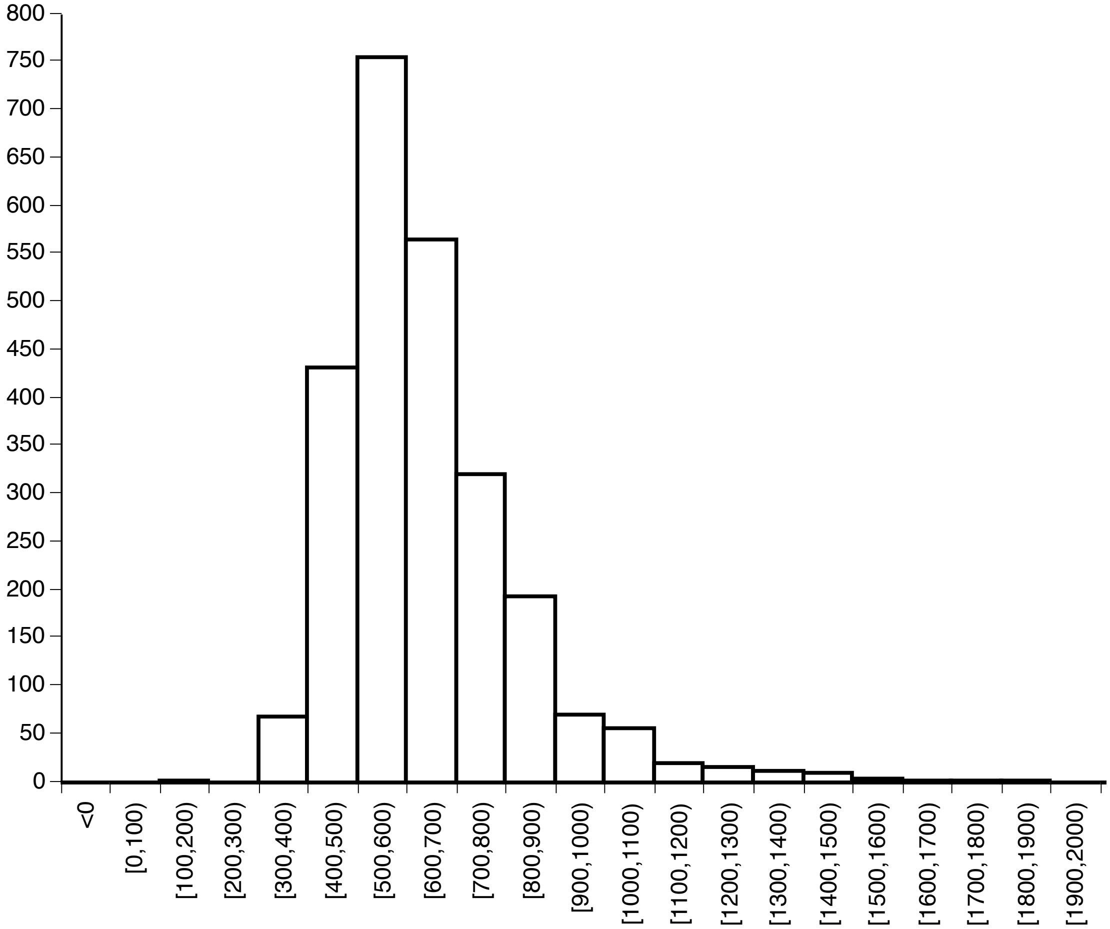
Human Data: 50% Validity, Invalid Trial Types



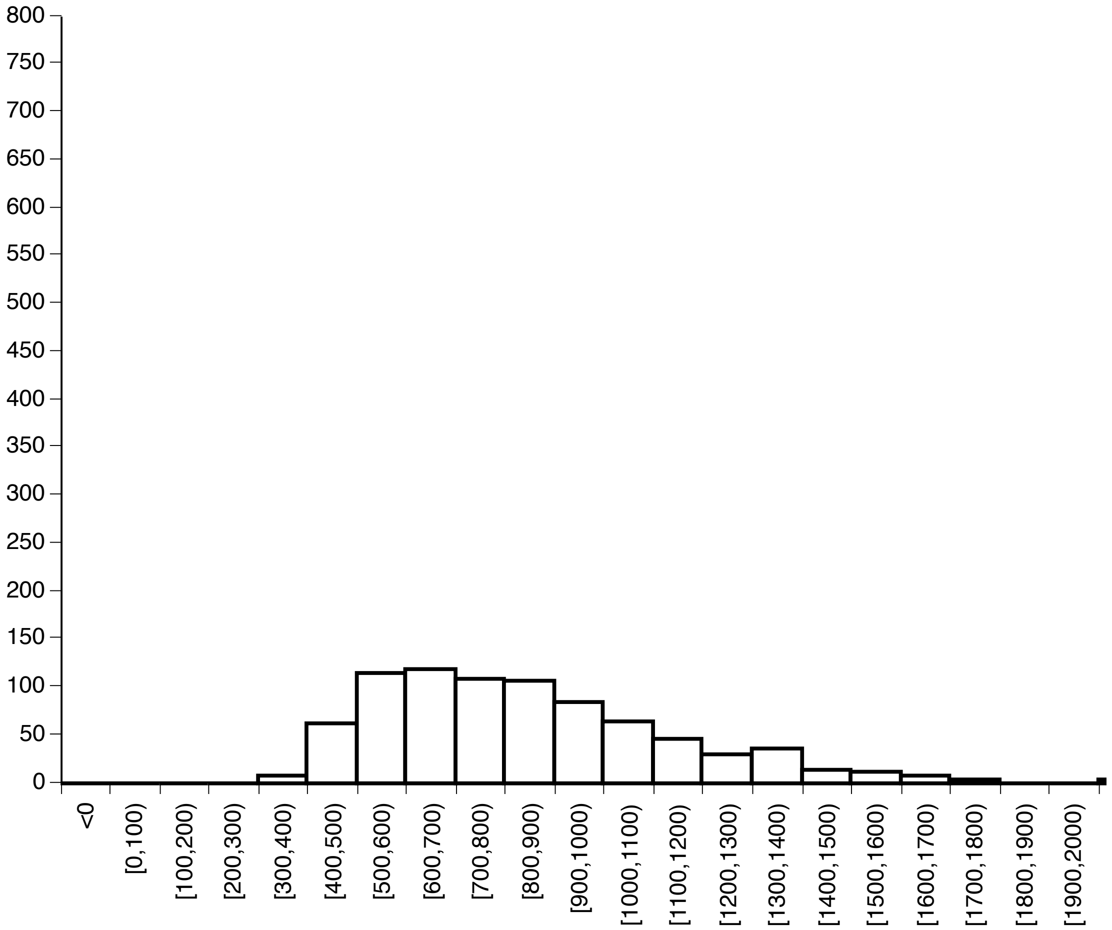
Human Data: 62.5% Validity, Valid Trial Types



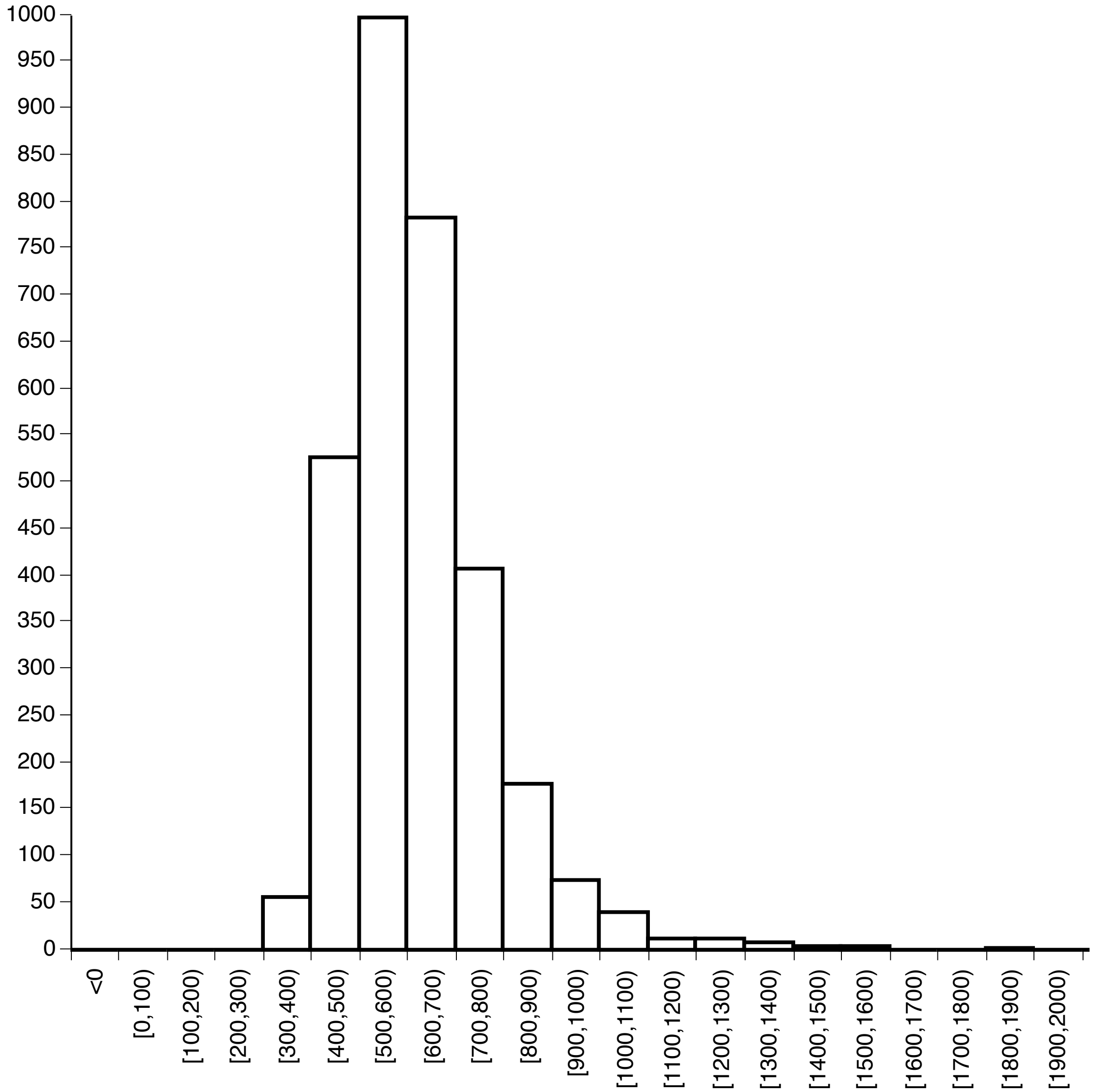
Human Data: 62.5% Validity, Invalid Trial Types



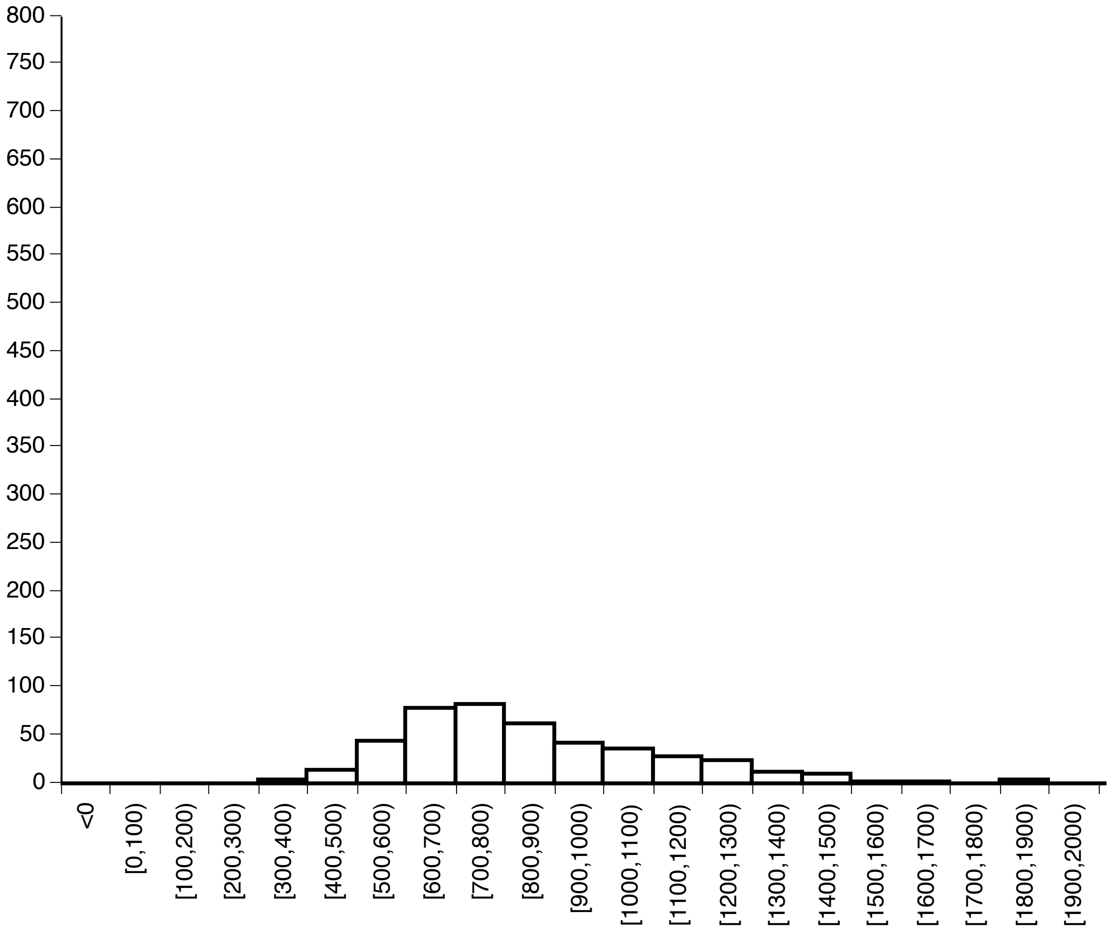
Human Data: 75% Validity, Valid Trial Types



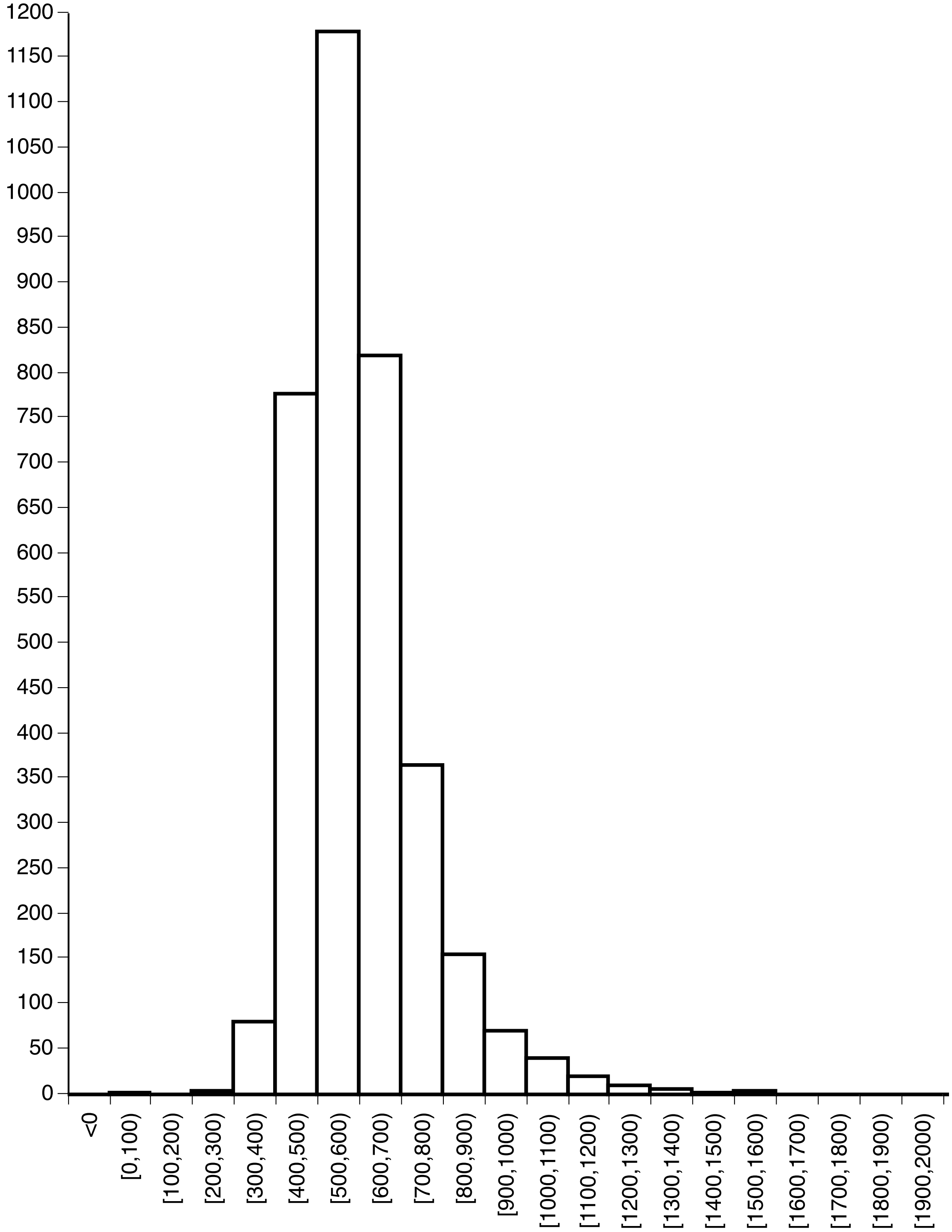
Human Data: 75% Validity, Invalid Trial Types



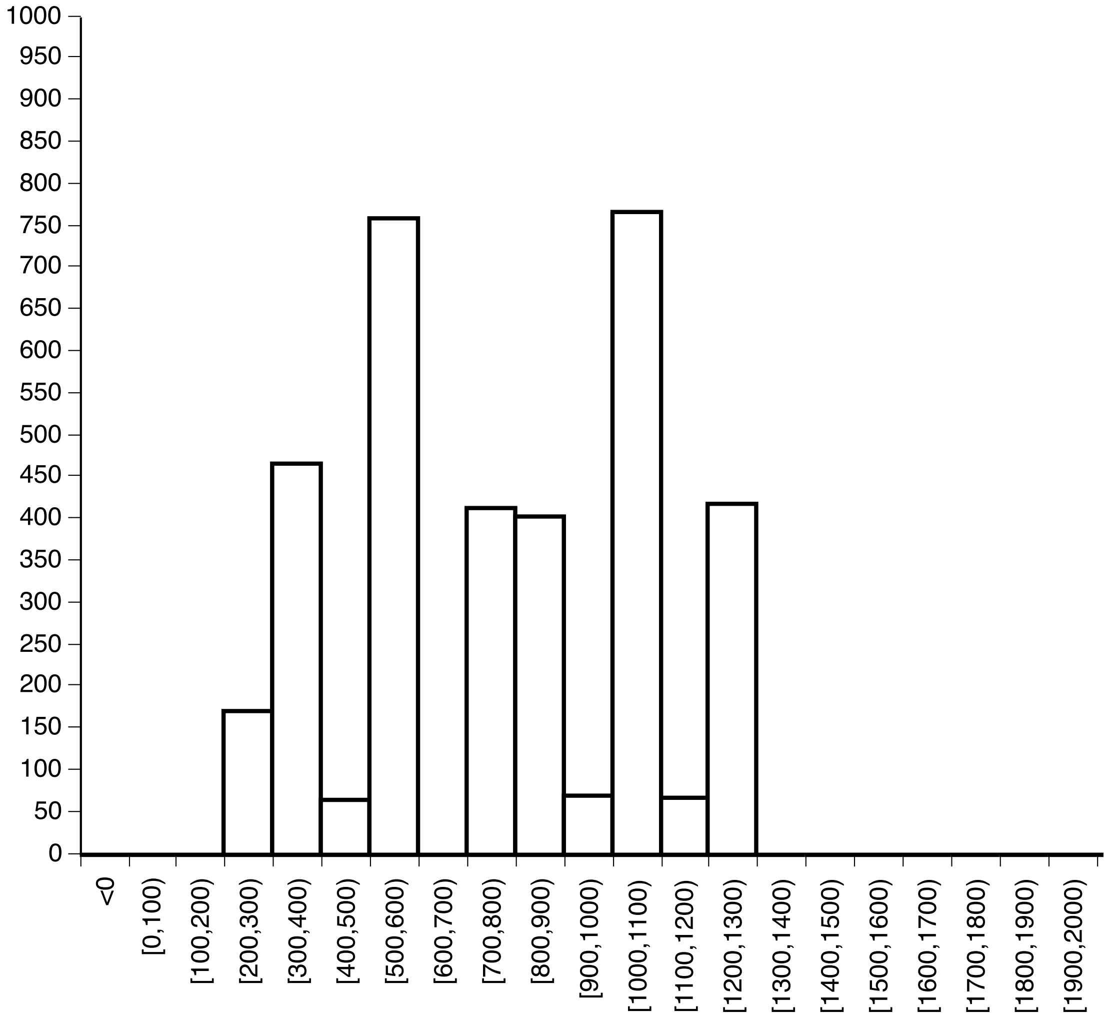
Human Data: 87.5% Validity, Valid Trial Types



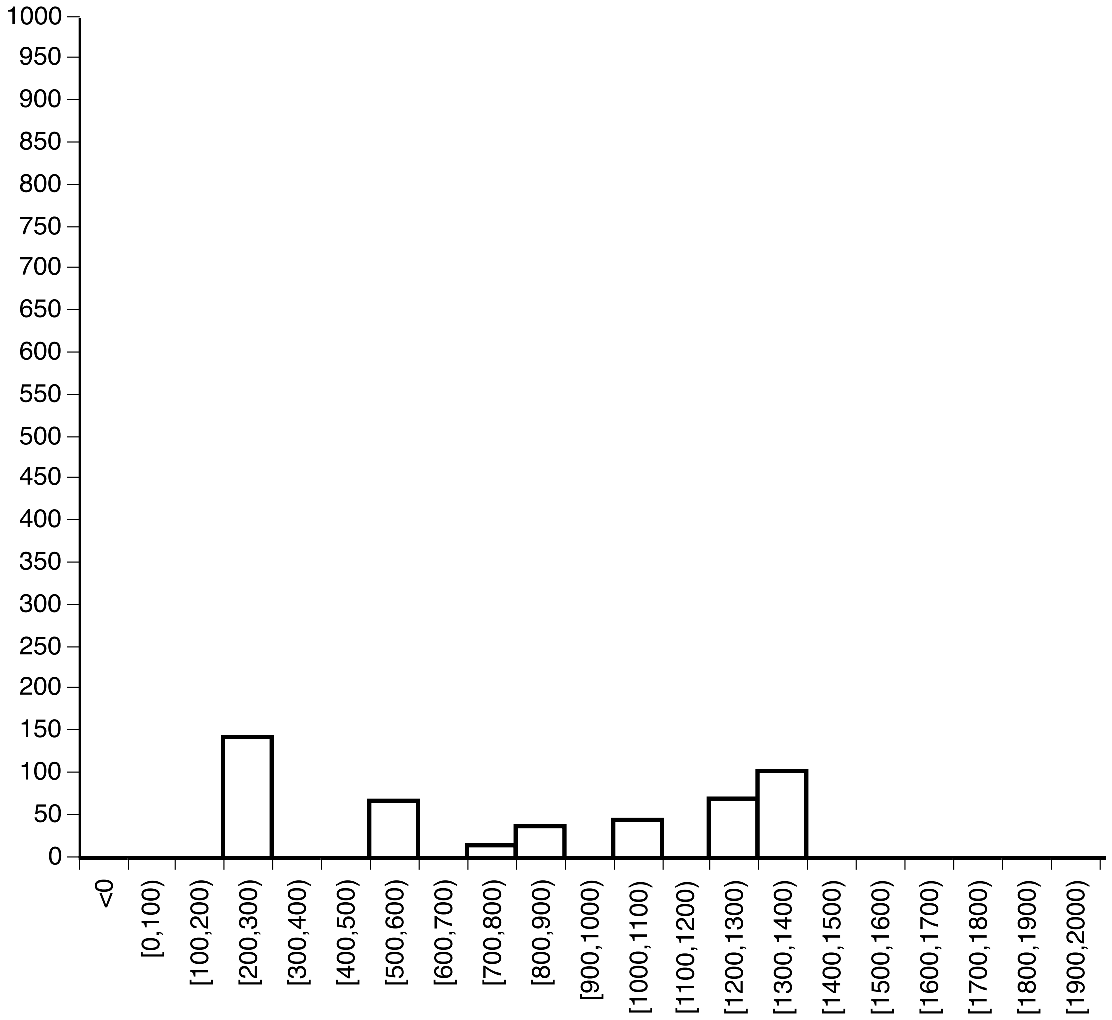
Human Data: 87.5% Validity, Invalid Trial Types



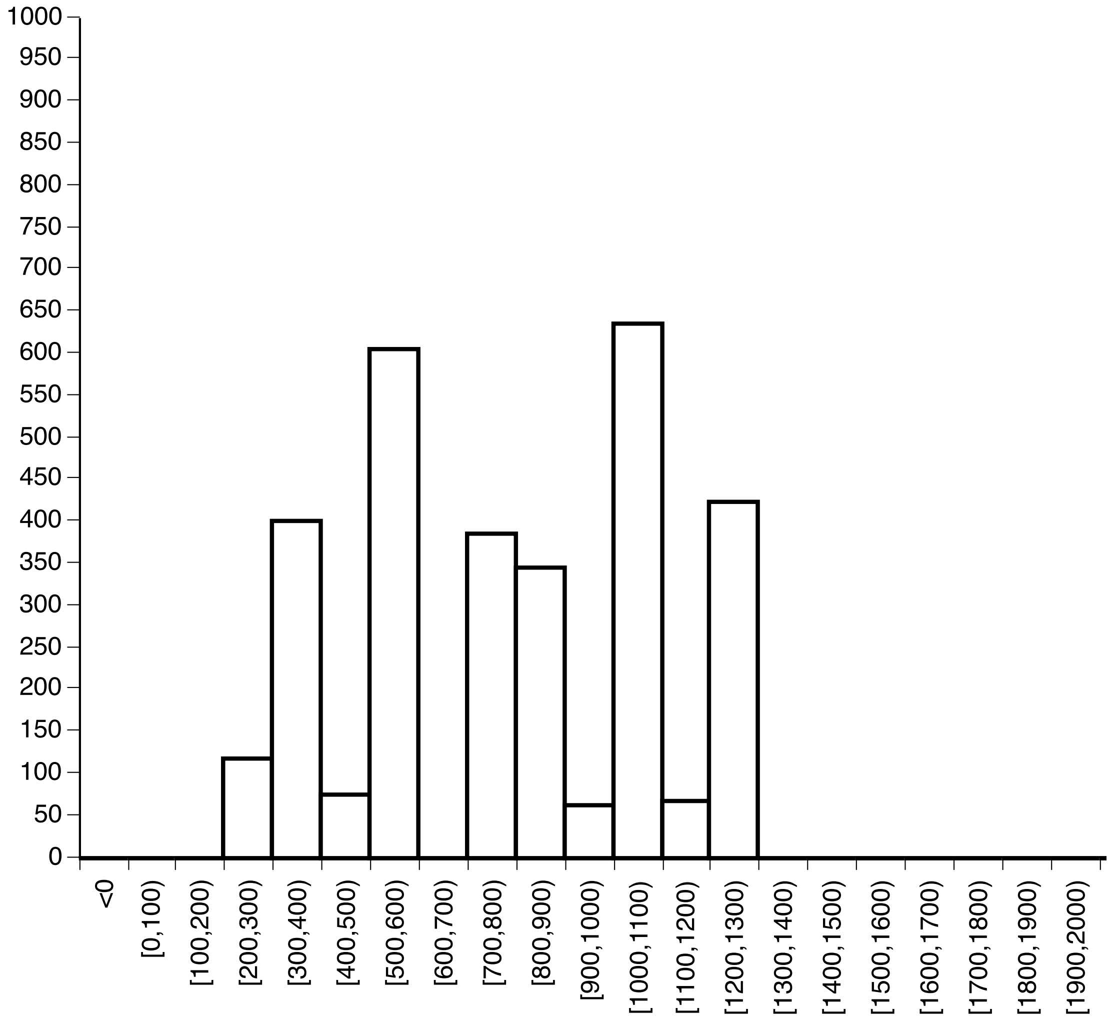
Human Data: 100% Validity, Valid Trial Types



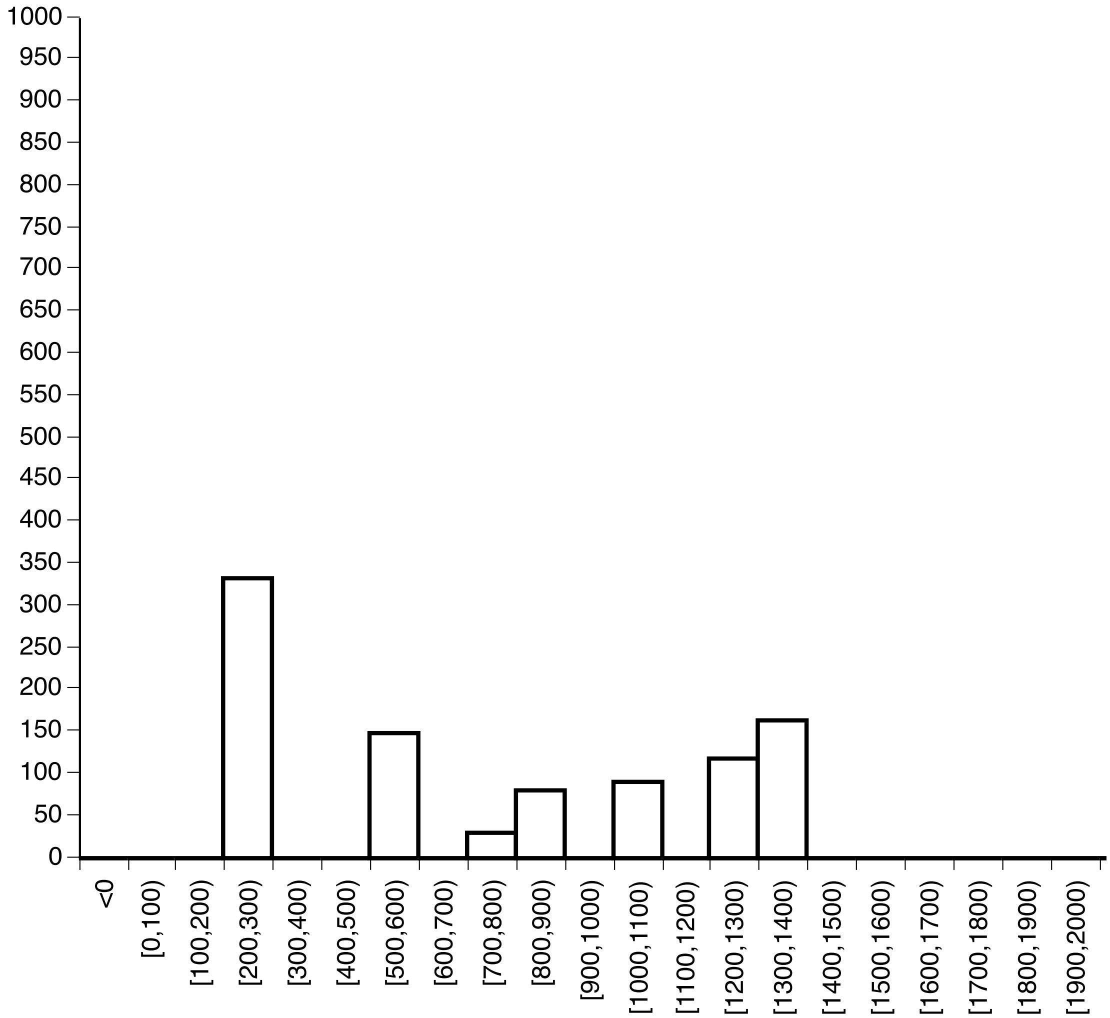
Experiment 1 ACT-R Model: 0% Validity, Invalid Trial Type



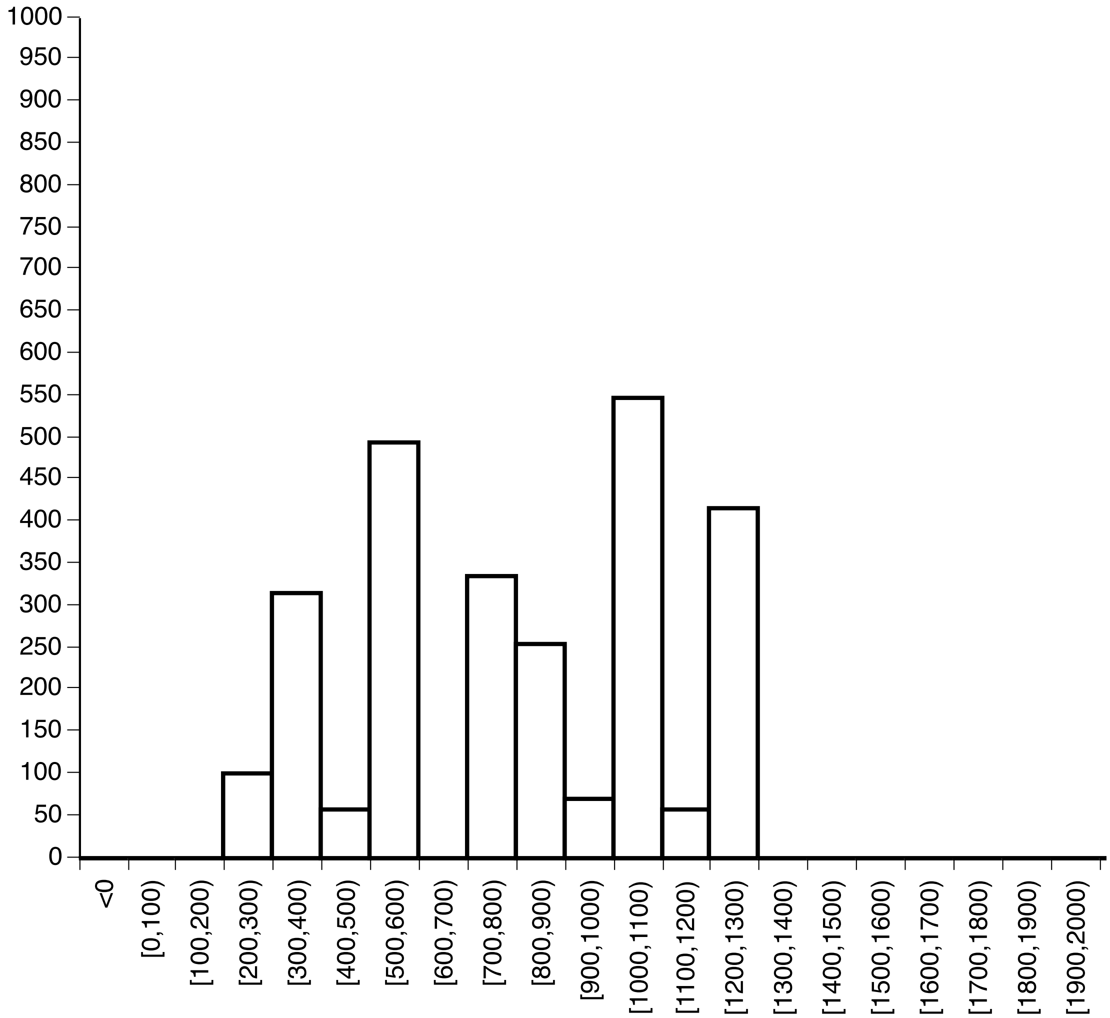
Experiment 1 ACT-R Model, 20 model runs: 12.5% Validity, Valid Trial Type



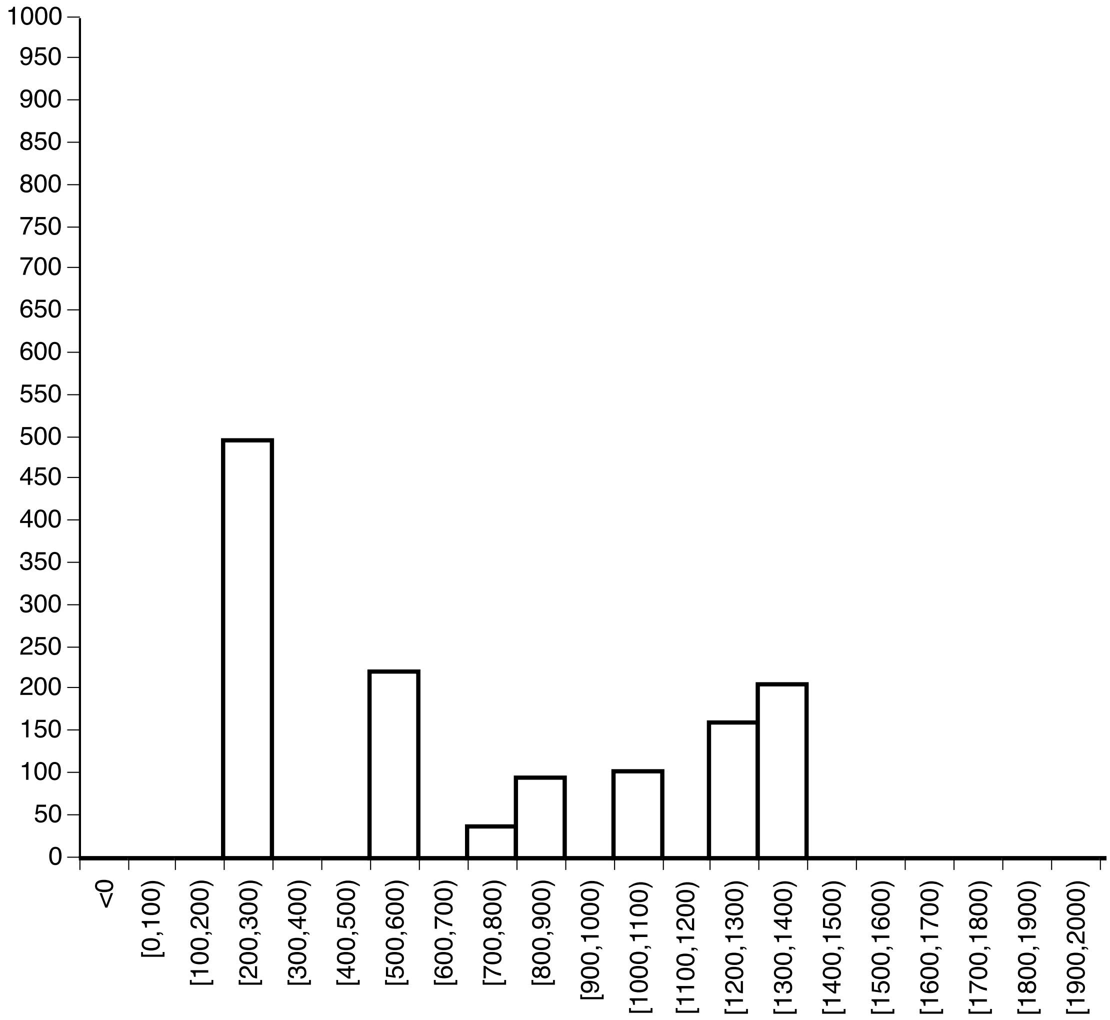
Experiment 1 ACT-R Model, 20 model runs: 12.5% Validity, Invalid Trial Type



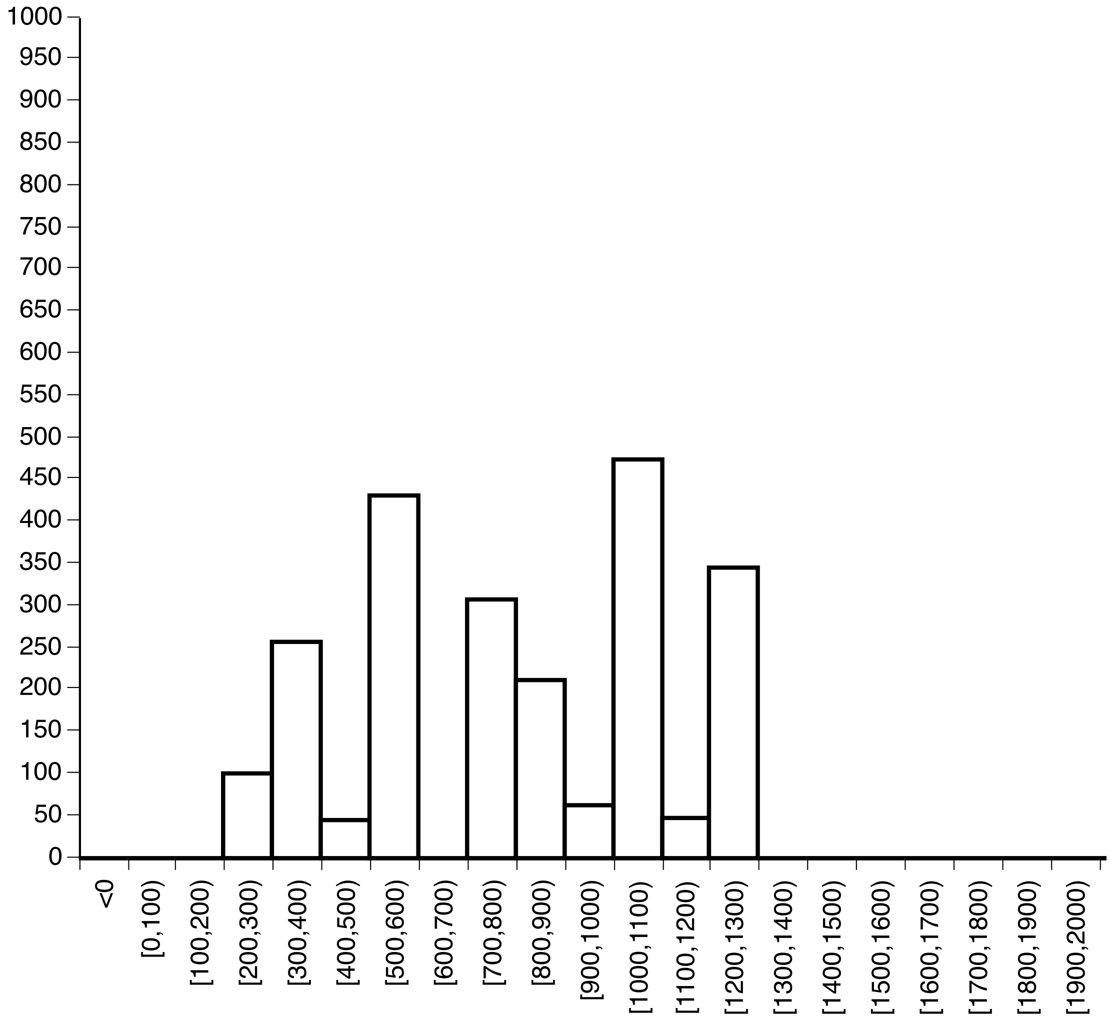
Experiment 1 ACT-R Model, 20 model runs: 25% Validity, Valid Trial Type



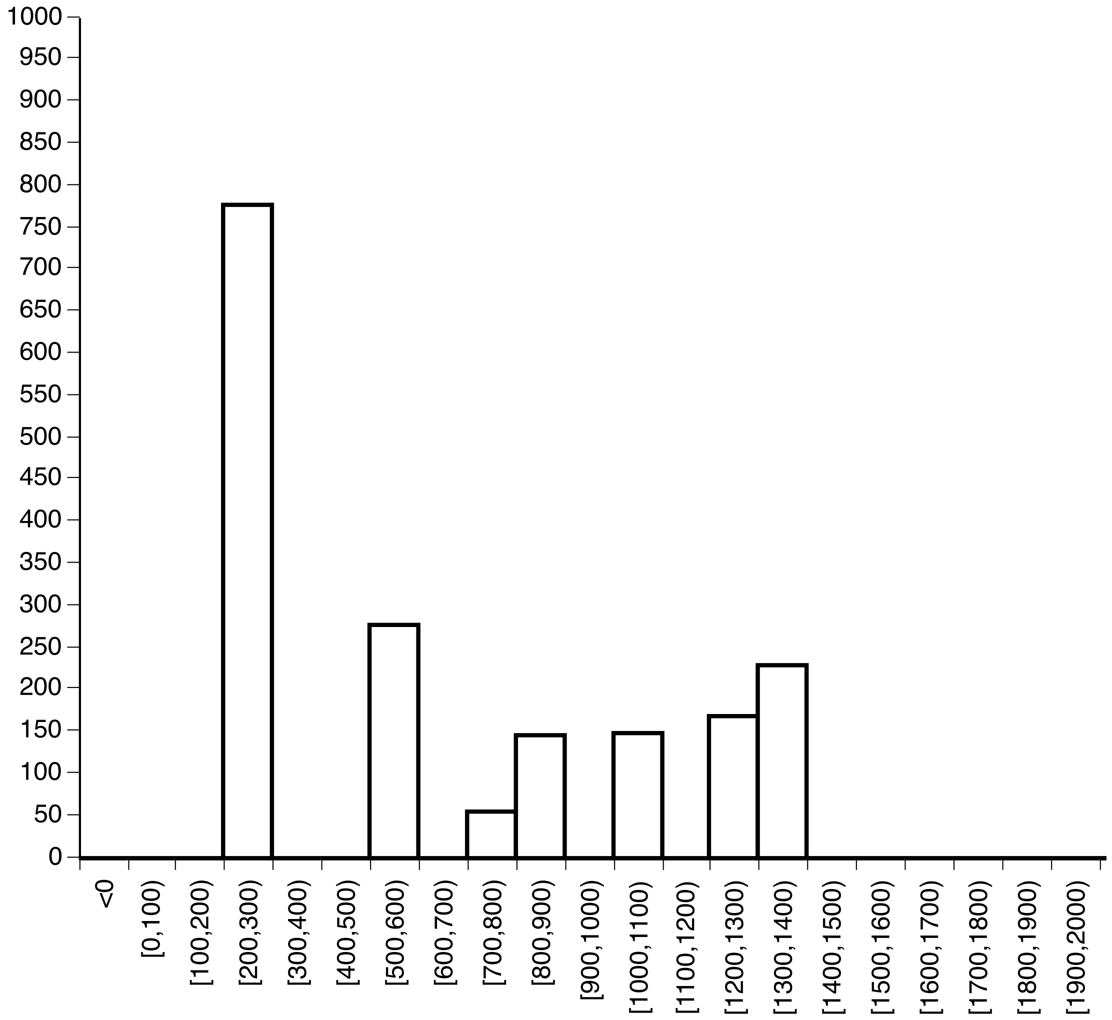
Experiment 1 ACT-R Model, 20 model runs: 25% Validity, Invalid Trial Type



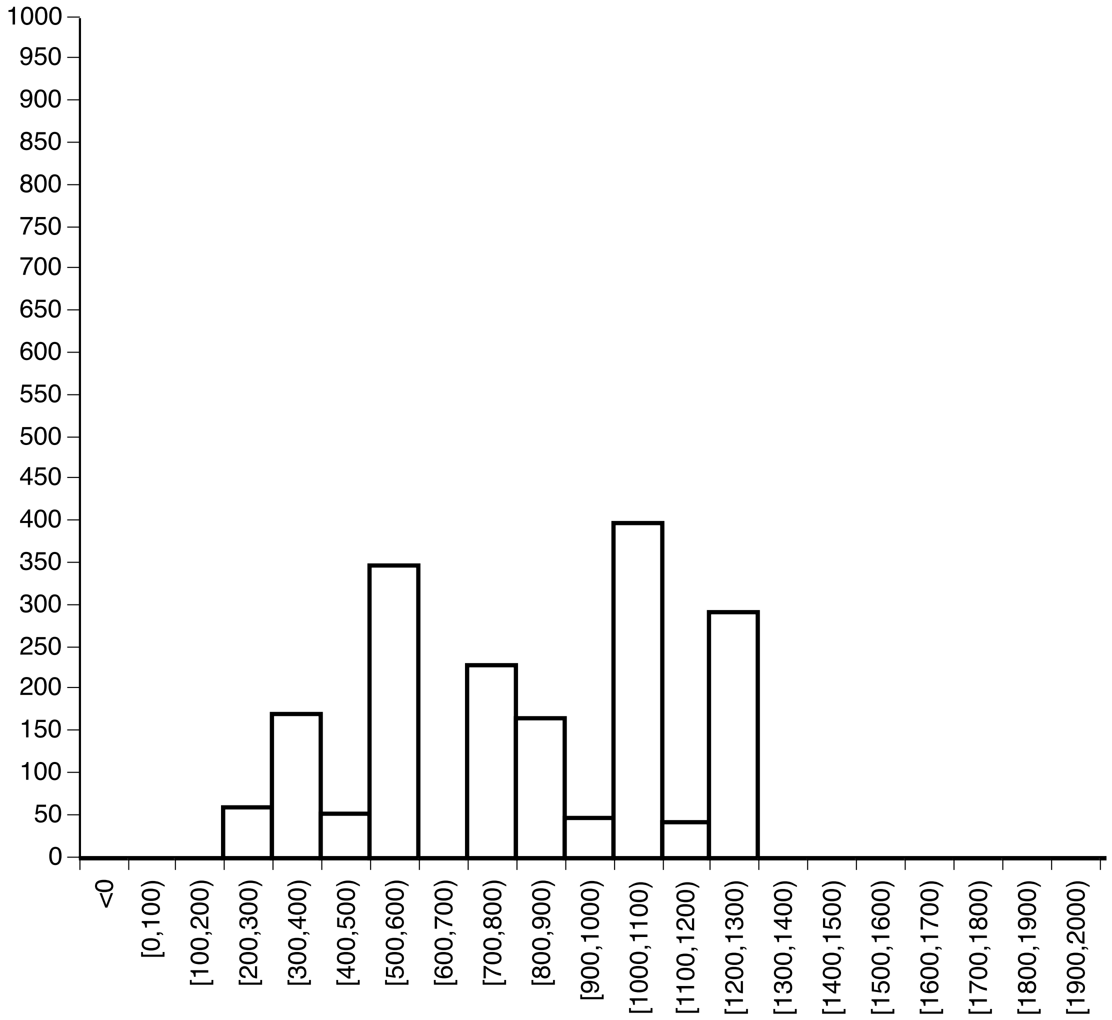
Experiment 1 ACT-R Model, 20 model runs: 37.5% Validity, Valid Trial Type



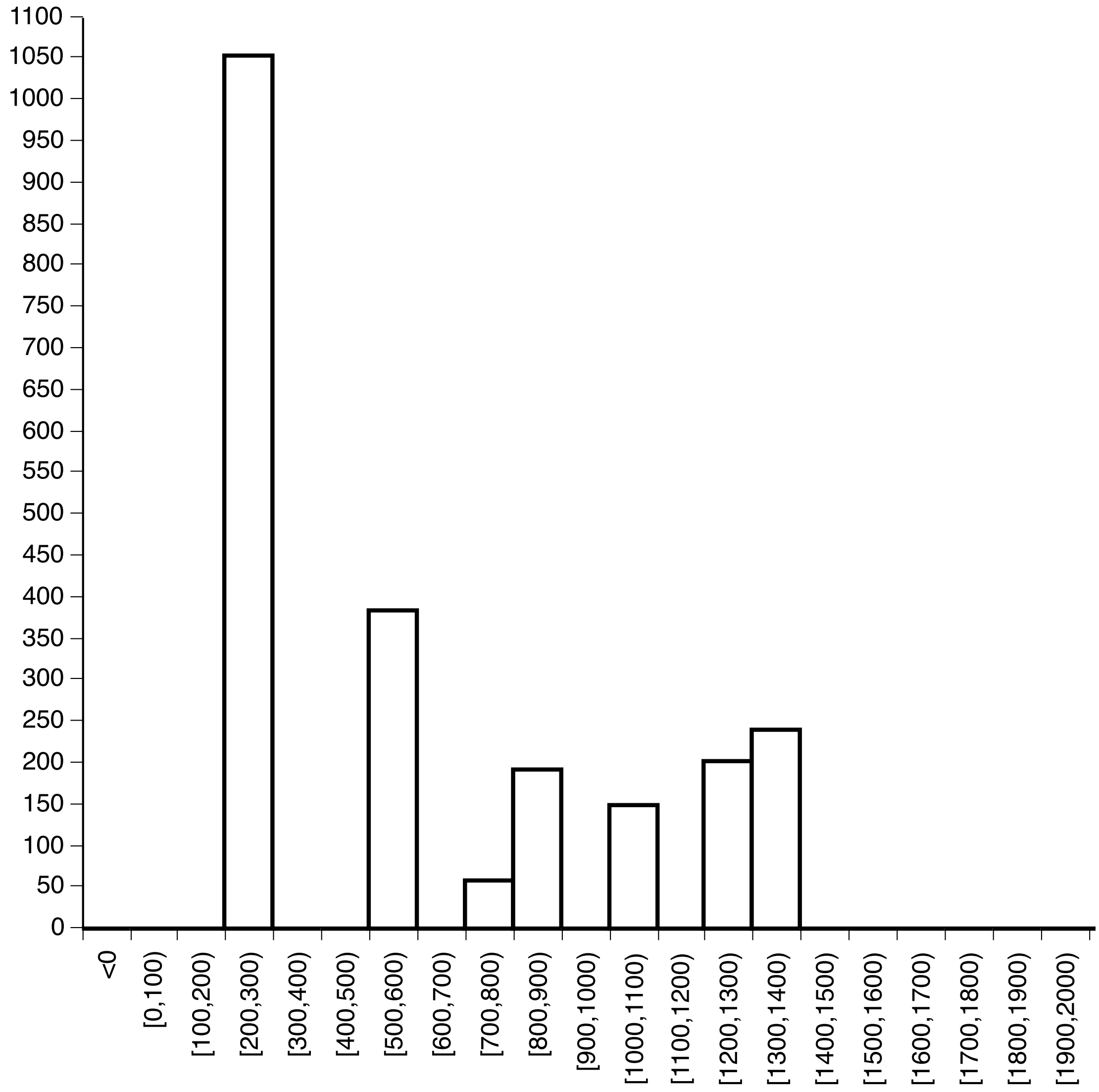
Experiment 1 ACT-R Model, 20 model runs: 37.5% Validity, Invalid Trial Type



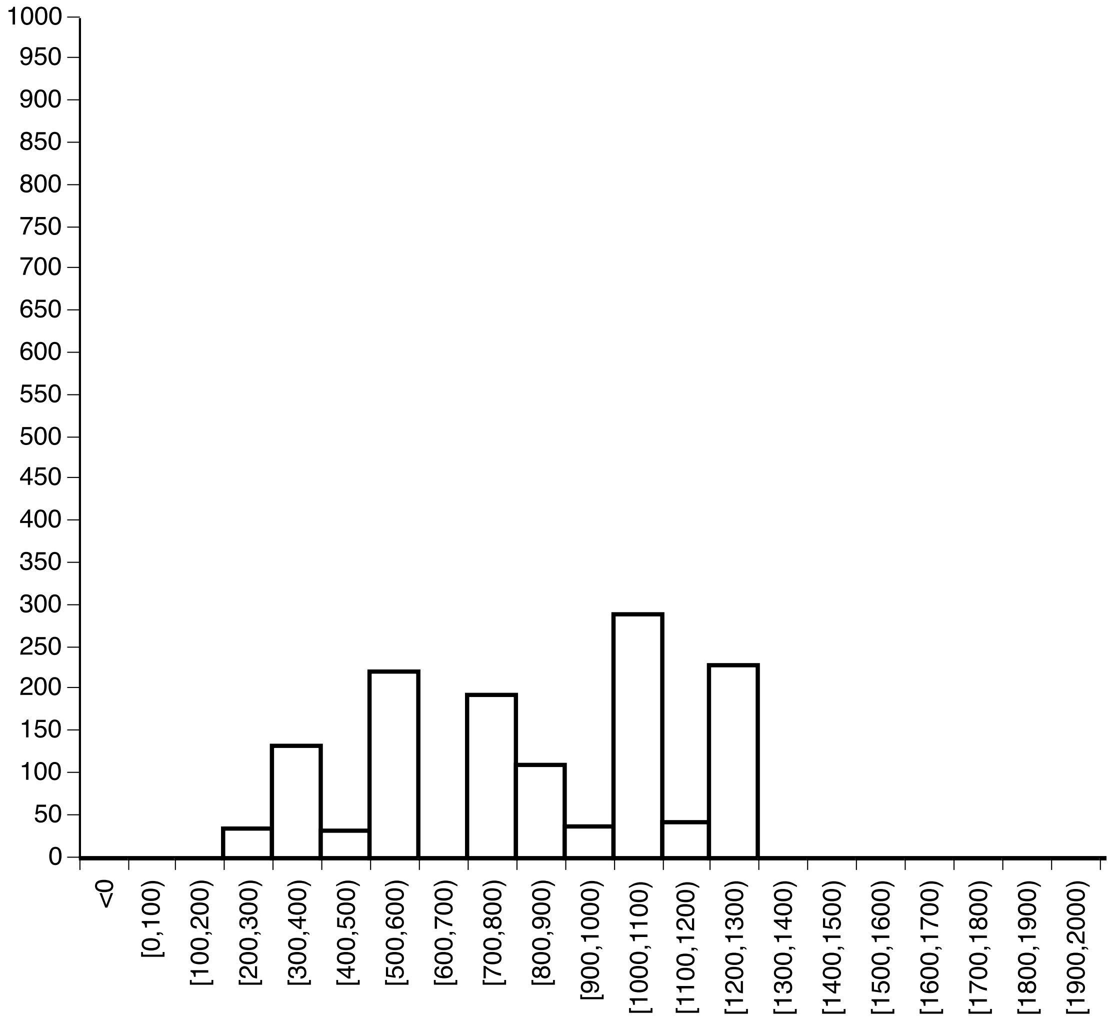
Experiment 1 ACT-R Model, 20 model runs: 50% Validity, Valid Trial Type



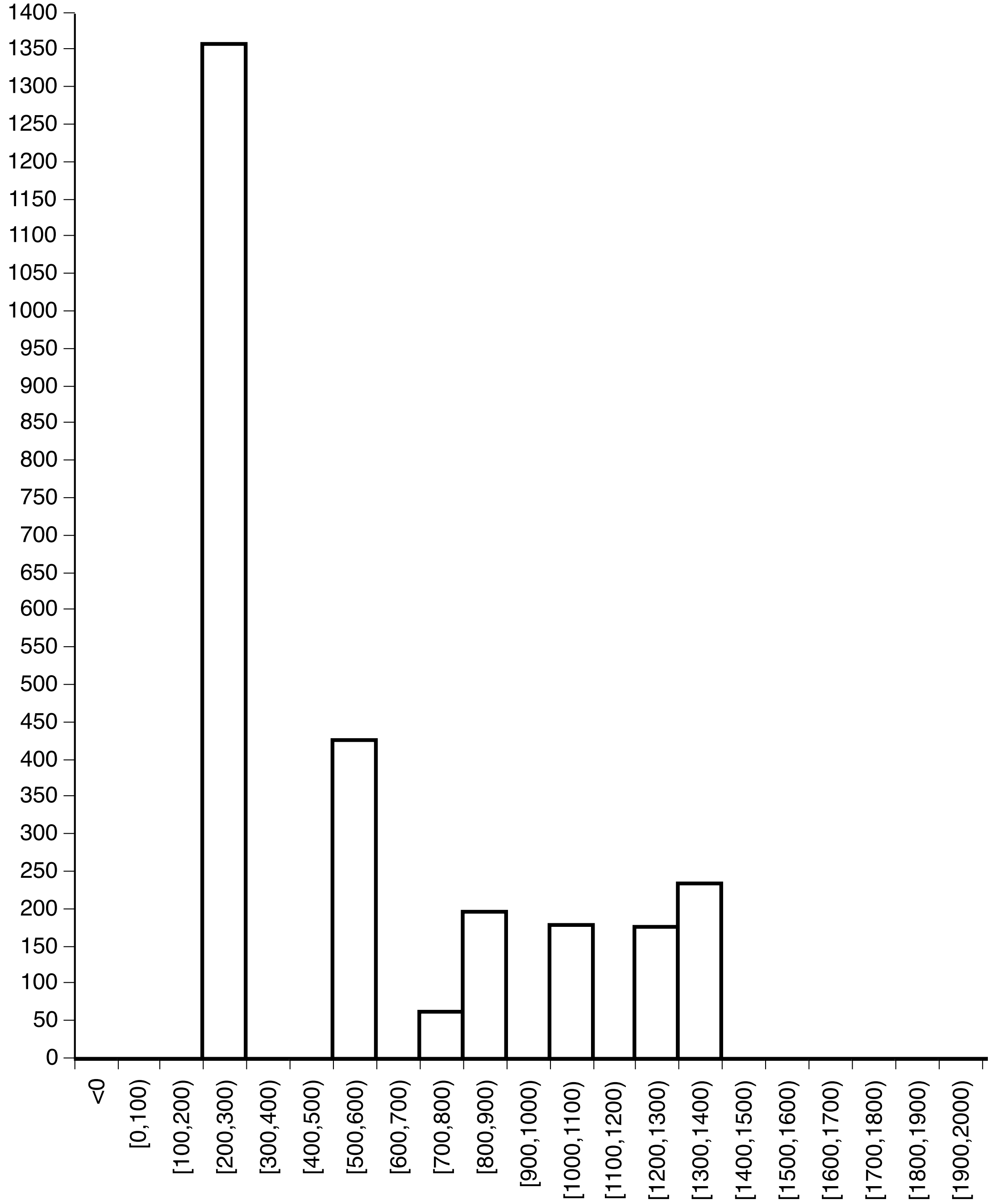
Experiment 1 ACT-R Model, 20 model runs: 50% Validity, Invalid Trial Type



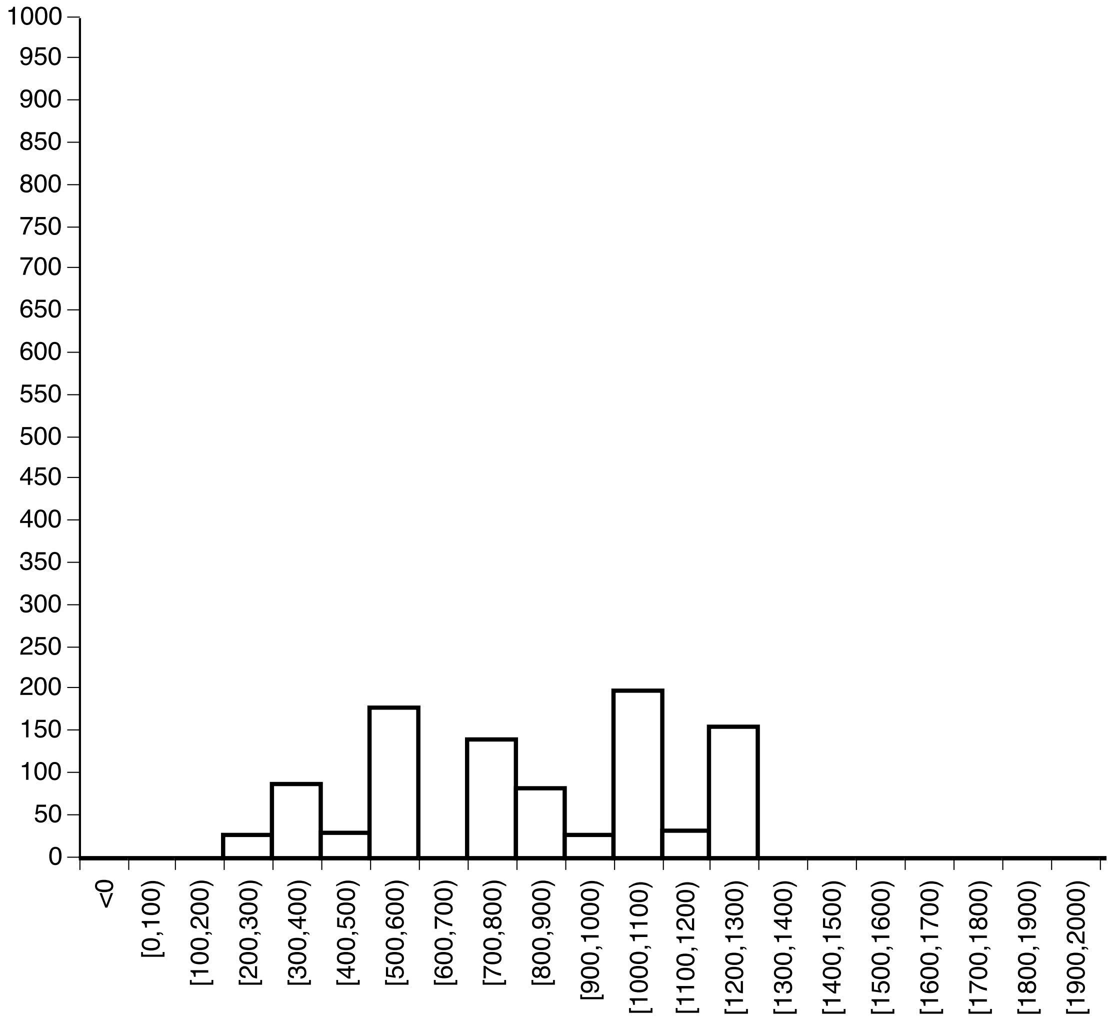
Experiment 1 ACT-R Model, 20 model runs: 62.5% Validity, Valid Trial Type



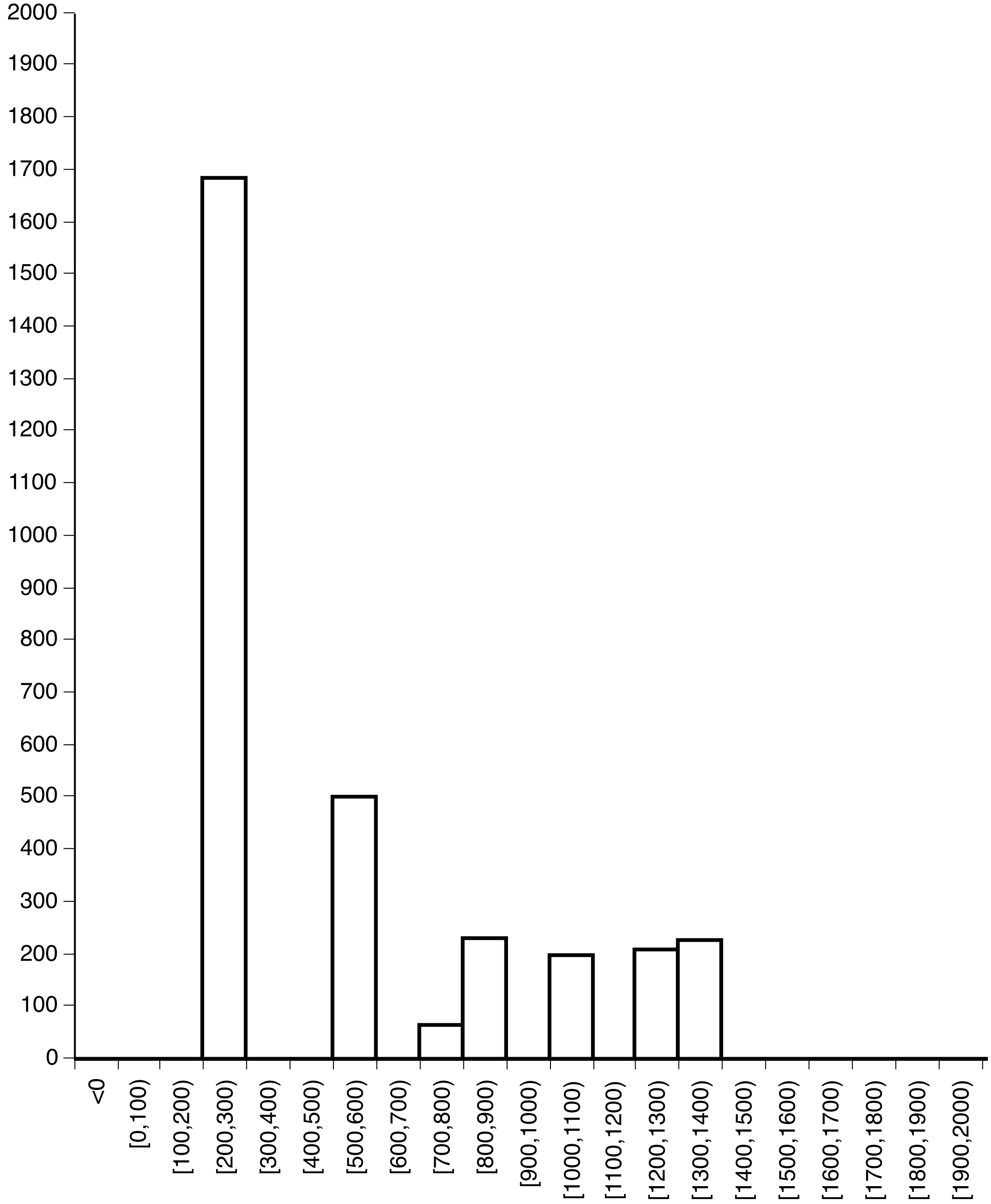
Experiment 1 ACT-R Model, 20 model runs: 62.5% Validity, Invalid Trial Type



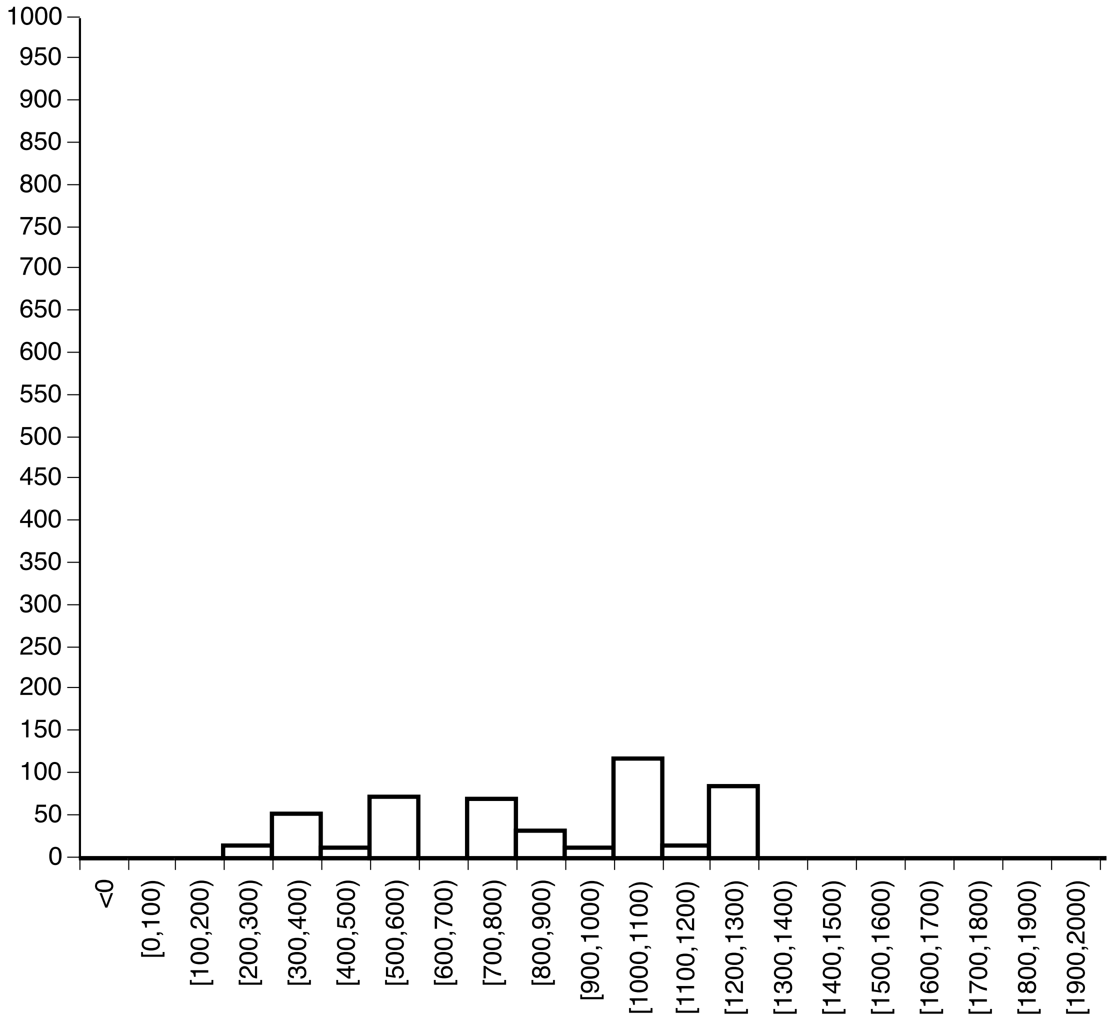
Experiment 1 ACT-R Model, 20 model runs: 75% Validity, Valid Trial Type



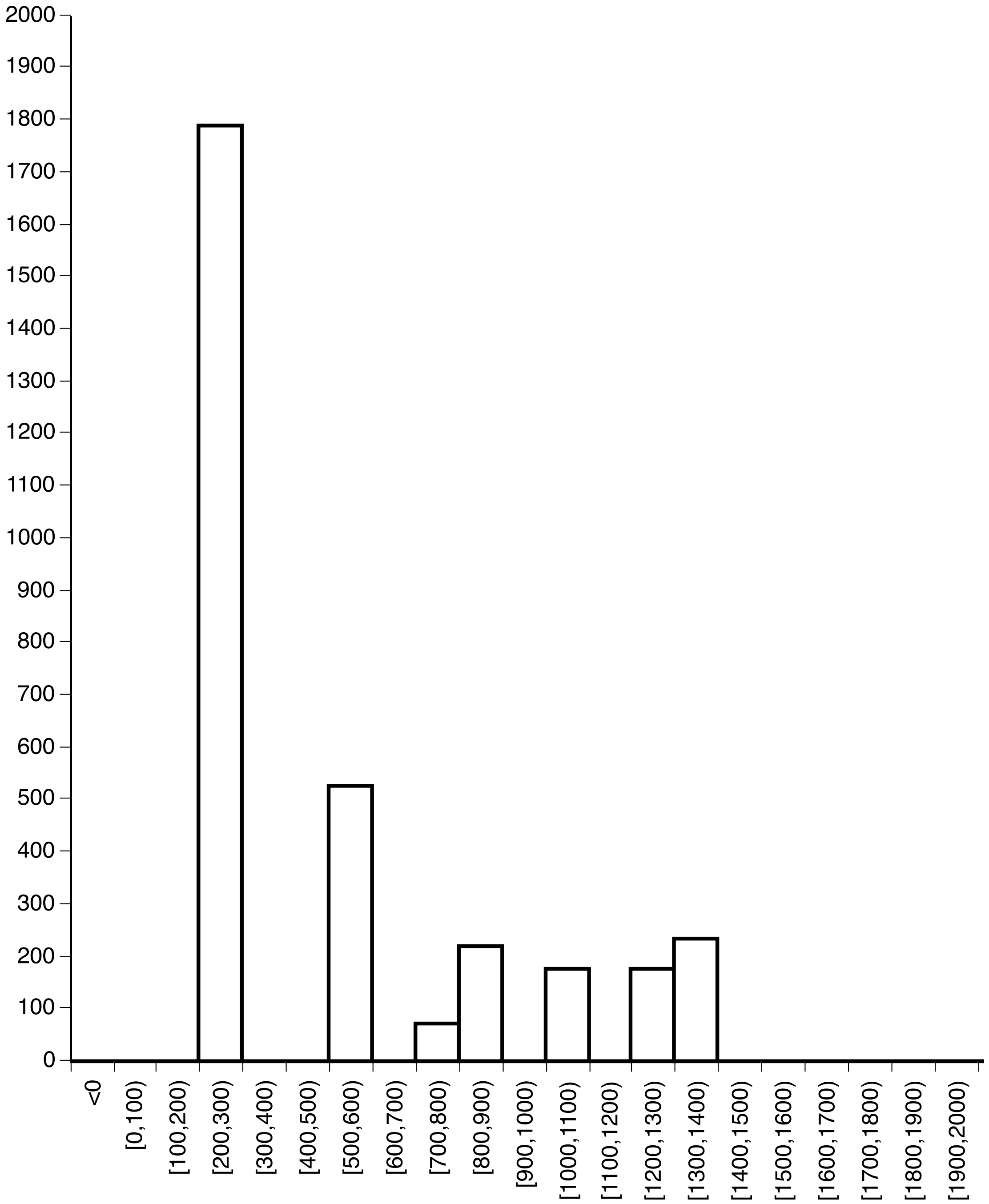
Experiment 1 ACT-R Model, 20 model runs: 75% Validity, Invalid Trial Type



Experiment 1 ACT-R Model, 20 model runs: 87.5% Validity, Valid Trial Type



Experiment 1 ACT-R Model, 20 model runs: 87.5% Validity, Invalid Trial Type



Experiment 1 ACT-R Model, 20 model runs: 100% Validity, Valid Trial Type