

Doozandeh, P. (2016). Quantification of human confidence in functional relations. *Cognitive Systems Research*, 40, 18–34.

Quantification of Human Confidence in Functional Relations

Pooyan Doozandeh

Abstract

What makes people infer that two continuous-valued entities are functionally related? Involving factors influencing human confidence in the existence of a functional link between two supposed variables has not so far been discussed in function learning literature. By examining this problem and based on relevant results from cognitive psychology, I propose a hypothesis according to which human confidence in a link between cue and criterion is affected by three factors: The difficulty of functions, the level of noise in observed data, and the sample size. Here, the formalization of this hypothesis forms a novel mathematical model of function learning which can also be used for predictions; so the resulting model receives cue-criterion pairs of a supposed relation and produces two outputs: Confidence and predicting function. In an experiment, the performance of a computational implementation of the model is compared with human data. The results show that the model is successful in tracking changes in human confidence. A close correspondence between the predictions of the model and humans was also achieved.

Keywords: Function learning; Function recognition; Confidence measurement; Mathematical model

1. Introduction

People can learn functional relations in real-world dynamic environments. This ability is used so frequently that we hardly even recognize it; it is the prerequisite of our judgments in multifarious contexts. In everyday life, we may perceive it in establishing the relationship between turning the volume knob of a radio and the change in the sound intensity of it; or the amount of coffee intake and increased wakefulness. In medical research, scientists may investigate how, for instance, a change in the amount of cigarettes smoked per day would change the probability of lung cancer or heart attack. In human sciences and economy, a simple example can be the notion that the increase in cash injection together with a decrease in production, would lead to an increase in inflation rate. In short, learning functional relations is essential in human judgment and a prerequisite for human knowledge. Establishing functional relations can also be thought of as cases of causal induction with both variables of cause and effect take continuous values.

The existence of a single mental mechanism behind all instances of human function learning is the shared assumption of cognitive psychologists. From this view, function learning is a mental system that, at first, measures the magnitude of two variables¹, namely cue (cause) and criterion (effect); for example, the degree of the pressure of gas pedal is the value for the cue and the RPM² of the engine is the value for the criterion. In the next step, the learner would sample the supposed relation by considering certain pairs of values of the cue and at those values, measuring the corresponding values of the criterion. Having a number of cue-criterion pairs, the mental system receives them as input. The output is a mathematical function (either explicit like algebraic rules or implicit like artificial neural networks) which is fitted on input pairs, as the final judgment that is generalized over the entire possible values of the cue. This generalization of the values of criterion from observed samples of the cue is also called prediction; since the resulting function can predict the values of criterion for unobserved cues. The main subject of the literature of modeling in function learning has so far lain in the question of what theoretical system we can devise that would have the human capability in such predictions.

Before investigations on human power in prediction, I think we should first take one step back and address a more fundamental question: What makes people infer that there exists a

¹ There can be more than two variables, both for the cue and criterion. For the sake of simplicity and plotting the pairs of cue-criterion, the model and experiments in this article work with single cue tasks.

² Revolution per minute

functional (causal) link connecting two variables of cue and criterion? In other words, in light of the assumptions of cognitive psychology, if a human observer is presented with a set of cue-criterion pairs, what factors affect his/her confidence in the existence of a functional link between the cue and criterion?

By focusing on cue-criterion pairs, the present study aims for introducing a measure that can quantify human confidence in the existence of functional relations. This measure is then embedded in a novel mathematical model of function learning that can also be used for prediction. The model integrates past findings in cognitive psychology and uses the rule-based approach (Carroll, 1963; Brehmer, 1974; Koh & Meyer, 1991), combining knowledge before and after observations.

In order to assess the existence of a functional relation between the cue and criterion, the model of this research receives cue-criterion pairs as input. Then, a number of preexisting mathematical functions are fitted on input pairs and for each parameterized function if the goodness of fit is higher than a threshold, the model calculates a confidence measure. In the end, the function with the largest confidence measure is chosen. This measure is the confidence in the existence of a functional link between the cue and criterion and is then passed as the first and primary output of the model. The corresponding rule of the function is the other output that can be used for predictions.

The next section examines previous relevant works in causal induction and function learning. In Section 3, the measure of confidence and the general model are formulated and presented and then, the description of different parts of the model follows. Section 4 explains experiments with human participants and Section 5 includes a general evaluation of the model's performance. Section 6 discusses abstract issues of function learning and causal induction with respect to the problems and shortcomings of the model.

2. Background

Researches on human causal induction with binary-valued variables of cause and effect have been both older and more popular than continuous-valued causal induction (or function learning). This is in spite of the fact that for human judgment, learning, and decision making, function learning is as equally, if not more, important. Upon closer examination, it becomes obvious that these two fields have similarities and many examples of binary-valued causal

induction can be thought of as examples of function learning. For example, Griffiths & Tenenbaum (2005) give an example of a binary-valued causal induction between injection of mice with a certain chemical and the expression of a particular gene. Viewed closely, it can be an example of continuous-valued causal induction in which the amount of chemical and the length of the gene can take continuous values. We can similarly think of other famous examples of causal induction like smoking and lung cancer, coke consumption and diabetes, etc.

While most research in binary-valued causal induction had been centered on assessing the strength of the relation between cause and effect, Griffiths & Tenenbaum (2005) were first to recognize the importance of questioning the existence of any causal link between cause and effect and they presented a rational model for this end. Six years later, Griffiths et al. (2011) formulated a Bayesian model for causal induction that could both incorporate prior knowledge and address the problem of assessing the existence of a causal relation between two binary-valued variables of cause and effect.

In the literature of function learning, however, there has not been a similar attempt in creating a model that can determine the existence of a causal relation between continuous-valued variables of cause and effect. This investigation seems to be absent from both theoretical psychology and cognitive or rational modeling. For example, in the case of drinking coffee and wakefulness, how do we infer if there exists any causal link between them? And how can we quantify human confidence in the existence of the link?

Models in the history of research in function learning are usually divided in two groups. The first, which is commonly called the rule-based group of models, assumes that the task of learning a function is conducted by approximating explicit mathematical functions, or rules, on observed cue-criterion pairs (e.g., Carroll, 1963; Brehmer, 1974; Koh & Meyer, 1991; Narain et al., 2014). Similarity-based models, as the second group, argue that functions are learned associatively and novel inputs are being predicted by their degree of similarity with observed values (e.g., DeLosh et al., 1997; Busemeyer et al., 1997). There have also been attempts at presenting hybrid models (McDaniel & Busemeyer, 2005; Kalish et al., 2004) or achieving reconciliation between the two approaches of modeling (Lucas et al., 2015).

None of the stated researches in function learning investigated the existence of a link between cue and criterion. The only work with this aim is recently conducted by Schulz et al. (2015), in which assessing the existence of a functional link between variables, or recognizing

functions, is called “perceived predictability of functions”. The study uses a previously defined Bayesian model of function learning with Gaussian processes (Griffiths et al., 2009; Lucas et al., 2015). Schulz et al. (2015) attempted to show, experimentally, that the “smoothness” of functions, rather than noise or sample size, is a major factor that determines the predictability of functions. In other words, if we plot a set of cue-criterion pairs on a plane, the resulting shape must resemble a smooth function to be easily recognized as a functional relation.

While authors in Schulz et al. (2015) clearly distinguish between smoothness and noise, there seems to be a difficulty in defining such a difference in sampling a function. Let us think of one specific curve; how exactly can we differentiate between a set of points which were sampled from that curve with noise, and another set of points sampled from a non-smooth instance of the same curve?³ It seems that noise and non-smoothness go completely hand in hand.

3. The model

The primary goal of this research is to present a measure that quantifies human confidence in the existence of a functional link between two continuous-valued variables. In this section, a generic model of function learning is introduced that contains a confidence measure at the heart of itself. For every set of cue-criterion pairs, the resulting model produces a confidence level and a predicting function, provided that it can recognize a link from input pairs. This idea is implemented in a mathematical, rule-based model of function learning which utilizes and incorporates some of the past psychological findings in function learning literature.

The exact values of some parameters of the model are tuned up on the data of a small training set. So in the rest of this section, the confidence measure and the resulting structure of the model and its parameters are introduced and the exact calculations of the values of those parameters are discussed in Section 4.

3.1. Confidence measure

In order to introduce a measure of confidence in the existence of a functional link between two variables, first we have to determine what factors influence human confidence in real-world

³ To better understand this objection, as authors suggest, go to -- <http://bit.ly/1CtXfMA> -- and with a certain sample size, set both Variance (noise) and Smoothness on minimum and click on Generate; and then set them on maximum. If authors’ clear-cut distinction between noise and non-smoothness is valid, there must be a recognizably visual difference, but it does not seem so.

function learning tasks. Since there have not been much behavioral experiments or theoretical studies in psychology with this exact subject, past works with some degrees of relevance are investigated and interpreted with respect to the goal of the present research. Thus, while a body of previous psychological studies is used, the selection of the effective factors and their combination in the final measure of confidence are intuitive and hypothetical. I propose to use three factors in measuring confidence; these factors are difficulty of functions, noise, and the sample size.

3.1.1. Functions and their difficulties

The first factor which is considered as affecting human confidence is the difficulty of functions. The assumption is as follows: People tend to be more confident in the existence of a functional link between variables that are related through easier functions. In different terms, if we plot the cue-criterion pairs of a real-world functional relation, confidence for a link between the cue and the criterion is more if the resulting shape resembles an easy curve. This requires the presupposition that people can distinguish between the standard shapes of different functions in function learning tasks; their confidence and learning are changed with respect to the change in the function (or at least it seems to be so). This claim is supported by a large body of experimental results in psychology (see Busemeyer et al. (1997) as only one reference to many studies in the subject).

The order of the difficulty of functions is derived from Busemeyer et al. (1997) in which they presented ten principles as constraints for future models of function learning. Based on those principles, they concluded that the general order of the difficulty of functions is as follows: “cyclic > non-monotonic⁴ > monotonic decreasing > monotonic increasing > linear”. Here, I used the results of Busemeyer et al. (1997) with some exceptions which will be discussed in Section 6. A numeric value is assigned to each function as an indicator of their respective easiness. Functions, their order of difficulty, and the numeric values indicating their easiness, as would be used in the model, are shown in Figure 1.

⁴ In a given range, monotonic functions either increase or decrease, but never do both. For example, linear and power functions are monotonic while quadratic and periodic functions are not.

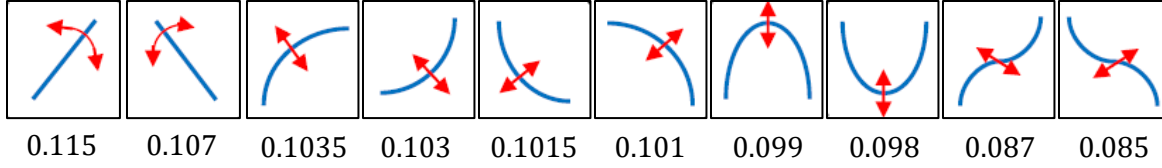


Fig. 1. The order of the easiness of functions (from left to right) and their respective values of ease, as loosely derived from Busemeyer et al. (1997). Red arrows indicate that the curve of the functions can slightly change direction or shape with respect to the input pairs on which they would be fitted.











Values indicating the ease of functions change from 0.115 for the easiest function to 0.085 for the most difficult. Since these functions will be fitted on the cue-criterion pairs of the input, numeric values can be thought of as each function's possibility of being chosen. An optimization was needed for assigning such numeric values. They must have been assigned with respect for both, setting a reasonable difference to distinguish between functions, and leaving sufficient chance for more difficult functions to be chosen. With this goal in mind, I first assigned a semi-arbitrary set of values and then slightly adapted those values to the results of the training experiments.

The ten functions as are presented in Figure 1 are among those that are frequently used in many of the modeling attempts in function learning literature. For example, Lucas et al. (2015) included the following types of functions in their most comprehensive model: Positive Linear, Negative Linear, Quadratics, and Nonlinear (nonlinear relationships are fitted by a radial basis function) with the respective prior probabilities of 8, 1, 0.1, and 0.01. The model which will be presented in this section is a generic model and is supposed to work in a wide variety of settings. From this perspective, although there exist different or more complex functional relations which can be recognized by humans, since they are infrequent in real-world learning tasks and are specific to certain environments, they are not included in the model; further reasons for this decision are discussed in Section 6.

The presented functions will be fitted on input pairs (observations) later in the model. For this end, the standard rules of these functions, how they are separated, and their degrees of freedom from standard curves must be clearly specified. This is accomplished by restricting the parameters of these functions, as shown in Table 1.

Table 1

Functions, their rules, and how their parameters are restricted.

<i>Functions</i>	<i>Rules</i>	<i>a</i>	<i>b</i>
1 	$ax + b$	$(0, +\infty)$	$(-\infty, +\infty)$
2 		$(-\infty, 0)$	$(-\infty, +\infty)$
3 	ax^b	$(-\infty, 0)$	$(-\infty, 0)$
4 		$(0, +\infty)$	$(0, +\infty)$
5 		$(0, +\infty)$	$(-\infty, 0)$
6 		$(-\infty, 0)$	$(0, +\infty)$
7 	$ax^2 + bx + c$	$(-\infty, 0)$	$(-\infty, +\infty)$
8 		$(0, +\infty)$	$(-\infty, +\infty)$
9 	$ax^3 + bx^2 + cx + d$	$(0, +\infty)$	$(-\infty, 0)$
10 		$(-\infty, 0)$	$(0, +\infty)$

Note. Inside each parenthesis in columns of a and b, the first number is the lower bound and the second is the upper bound. Note that the parameters of c and d for quadratic and cubic functions have no constraints, so their range of values is $(-\infty, +\infty)$.

3.1.2. Noise and sample size

Having defined constrained functions and their respective difficulties, which are considered generally as the first factor, there are other two factors which were also mentioned in Schulz et al. (2015). The argument behind the noise, as the second factor, is that for a set of cue-criterion pairs, the higher the level of noise (with respect to the standard curves of ten specified functions), the less confidence there is in the existence of a functional link between the variables of cue and criterion.

In the literature, noise is often referred to by its reverse equivalent, or the “goodness of fit”, and is calculated through various methods (like SSE⁵, MAE⁶, RMSE⁷, etc.). In calculating the goodness of fit in the model that follows, it is important for the chosen method to be unaffected by two elements: The range of change, or the magnitude of the numeric values of variables (since the model is not restricted to work in a given range); and the number of pairs, or the sample size (so that the factor of noise would be distinct from the sample size). The exact method in the following model for measuring the goodness of fit is defined within the framework of the computational implementation of the model and will further be discussed in Section 4.

The number of input pairs, or the sample size, is the third factor that influences confidence. For every set of input pairs from observations in a function learning task, human confidence in the existence of a functional link between cue and criterion is increased by the growth in the number of observed pairs. The nature of this effect is unlike the two previous factors; since it seems that the effect of sample size changes with respect to the number of input pairs. For example, the increase in human confidence when we observe the fourth pair of a relationship is higher than when we observe the tenth pair. So here, the effect of sample size on the overall confidence must be in a way that is increased rapidly at the beginning and is mediated with more pairs. For this reason, if we should define a term for the effect of sample size on the overall confidence, the term must be a negative exponential term and the exact number of exponentiation will be calculated with the data of training set.

3.1.3. Formula

Three factors affecting the confidence in the existence of a functional relation between the two variables of cue and criterion have been hypothesized. They are difficulty of function, noise, and sample size, and now, the intended mathematical model requires a method to make a formal combination of these factors in a confidence measure. I suggest the simplest method and use a linear combination of the three factors. Since combining these factors to calculate confidence is a new proposal, the importance of each of these factors in overall confidence with respect to other factors must first be investigated. This goal is accomplished in the formula by multiplying a coefficient to each factor and after assessing the results of the training experiments, assigning

⁵ Sum of square error

⁶ Mean absolute error

⁷ Root mean square error

numeric values to the coefficients. Therefore, the combination of the three factors for calculating confidence is as follows:

$$Confidence(f) = (Difficulty(f) \times k1) + (GOF(f) \times k2) + (n^x \times k3)$$

$f:1 \rightarrow 10$

where f is the list of ten previously defined functions, GOF stands for the goodness of fit, n is the number of input pairs, x is the exponentiation factor, and $k1$, $k2$, and $k3$ are the coefficients for the importance of the three factors. The numeric values of x , $k1$, $k2$, and $k3$ will be assigned after gaining data from training experiments.

The presented measure is designed to calculate the confidence for every pre-defined function. This is because, after the observation of input pairs, the numeric values of the two involving factors in confidence, namely the difficulty and goodness of fit, are specific to every function. So for every set of input pairs, this measure yields many confidences for various functions and then, as is discussed below, the mathematical model chooses the largest confidence together with the corresponding function.

3.2. Mathematical model

Supposing the coefficients of the confidence measure are calculated, for every set of observations (or input pairs) we need the numeric values of difficulty, goodness of fit, and sample size to derive a number for confidence. In fact, the confidence measure, as the core of the model, lies at bottom of the steps that are needed to calculate a quantity of confidence and there must be a general framework in which the previous steps, needed to calculate the three factors, are systematically established. These steps form the general structure of the mathematical model, as depicted in Figure 2.

Of the three factors affecting confidence, difficulty of functions, unlike the goodness of fit and sample size, is the observer's knowledge before observations; so the ten curves and their respective difficulties are embedded in the model as the starting step. After the arrival of the input pairs, these curves are fitted on observed pairs and the rule of every curve is tuned on observations, as much as the degree of freedom of functions allows (see Table 1). At this step, it is also possible to calculate the goodness of fit for every curve.

Now, for every function, there is a difficulty, a goodness of fit, and a number for sample size; so the ingredients for computing confidence for every function are at hand. It must also be

noted that the model contains a filter for functions; those with low levels for the goodness of fit are blocked and the confidence is only computed for those that pass this threshold. This filter is designed so that the model would not need to compute confidence for curves that have no similarity to input pairs. In this way, if a computational implementation of the model is required, the operating complexity of this mathematical model is reduced. To define this filter, first it must become clear, on average, how much goodness of fit is required for humans to detect a relationship (or link) between the cue and criterion; so the exact level for the threshold of this filter is another element of the model which is tuned on the data of training set.

In the end, the model has one or many candidate functions, together with each function's confidence. These measures of confidence are compared and the largest confidence is chosen as the final judgment of the model. Since for this confidence there exists a corresponding function which was parameterized on input pairs, the rule of this function can also be used for predicting unobserved cues and is passed as the other output of the model.

4. Experiment

This section includes behavioral experimentation on human participants with two goals: to assign specific numeric values to the parameters of the model, and to test the performance of the resulting model. The main focus of this section, however, is on the performance of the model on test set. This is because the general structure of the model is derived from previous knowledge that came from relevant studies in psychology and resulted in a mainly hypothetical model. This has reduced model's sensitivity to the data of the training set. In contrast, associative approaches and models based on artificial neural networks are entirely or to a large extent dependent on the training set, and compared to those models, the role of the training data on the presented model of this research is marginal.

4.1. Computational implementation

Before experimentation on human participants, it would be more accurate and easier to have a systematic method to calculate the outputs of the model for every set of input pairs. Since fitting constrained functions (optimizing parameters) and computing the goodness of fit for every set of input pairs and every function were too cumbersome to be performed by paper and pencil, a software tool was used for these computations. I used the Curve Fitting Library of MATLAB

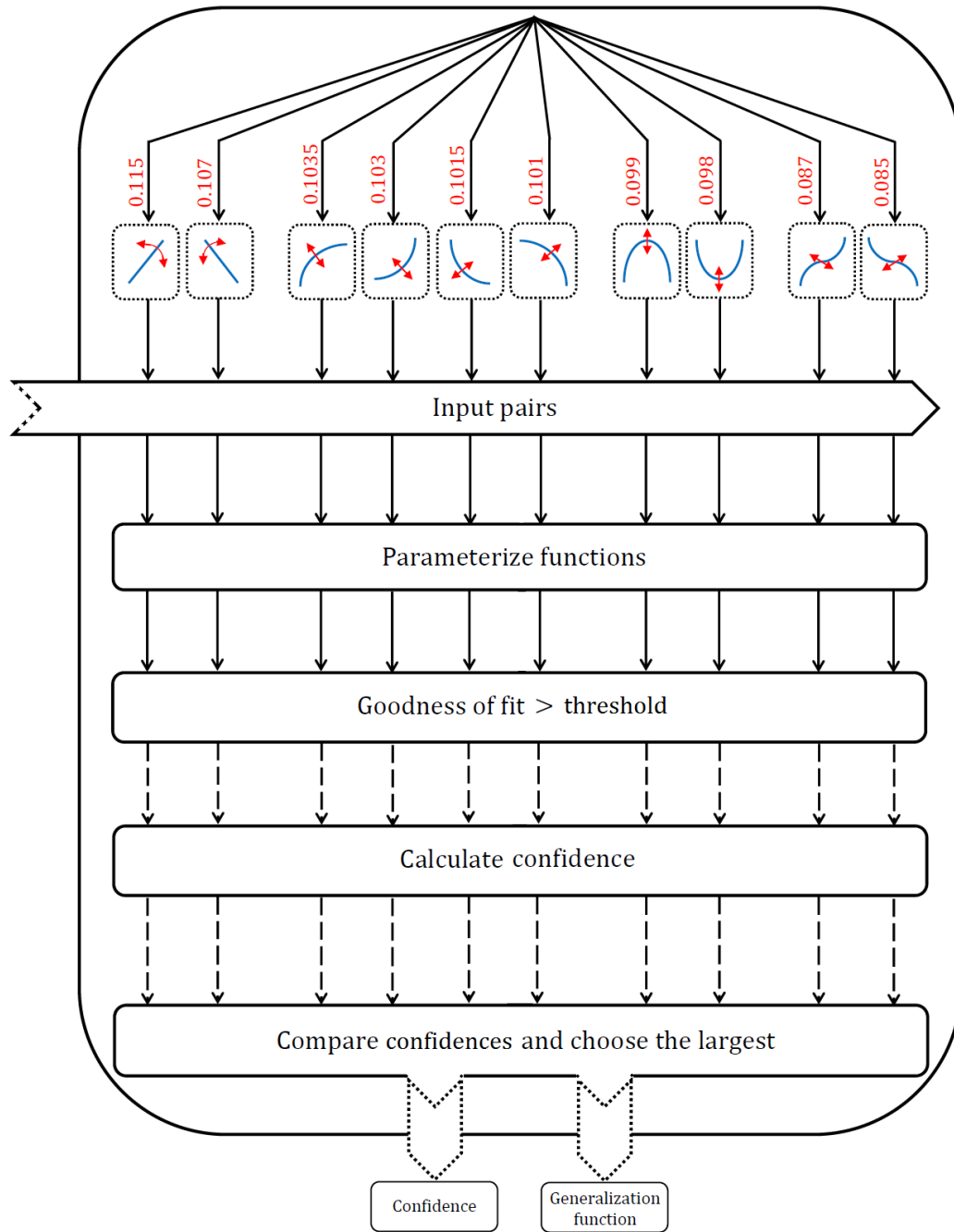


Fig. 2. An abstract graphical depiction of how the model works. It contains ten pre-defined functions with constrained parameters to limit their range of change (red arrows) and a number is assigned to each function that indicates the respective easiness. The model receives input pairs and then parameterizes, or fits, functions on those pairs. For each parameterized function, the goodness of fit is measured and if it is more than a threshold, the function is passed to the next step. Dash lines indicate that after measuring the goodness, not all functions are sent to next steps and those with low similarities with input observations are blocked. For remaining competing functions, the model computes confidences and at last, the largest confidence is chosen as the model's confidence in the

existence of a functional, or causal, link between the cue and criterion. The rule of the corresponding function is also passed for future generalizations (prediction).

which is a library of models for data fitting. After setting the lower and upper limits of change in the value of parameters for each function through Fit Options, I set the option of Robust to Bisquare which lowers the effect of outliers on the final fit. Other options were set to default.

As a measure for the goodness of fit, the 'Adjusted R-square' of the Curve Fitting Library is chosen. This is because it satisfies the conditions that were mentioned earlier; it is unaffected by the range of the change of variables and the number of points (input pairs). Generally, for every fitted function this measure indicates the correlation between actual points and those predicted by the function. For every set of pairs and a parameterized curve, the Adjusted R-square would be less than or equal to one, and values closer to one are better fits.

4.2. Design

Twenty students were recruited out of which five were used for training experiments and the other fifteen for testing the model. They were second-year engineering students from a technical college in Gilan province of Iran and received monetary reward for their participation. They were all male with the age of 20 ± 1 .

Physical mediums for presenting cue-criterion pairs were three cell phones, running a small Java application which was especially designed for this experiment. The app used wireless communication for connections between two or more cell phones, all running the app. If two cell phones were connected with this app, one must be the host and the other, guest. The host user could change the value of the cue variable through a vertical bar, between 0 and 25 with step 1, and the guest user could observe a similar bar, as the criterion variable, that changed as a function of the host. The changing bar of the guest user's app also changed the intensity of a beeping sound; this was considered to better imply the functional relation, as against mere numerical changes in the two cell phones. The effect of change in the value bar of the host could be observed by the guest only after pressing a button by the host user. So, for every pair, the host user changed the value of his bar and when he pressed the button, the value of the bar for the guest user, which was the volume of the beeping sound, changed. An approximately similar screen of cell phone for host and guest is shown in Figure 3.

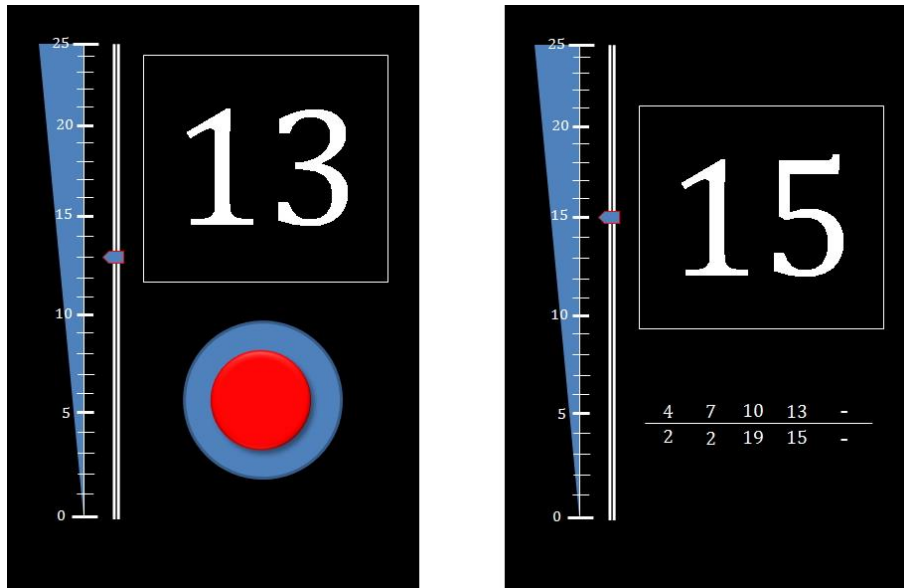


Fig. 3. Screen of cell phones, similar to the app for the experiment. The host user (left image) sets the changing bar to a value and then presses the red button. After a time lag, the guest user (right image) observes a change in his bar, which is the volume of a beeping sound. During and to the end of each trial, the cue-criterion pairs of previously presented values of that trial are shown for the guest user (at the bottom of guest user's screen), so at the end of that trial, the guest would have a better general picture of the changing pattern.

The way by which changes of values of the host changed the values of the guest must have been determined by the host user earlier. For this end, there were twenty specific functions among which the host user could choose in any trial, together with a variable for determining the level of noise with values between zero and five. Only one host user could connect to one guest user at a time; the host user could choose to disconnect from the guest and then, another host could connect to the guest.

Two host users were used during this experiment; Host1 as the primary host and Host2 the secondary. Before trials, participants were told about the general structure of the app, experiments, and were shown five functions out of twenty by which the host could manipulate the guest; these five functions were not used again during the actual test experiments. They were also aware that during the experiments and between every two trials, Host1 and Host2 could switch as the host user without participants' knowledge and during a single trial, no such switching was made.

Host2 was used as the background cause. In other words, he was there to imply the doubt to the participant of the possibility that Host2 was the cause of changes of guest values. This allowed quantifying guest users' idea about how much they are confident that Host1 was the cause. They were also told that Host1 changed the values of the bar only in increasing order, while Host2 could change the values in any order.

The primary host user, or Host1, and the guest user sat on the two opposite sides of a table, able to see the other's cell phone screen. Guest user was not allowed to touch the screen of his cell phone. Host2 sat with a hidden screen of his cell phone in a few meters distance and in every trial he acted as though he was changing the bar and pressing the button.

4.3. Task

There were twelve sets of pairs for every guest user. From the guest users' perspective, in every trial, it looked equally probable for Host1 and Host2 to control the values. Guest users could observe the screen of Host1, but that of Host2 was hidden. After observing each set of pairs and hearing the pattern of change in the intensity of the beep, first they were asked how much they were confident that Host1 was controlling their app; they were asked to quantify their confidence level from zero (no confidence) to ten (definitely confident). If their confidence was more than five, then they were also asked to predict the criterion value of three unobserved cues, one for interpolation (inside the range of observed values for cue) and the other two for extrapolation (outside the range of observed values for cue). After the presentation of every set of pairs, cell phones were gathered by the primary host user and the apps for all of three cell phones were reset, making it look probable for the guest user that Host1 and Host2 could switch their manipulative role. In abstract terms, the question of whether there exists any functional relation between cue and criterion is translated in the experiment as whether there is any relationship between Host1 and guest.

4.4. Results

4.4.1. Training set

In previous section, one assumption was that humans could only recognize the existence of a functional relationship between two variables only if the goodness of fit of input pairs, with respect to the standard curves of pre-defined functions, was higher than a threshold. In order to

determine the exact value of this threshold, the results of the training set were evaluated and the conclusion was that, on average, if a set of cue-criterion pairs are to be recognized as functionally related, the Adjusted R-square (as the measure of the goodness of fit) of a fitted function on those pairs must be higher than 0.8 (Adjusted R-square is less than or equal to 1). So in the model, when functions are fitted on input pairs and their goodness of fit are calculated, those with Adjusted R-square lower than 0.8, which have low similarity with input pairs, are discarded and the confidence is only computed for the remaining functions that can pass the threshold.

Another parameter which was supposed to be derived from the data of training set was the exact value of exponentiation for the factor of sample size. The hypothesis for the exponential form of the effect of sample size was observed in the training data and resulted in the conclusion that, at least with respect to the nature of the experiments of this research, participants' confidence as a function of the sample size increased rapidly in the first few pairs (like three to four pairs) and this increase was lowered with the further growth of the sample size (like insignificant increase for more than ten pairs). So, the exact number for x , as the power of the number of pairs (n), is computed as -1.7 .

Finally, the importance of each of the three proposed factors on the confidence must be determined. This is represented in the formula of confidence with the coefficients of k_1 , k_2 , and k_3 which must be numerically valued. For this end, several sets of input pairs were presented to the participants of the training experiments and their confidence was measured. Then, for a preliminary evaluation, the coefficients of the model were valued in a way that the numeric importance of all three factors was equal; the performance of the model was measured in this way. Later, the coefficients were changed so as to increase the importance of one with respect to the other two factors. This was done for all three factors and in each case, the performance was measured. After comparing the results of each of these settings, the conclusion was that, increasing the importance of one factor against others reduced the model's performance and thus, the model would yield the best performance if the three factors had equal importance on the general confidence.

As a result, three coefficients must be valued in a way to give equal numeric importance to the three involving factors in confidence. So, at first the factor of goodness of fit was chosen as the starting factor and its coefficient (k_2) was set to the arbitrary number of 3. Since the value

for the goodness of fit (Adjusted R-square) varies between 0.8 and 1, the second factor of the formula would vary between 2.4 and 3. The other two coefficients must be valued so to give the same range of change (0.6) to the first and last factors; so, the first coefficient ($k1$) is set to 20 and the last coefficient ($k3$) is set to 4.5.

Having assigned the values of the sample size exponentiation factor (x) and the three coefficients of $k1$, $k2$, and $k3$, the resulting formula of confidence would be as follows:

$$Confidence(f) = (Difficulty(f) \times 20) + (GOF(f) \times 3) - (n^{-1.7} \times 4.5)$$

$f:1 \rightarrow 10$

Adding two scaling factors to bring the confidence measure between one and ten would yield the final formula as:

$$Confidence(f) = \left(((Difficulty(f) \times 20) + (GOF(f) \times 3) - (n^{-1.7} \times 4.5)) - 3.2 \right) \times 5$$

$f:1 \rightarrow 10$

4.4.2. Test set

As the result of training data, now the threshold for filtering functions and coefficients for calculating confidence is at hand. So, the final parameterized model is ready for use. A set of novel input pairs, as test experiments, are designed to make an initial evaluation of the resulting model. These pairs are shown to a group of fifteen participants and the comparison between their performance and that of the model, as the overall result of experiments, is shown in Table 2.

The order of the presentation of pairs during the experiment was as shown in Table 2. In model's confidence, zero means none of ten functions had a goodness more than 0.8 (trials 4, 5, 8, and 9), so they could not pass the filter. To evaluate the generalization of the model, if any, even one, of participants had a confidence more than or equal to five, he was then asked to predict the criterion of three unobserved cues, one for interpolation and the other two for extrapolation. It must also be noted that if the predictions of the model or participants for criteria was negative, it was bound to 0 (like trial 3) and its upper boundary was 25 (like trials 1 and 10).

One important result which is not included in Table 2 is the correlation between the human confidence and the model. In other words, how good the model could track changes in human confidence, during the trials. For this end, the mean absolute error between the columns of human and model's confidence is computed as 0.78 (between 0 and 10). Another measurement is the performance of the model in all predicted pairs, taken together, for

Table 2

The results of experiments on test set.

Presented data					Confidence		Generalization performance											
					Human	Model	Data space	Data	Model	Human	STD	MAE						
1	6		10	14				4	25	20.33								
	14		0	14				11	0.87	0.33								
								16	25	20.33								
2	6		10	14				2	2	2								
	6		10	14				11	11	11								
								18	18	18								
3	4	8	12	16				2	0	0.38								
	5	15	15	5				10	16.25	16.75								
								18	0	0.38								
4	4	7	10	13				-	-	-								
	2	2	19	15				-	-	-								
								-	-	-								
5	4	7	10	13				-	-	-								
	14	10	16	4				-	-	-								
								-	-	-								
6	5	7	9	11				4	17.54	19.8								
	17	13	10	10				7	3	10			10	10				
								16	2.45	1.13								
7	3	5	7	9				1	0.84	0.93								
	1	1	1	1				2	4	7			12	10	1.67	1.33		
														19	20.19	18.93		
8	5	6	7	8				-	-	-								
	5	6	7	8				0	5	1			1	12	9	-	-	
														-	-	-		
9	5	6	7	8				4	-	0								
	1	9	15	19				2	2	19			15	9	1	-		
														16	-	0		
10	5	6	7	8				3	25	20.2								
	18	18	14	8				10	6	4			4	3	2	10	7.57	8.13
														17	0.6	0.73		
11	5	6	7	8				4	16.61	17.4								
	12	7	4	2				1	1	2			4	7	12	10	0.17	0.6
														16	16.61	17.4		
12	5	6	7	8				4	20.84	19.36								
	16	13	11	10				10	10	10			9	7	4	10	10	10
														16	0	1		

Note. The table has three parts. The first is the order and values of input pairs that were observed by participants and the model. In each trial, pairs were presented from left to right (in increasing order of the cue). The second part is the

column of confidence for humans and the model. In every trial, the human data is the average confidence of fifteen participants together with the standard deviation, and the model's confidence is computed by the presented formula of confidence. The remaining columns are the results of generalizations. In the column of Data space, blue stars are the presented data points and the red curve is the model's generalization. Trials 4, 5, 8, and 9 had goodness less than 0.8, so the model filtered all of ten functions and did not produce any confidence and generalization function. The next three columns are the asked cues (Data), model's response (Model), and the average of participants' responses (Human) asked only from those who had confidence more than five. In the last two columns, STD contains standard deviations of participants' predictions, and MAE is the mean absolute error between columns of Model and Human. Note that the MAE in the last column shows the performance of model's generalizations for every trial and its range is between 0 and 25 (to be clearly seen, it is scaled between 0 and 4).

interpolation and extrapolation tasks; the mean absolute error for all of model and human's criteria is approximately 1.1 (between 0 and 25).

5. Performance

The measure of confidence in a functional link and the model as a whole were originally conceived to operate in a task similar to, but not exactly like, the presented experiment. The results of the experiment were, nevertheless, better than what was initially expected. In spite of many shortcomings in the experiment for a complete evaluation of the model (like the small number of participants, their equal age and sex, limited number of trials, etc.), there are some noteworthy points in analyzing the results of the experiment.

5.1. Confidence

The primary goal of the research, which is measuring human confidence in the existence of a link between cue and criterion, was to a large extent successfully attained. With mean absolute error of 0.78, the model had an acceptable general performance in tracking human confidence during the entire experiment.

Trials 1 and 3 were the two challenges of the model and the respective differences of 2.32 and 1.77 between the average human and model's confidence were comparatively high. That is probably because of a small sample size and relatively high difficulty of functions. This problem is mediated in other nonlinear trials like 7, 11, and 12, where the sample size increases. An easy solution would be to increase the numeric importance of the factor of sample size in calculating the confidence (increasing the value of the coefficient k_3); but this only solves the problem of

difficult functions and for easier ones, like linear, the sample size does not play the same important role. Thus if the model was to use one single formula of confidence in its minimalist approach, a trade-off is required in the importance of sample size between easy and difficult functions and it seems that in the rule for confidence this trade-off is reached.

Trials 4, 5, 8, and 9 pose another problem to the model. In these trials, the model's confidence remains zero while human confidence varies between 0.2 and 2. In general terms, the model is unable to track subtle changes of human confidence in trials with average small confidence. One probable explanation can be that some participants tend to recognize the presented patterns easier and with more confidence while others are more skeptical. For example, in trial 8 participants saw a linear increasing order in the first four pairs; this gave participants a level of confidence in the existence of a link. The six pairs that followed divided the participants to two groups; some who still thought there may be a link and others who rejected any link. Likewise in trial 9, some participants recognized the difficult pattern while others considered it to be a complete noise. Accounting for individual differences and defining a clear-cut distinction between noise and recognizable patterns are two challenges for many studies in psychology, cognitive science, and machine learning.

A relevant issue is the observation of comparatively high standard deviations in participants' confidence for more difficult functions. Trials 1, 3, 11, and 12 which are difficult functions had respectively the highest STD among trials; while easier functions of trials 2, 6, and 10 had smaller STDs. This shows that people are more unanimous in learning easy functions.

5.2. Prediction

Although prediction or in other terms, transfer phase or generalization, was the secondary goal of the study, the model had a successful performance in predicting criteria for unobserved cues, at least with respect to the trials of the experiment. This performance is measured by the mean absolute error of all presented pairs, which is approximately 1.1.

To further analyze the generalization performance, the MAE of each trial is presented in Table 2. In all trials, increasing and decreasing trends were successfully tracked. In trial 10, however, the model chose a decreasing power function while most participants probably recognized a linearly decreasing function. Although this problem is not explicitly evident in trial 10, the recognition of different functions by the model and humans would result in a major

deviation of predictions in extrapolation, especially for pairs in further distances of cues from the presented data. One solution is to assign proportionately larger values of easiness (smaller difficulties) to linear functions, but this may only solve this single case; it downplays nonlinear functions and would result in recognizing most sets of input pairs as linear functions. So, in assigning numeric values for difficulties among functions, another trade-off is needed that reconciles between appreciating the relative easiness of linear functions and leaving enough chance for nonlinear functions. If this trade-off is reached, problems like trial 10 are inevitable.

Another challenge of the experiment design in generalization is the case of trial 9. What if one or many participants' confidence is more than five but the model's confidence is zero? In trial 9, one participant recognized the difficult pattern and his confidence was more than five; so, according to the procedure, he was then asked to predict. The model, however, was unable to recognize the pattern, since no function was considered in advance as a pre-defined function to address a set of pairs similar to those in trial 9. Including more complex functions for difficult sets of pairs would solve this problem but as mentioned earlier, such model would face so many further issues, like the risk of recognizing noise as functions.

5.3. Model's sensitivity to modifications

After forming the hypothesis of how to measure confidence and tuning the coefficients to the training set, the model was tested and was comparatively successful with respect to the test set. Now, the question is: Is it possible to have identical, or nearly identical performance if some parts of the model, or its parameters, are modified?

In order to fully answer this question, all of the considered parameters and variables must be modified with respect to other parameters and in each setting, model's performance should be measured and compared with the already presented setting. Since doing so is too cumbersome and outside the general goal of this article, only the performances of the two variations of the model are compared with the standard model. These modifications are made to the confidence measure as the heart of the model.

5.3.1. A confidence solely based on the goodness of fit

In the first variation, instead of using the three factors which were involved in the calculation of the confidence, only the factor of the goodness of fit, or noise, is used as the only effective factor

in the final confidence. For this end, in the presented formula of confidence, the first and the last terms are omitted and only the second term is used. As mentioned earlier, the goodness of fit of the computational implementation of the model is measured with the Adjusted R-square of the Curve Fitting Library of MATLAB. Because the confidence is measured between one and ten and the Adjusted R-square is a number less than or equal to one, the Adjusted R-square is considered between zero and one and is then multiplied by ten. As a result, all parts of the model remain intact, except the confidence measure which becomes as follows:

$$Confidence(f) = GOF(f) \times 10$$

$f:1 \rightarrow 10$

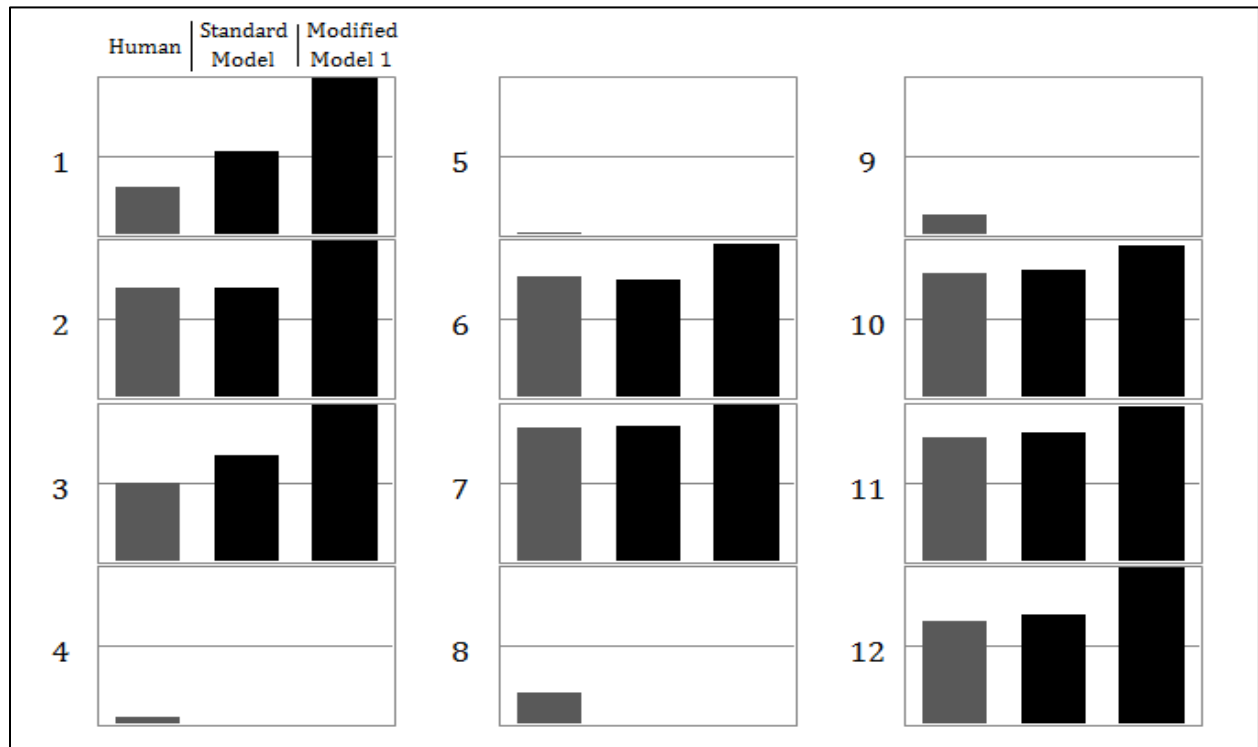
The model which is resulted from this modification is presented with the same twelve sets of pairs that the participants and the standard model perceived. Table 3 depicts the performance of the modified model in those observations, in terms of confidence. As can be seen from the table, the performance of the modified model is significantly reduced. It is interesting to note that in most cases, the confidence of the modified model is nearly binary; this is due to two reasons: The existence of the threshold by which functions with Adjusted R-square less than 0.8 are filtered, and the absence of the other two factors (difficulty and sample size) which seemed to balance the final confidence as a result of their effects. In fact this shows the importance of the other two factors in tracking subtle changes in confidence and, in effect, justifies their existence.

To be specific, the mean absolute error of the confidence of the modified model is 2.48 (out of 10). The standard model, with the MAE of 0.78, has therefore a better performance than the modified version. With regard to prediction, in two of the twelve trials, the functions that were chosen were different than those in the standard model. In trial 6, instead of a negative linear function, a negative cubic polynomial (right-most curve in Figure 1) is chosen, and in trial 7, instead of a power function, a positive cubic polynomial is recognized. This is while in those trials, the MAE of the prediction of the standard model (or generalization as shown in Table 2) was small and the prediction of human behavior was quite successful. As stated earlier, although this recognition of different functions by the modified model does not result in major difference with the standard model in prediction in the given range of the experiment, if we try to test the generalization performance in further ranges for the value of the cue (longer extrapolation performance), the reduction in generalization performance becomes more obvious.

Another result of testing this modified model in prediction is that, if we depend only on, or give higher importance to, the factor of goodness of fit in the calculation of confidence, it becomes more likely for most sets of input pairs to be recognized by one of the cubic polynomials; so, by using the factor of difficulty we can curb cubic polynomials in fitting input pairs and therefore, only sets of pairs that are highly similar to the curves of cubic polynomials are recognized by them.

Table 3

Comparison between the confidence of the standard and a modified model that only depends on the goodness of fit.



5.3.2. Omitting the effect of sample size

In the other variation of the model, both factors of difficulty and the goodness of fit are used to calculate confidence and the difference of this variation with the standard model is the absence of the effect of sample size in confidence. If the coefficients of the factors of difficulty and the goodness of fit satisfy equal numeric effect of these two factors on the confidence, and if the resulting confidence is scaled between one and ten, the formula of confidence would become as follows:

$$Confidence(f) = \left(((Difficulty(f) \times 20) + (GOF(f) \times 3)) - 4 \right) \times 8$$

$f:1 \rightarrow 10$

As can be seen in Table 4, the performance of a model with this formula of confidence is also reduced, although not as much as the previous modified model. The mean absolute error of confidence between this model and human performance is 1.44 (compared to the MAE of 0.78 of the standard model). In addition, compared to the previous variation, all of the functions that are chosen are those of the standard model, since unlike difficulty and goodness of fit, sample size is measured regardless of the type of function that is chosen.

5.3.3. Conclusion

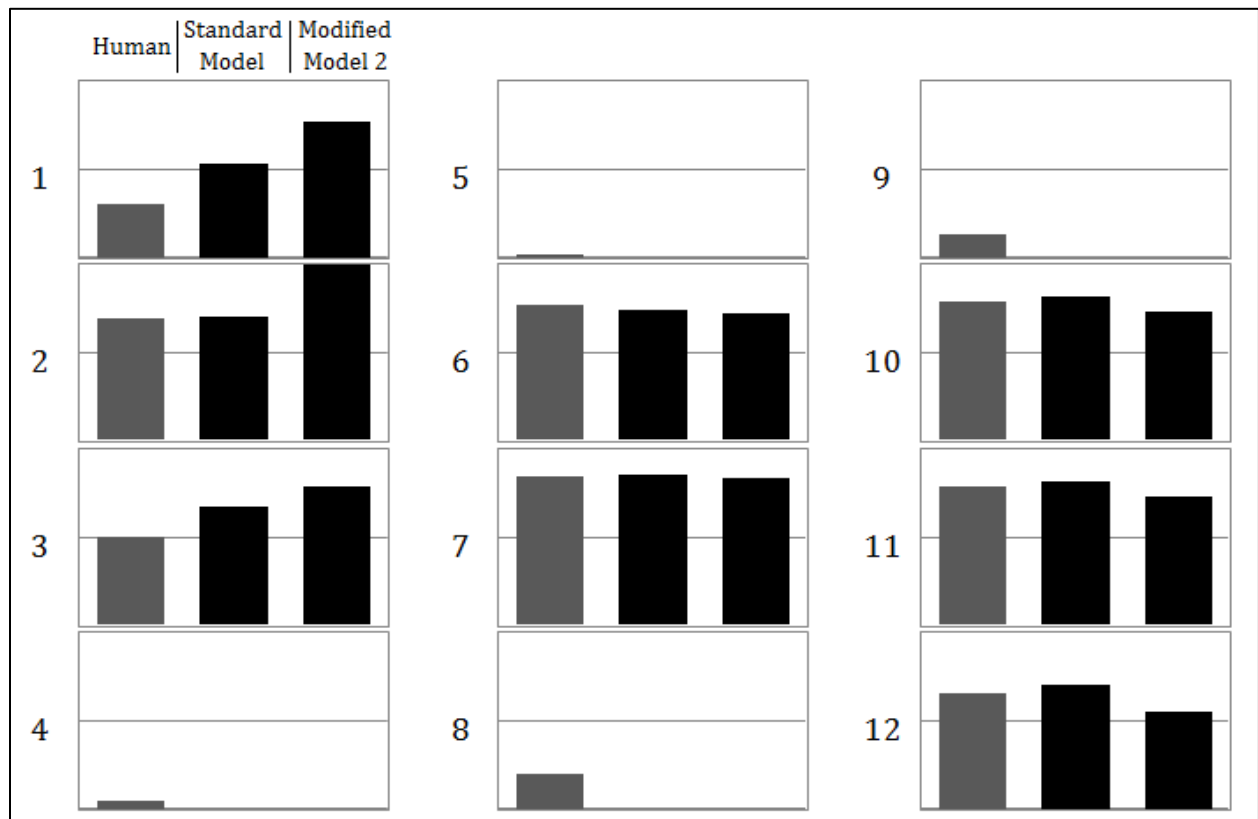
From the standard model, two variations were derived in which the confidences were measured in two different ways. In the first variation, the two factors of difficulty and sample size were omitted and the MAE of confidence was measured as 2.48. For the second variation, the factor of difficulty was added and the MAE reached 1.44. And finally if all three factors are considered together, the MAE becomes 0.78. From these two variations of the model, some important results can be gained.

First, these results indicate that, although the coefficients of the three factors were measured so as to have the same numerical effect for each factor on the final confidence, the importance of the role of the three factors is not equal. Specifically, the goodness of fit is the pivotal factor and the confidence measure can barely be imagined without this factor. This is because it would be hardly imaginable to use the factor of difficulty without the factor of goodness. Also, it would not be possible to define the filter that we used in the model. Compared to the goodness of fit, the other two factors, namely difficulty and sample size, can be seen as contributing factors for increasing the performance of the model. It must also be noted that the factor of difficulty has a role other than increasing performance; it differentiates between functions and allows easier selection of easy functions. Without its role in the experiment, in two trials, functions other than those of the standard model were chosen. As mentioned, the importance of this role of difficulty may not be much obvious with respect to the experimentations of this research; it only shows itself when we test the extrapolation performance in larger ranges for the cue.

Likewise, it may be tempting to underestimate the role of sample size in confidence, based on its relatively marginal influence on increasing performance. Upon inspecting the results of the second modified model, we can recognize that in trials with average sample size (like 6 and 7) and those in which the model does not recognize a relationship (4, 5, 8, and 9), the importance of sample size is insignificant; but in trials with small or large sample size (like trials 1, 2, 3, 10, 11, and 12), it's influence is more evident. If different experiments with more extreme cases (very small or very large sample size) had been conducted, the positive impact of the factor of sample size on increasing the performance of the model would have been more obvious.

Table 4

Comparison between the confidence of the standard and a modified model in which the effect of sample size is absent.



As a result, although the importance of the role of the three factors on final confidence is not equal, each factor is making a significant contribution in the general model. So, the inclusion of these factors helps in building a more robust model, especially when it is tested on more diverse, or more extreme, trials. This may not be much obvious in the limited experiments of this research, but even here, dropping the factors of difficulty or sample size resulted in noticeable decrease in the performance of the resulting models.

6. General discussion

We have seen that a measure based on difficulty of functions, goodness of fit, and sample size could successfully be used to quantitatively model human confidence in the existence of a functional link between two candidate variables of cue and criterion. The calculation steps of this measure then formed a mathematical, rule-based model of function learning. The resulting model receives a set of pairs as input and gives confidence and prediction function as outputs.⁸ Because this topic has been quite absent from function learning literature, it has implications on relevant debates and can help to shed light on some of the choices in the field.

6.1. The order of the difficulty of functions

As is mentioned in Section 3, pre-defined functions of the model and their respective difficulties are loosely derived from Bussemeyer et al. (1997) with two exceptions from that research. They argued that models of function learning must be capable of learning cyclic, or periodic, functions. Whether humans are capable of learning cyclic functions is a moot point. Byun (1995) argues that humans can learn cyclic functions, like sine or cosine, while Kalish (2013) shows that “... learning periodic functions are extremely difficult” and in rare occasions, if not impossible. In the presented model of this article periodic and other complex functions are not included, for reasons presented in Kalish (2013) and two other reasons. Input pairs in trials 5 and 9 could be approximated by certain cyclic functions, while the results show low levels of human confidence in the existence of any functional relationship in those trials. This observation supports the hypothesis that humans have difficulty in recognizing cyclic functions and in effect, agrees with the conclusion of Kalish (2013). In addition, if we are to fit functions on input pairs, any set of

⁸ The presented model is not restricted to work with single-cue tasks. Defining functions, fitting them on incoming pairs and parameterization, and other crucial parts of this model can easily be extended to work with multiple-cue tasks.

noise (or points without regularity) can be closely approximated with a cyclic function. It is not clear how to exactly distinguish a complete set of noise with a sampled cyclic function.

Another exception is that in the model, two curves that can take monotonic shapes, cubic polynomials, are considered to be more difficult than two non-monotonic functions of quadratic polynomials. This ignores some of previous findings in psychology since non-monotonic curves are generally thought to be more difficult than monotonic curves. For example, Delosh (1995) shows that both linearly decreasing and exponentially increasing functions are learned faster than non-monotonic quadratic functions. Although in most of those experiments monotonic curves were easier than non-monotonic ones, it is important to be cautious in making a general rule. We cannot generalize those findings to all functions and there may exist some complex monotonic functions that are more difficult than quadratic functions. According to my knowledge, there has not been similar experimental research to show whether non-monotonic quadratic functions are more difficult than monotonic cubic polynomials. Besides, in the current research if a cubic polynomial is treated easier than any other functions, it would be highly probable that it outperforms those functions in fitting, replacing all functions with smaller values of easiness. Now that cubic polynomials have the smallest easiness, only sets of input pairs that are exclusively similar to cubic polynomials are recognized as such.

6.2. Rule-based vs. associative approach

The assumption of this research is that three factors of difficulty, noise, and sample size influence human confidence in the existence of a functional relation, or link. Instead of the rule-based approach of this study, if an associative model was used with the same assumption, there would have been difficulty in defining noise, which is the pivotal factor among the three factors in the confidence measure. In fact, one of the challenges of associative models is their inability to differentiate noise from a learnable function.

Another problem with associative models is their lack of clarity; the internal structure of associative models, upon training, are complex and must be translated in order to be understood, while rule-based models are easier for comprehension and manipulation. Associative models are also highly sensitive to training samples as their primary source of learning, while in the presented model, only some parameters are tuned to the training data. In other words, the general rule-based approach of this study provided the extensive use of prior knowledge, both in

embedding pre-defined functions and shaping the general structure of the model. Setting a threshold for the goodness of fit and measuring noise, defining difficulties of functions, and presenting a formula for calculating confidence are among the instances of previous knowledge embedded in the model of this research and it seems that if in the future, further previous knowledge is needed in a model, associative models will have serious difficulty in accounting for them.

6.3. Algebraic vs. Bayesian modeling

With respect to the presented model, one question might appear in mind: Why not using a probabilistic, like Bayesian, model for the same purpose? The final goal of the presented model is to measure human confidence; this confidence can also be translated to the measurement of human doubt in certain tasks and if so, the rules of the probability theory, and for example Bayesian modeling, is the first standard tool that comes to mind. In the domain of modeling human cognitive processes and especially in judgment and decision making, most of the attempts in recent years have been focused on Bayesian modeling. Although the presented model is in some respect similar to a standard Bayesian approach, it is an algebraic model which does not conform to the standard rules of probability.

6.3.1. Prior knowledge and present data

In most of previous attempts at modeling function learning, knowledge before observations, or prior knowledge, are conceived and included in models. For example, in an attempt to account for the frequency of encountering different functions in the world, or addressing human inductive bias, Lucas et al. (2015) and Schulz et al. (2015) considered a body of knowledge as prior probabilities. Also in binary-valued causal induction and specifically in Griffiths et al. (2011), researchers focused on a specific task of causal inference, namely the *Blicket Detector*, and tried to approximate human causal inference from observations.

The goal of Griffiths et al. (2011) can be considered as a specific instance of the general objective of the present research; the authors attempted to quantify human confidence of which object was the cause (or *Blicket*), having seen the detector's reaction to the placement of the two objects of A and B. Then, through a Bayesian model, the prior knowledge was combined with the present observations, or likelihood, in order to model human confidence, or doubt.

Upon reflection, it becomes clear that the specific causal task that Griffiths et al. (2011) were working on, or the Blicket Detector, was perfectly suitable to the structure of a Bayesian model and therefore, could easily be formulated into the language of the Bayes' rule. Their final model could solely be used for their specific causal inference task, or at most, any task in the form of the experiments on the Blicket Detector. In general, for any problem or task to be modeled with a Bayesian approach, it should conform to the conditions of Bayesian modeling. Two important of such conditions are: The problem should take the form of the logico-probabilistic hypothesis space of Bayesian models, and the observed data should be able to update the probability of each hypothesis.

The model of the present research also contains prior knowledge and the primary output of the model is human confidence. Although these elements remind us of using a standard Bayesian model, the exact goal of the current research does not allow the problem to be translated into a Bayesian model. What we ask the final model of this article is: How much confident are you that there is a causal, or functional, link between the two variables of cue and criterion? If we are to create a Bayesian model with this question, the only imaginable hypothesis space would contain two logical alternatives; either the cue is linked to the criterion or it is not. The biggest obstacle of this scenario is how to update prior probabilities with the observed data in a way that is used in the presented model of this research?

Therefore, it is possible to conceive models that contain previous knowledge which cannot take the form of a standard Bayesian model. The knowledge that I embedded in the model, or in other words, the corresponding values of difficulty for functions, are not in any sense, probabilities; they are the difficulty of learning various functions in the world, as derived from Busemeyer et al. (1997).

As another reason for deviating from the norm of using a Bayesian model, the proposed confidence measure is computed as a linear combination of its effective factors, together with their coefficients. This is not in accordance with the standard Bayes' rule in which prior probabilities are multiplied by the observed data, or likelihood. In the current research, because the prior knowledge (functions and their difficulties) and the observed data (noise and sample size) are composed of non-probabilistic terms, their combination in the final confidence cannot follow from the standard probability or Bayes' rule.

From another perspective, most of the Bayesian models of causal learning that have so far been presented focused on specific causal tasks with a certain medium and format. This allows easy division of problems into their logical hypothesis space which is a prerequisite of Bayesian models. Although this results in more accurate models of human behavior through standard probability theory, it highly limits their performance in different causal tasks. The model of this article, on the other hand, is a very generic model which claims to work in a more abstract level, even though with higher generalization error among various function learning tasks. It is difficult to expect from the models in this level of operation to predict all possible logical hypotheses of problems in advance and build Bayesian models accordingly.

6.3.2. The assumption of a probabilistic mind

In general, the research of this article was started by the question of how to quantify human confidence in the existence of a relationship between two variables. In that direction and based on some previous investigations in psychology, the attempt in addressing the question resulted in proposing a hypothesis for choosing the three effective factors, their combination, and a general model of function learning. As a result, the first and the simplest implementation of that hypothesis, which was the algebraic formula of confidence, was used and in order to calculate each of the three factors, the general structure of the model was formed. It is important to note that the proposed hypothesis is the central subject of this study and it was formalized regardless of the possibility of whether it can be implemented by or conforms to a Bayesian model. As long as any other mathematical formula and a general model are faithful to that hypothesis, they are valid.

Based on previous investigations in the literature of probabilistic modeling and more abstract discussions in the domain of the probability theory, some would argue that human behavior can best be modeled by the probability theory (e.g., Griffiths et al., 2008; Griffiths & Tenenbaum, 2006), especially in the tasks that involve human uncertainty. Most of the models of human behavior which have been proposed in recent years are based on probability theory and in accordance with this claim. Without entering the discussions concerning the origin and extent of the probability theory, the claim above rests on the belief that human mind works entirely or to a large extent probabilistically or Bayesian, or at least can best be approximated by those

frameworks. Therefore, the probability theory is treated as a gauge on which the best human inference is based.

It is important to be conscious of such an important assumption and be open to the possibility of other assumptions and beliefs on which different approaches in modeling would be based on. As an example of a different belief as the starting point, some of other researchers argue that human cognition and behavior is fundamentally different from our formalizations as probabilistic rules (e.g., Tversky & Kahneman, 1974) and so, the Bayesian mind is an illusion, or at best an imaginary approximation to what is really happening in human mind. An approach to modeling derived from this perspective would not claim that the resulting model is depicting the human cognitive processes, but instead treats human mind as a black box system with specific inputs and outputs. So, the resulting model of such an approach utilizes tools from any field to build a mathematical system that tries to approximate human judgment and predict human behavior.

Moreover, there are evidence indicating systematic deviations of human behavior and judgment from probability rules (e.g. see Vul et al. (2014), although the authors tried to reconcile the mismatch inside the scope of Bayesian modeling). In those cases, we should not consider those patterns of human behavior as ‘mistakes’, ‘abnormalities’, or ‘exceptions’; instead they must be accepted as a part of natural human behavior and if we are to model it at any level, we should try to account for those patterns in our models, regardless of considerations like their compatibility with probability theory.

Since this argument is still an unresolved matter, treating the probability theory as a standard in modeling and expecting all proposed models to be in conformity with it, is unwarranted. It has the danger of eliminating those hypotheses that cannot be translated into the probability rules and forces various models to change their structures to be concordant with probabilistic or Bayesian models. Generic mathematical or algebraic models, like the model of this article, are highly flexible and allow easy changes of their structure with respect to any regular set of human behavior.

6.4. Limitations

In order to simulate human confidence in the recognition of functional relations, this study must be based on the assumption that human confidence can be quantified. Having recognized a

functional link between two variables (in model's term, if the confidence is more than zero), the model takes it for granted that there is a level of confidence in human judgment and this level can be translated as a numeric value. In fact, the psychological states of confidence and expectation are represented merely as a number and a mathematical rule. Even if we believe in mental mechanisms as computational processes, this ignores the complicated processes in human mind with respect to the real-world function learning tasks.

In addition to the problem of quantification and in evaluating the performance of the model in this article, participants' report of their own confidence was used as a reference for the model's assessment. This assumes that people are explicitly conscious of their own learning and confidence, while large swaths of psychological findings show that human learning is unconscious and implicit. With respect to this study, how can we be certain that each participant reflected and quantified his own confidence accurately? Supposing participants' scrupulousness, it is possible that at least some of them were not precise in their self-report for confidence. This is the problem of the experiment and in order to assess human post-learning confidence more accurately, we need a more comprehensive experiment design that can measure human confidence through post-learning behaviors in natural settings.

Aside from general problems of the model and experiments, and in addition to the issues discussed in analyzing the model's performance, there are some other specific limitations. In assessing the model's confidence in a trial, if the incoming pairs resemble a line with a very small slope, like 0.01, and if there is no level of noise, both the model and probably humans will recognize the relationship by testing it in larger ranges of cue. But if there is some noise and the range of values for cue is not that much to recognize the slope with respect to the level of noise, the model will recognize the line but humans will not be able detect any relationship, since the change of the value of cue does not result in a detectable increase or decrease in the criterion. With the same level of noise, if the slope were larger (or smaller) humans would not face difficulty in recognizing a linear pattern.

Another issue is the famous criticism on function learning and causal induction models for their indifference to the order of observations. The model of this research suffers from the same problem and for a certain set of input pairs, it would not change its judgment if the order of the presentation of those pairs were different.

Moreover, the experiment of this article was limited to a certain design of a certain physical system of functional relation. This is while the model is supposedly designed to operate in a wide variety of settings. There is no guarantee that the model would have the same successful performance if I had used a different system for functional relation than a physical system, if there was a different physical system, or even if the design of the same physical system was different. In other terms, the performance of the model in this article was specific to the physical system of cell phones and the objects and environment of the experiment design. Even subtle changes to the involving factors of the experiment would probably influence the model's performance.

One solution to most of the problems and limitations, which are not specific to this article, is to add additional parts to the model in order to account for subtleties and exceptions. In that case, different parts of a model, or different models, would be responsible for various contexts and different systems of functional relations. Instead of building a general and simple model for function learning and recognition, there would be numerous models, each for one system and context. A general model of function learning with that perspective would be a complex system of systems with the aim of simulating human function learning, in general. That view is in contrast to the goal of the present research which is to address function learning and recognition with a simple mathematical and computational model.

Although this solution would probably solve many problems, it would make the system too complex and even in that case, we would not be certain if it could address all cases of function learning tasks, since in that level of generalization, function learning is closely relevant to other elements of the mind, like logic, memory, and perception. Besides, how does such a general model recognize a specific system and context to use the part of itself responsible for that system and context?

The presented model, as is clear from the problems and limitations discussed, suffers from being too much average over real-world examples. In fact, although the model was successful during the experiment, it would probably have major errors when used in different contexts with different functional systems. The idea of our model stems from the assumption that a single mental mechanism is responsible for human function learning. But the assumption becomes unwarranted if we look at real-world examples and how humans operate in various situations. This is not merely the problem of function learning, but it seems to be a general

criticism on similar assumptions of cognitive psychology and artificial intelligence. So, the model of this article and similar models have a very limited usage in some domains. It is a simplified recommendation for average tasks and it does not claim to model human function learning in general.

6.5. Alternative perspectives and future directions

This study showed that in recognizing a link between two continuous variables, human performance can be considered as optimal or near optimal. In other words, if people do not clearly measure noise, difficulty of functions, or the number of pairs, their performance in function recognition and learning can be closely simulated with a system based on measuring those factors, at least with respect to the physical medium of cell phones and the design of the experiment. Aside from the view of optimality, the same measure of confidence can be implemented in a different way. What if we plot the incoming pairs on a plane and compare the similarity of the resulting shape with pre-defined functions' shape? In that case, the more similar the incoming pairs are to one of function's shape, the more confidence would be in the existence of a functional link.

One recommendation is to use an optical recognition system to recognize the plot. For example, we can train an artificial neural network for optical recognition of certain functions (like the ten functions presented in this article) and use it to detect those functions from input pairs that may be noisy or with a few samples. This suggestion is no longer a rule-based approach, since it would not assume that people are learning explicit rules. Instead, it would be an approach based on the assumption that people learn functional relations in the same manner that they visually recognize certain regular shapes; or that human performance in function learning can better be modeled with the help of optical recognition systems, regardless of the mechanisms of human mind. Relevantly, we can use the literature of human visual recognition of curve-like shapes and so, bring an analogy or make reconciliation between human recognition of visual and functional curves.

This article presented one solution to the problem of human confidence for links in function learning tasks. A helpful line of research in future would be to use a different approach, like an optical recognition system, for this problem and compare the results with the presented model of this article.

References

- Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, 11, 1-27.
- Bussemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts and D. R. Shanks (Eds.), *Knowledge, Concepts and Categories: Studies in Cognition* (pp. 408–437). Cambridge, MA: MIT Press.
- Byun, E. (1995). *Interaction between prior knowledge and type of nonlinear relationship on function learning*. Purdue University: Doctoral dissertation.
- Carroll, J. D. (1963). *Functional learning: The learning of continuous functional mappings relating stimulus and response continua*. NJ: Education Testing Service Princeton.
- DeLosh, E. L. (1995). *Hypothesis testing in the learning of functional concepts*. Purdue University: Master dissertation.
- DeLosh, E. L., Bussemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 968-986.
- Griffiths, T. L., Kemp, C., Tenenbaum, J. B. (2008). Bayesian models of cognition. In Sun, R. (ed), *The Cambridge Handbook of Computational Psychology* (pp. 59–100). Cambridge University Press.
- Griffiths, T. L., Lucas, C., Williams, J., & Kalish, M. L. (2009). Modeling human function learning with gaussian processes. In *Advances in Neural Information Processing Systems*, (pp. 553-560).
- Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and Blickets: Effects of Knowledge on Causal Induction in Children and Adults. *Cognitive Science*, 35, 1407–1455.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51, 285–386.
- Griffiths, T. L., & Tenenbaum, J. B. (2006). Statistics and the Bayesian mind. *Significance*, 3, 130–133.

- Kalish, M.L. (2013). Learning and extrapolating a periodic function. *Memory & Cognition*, 41(6), 886–896.
- Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychological Review*, 111, 1072-1099.
- Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 811-836.
- Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, 22, online.
- McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, 12, 24-42.
- Narain, D., Smeets, J. B. J., Mamassian, P., Brenner, E., & van Beers, R. J. (2014). Structure learning and the Occam's razor principle: a new view of human function acquisition. *Frontiers in Computational Neuroscience*, 8, article 121.
- Schulz, E., Tenenbaum, J. B., Reshef, D. N., Speekenbrink, M., & Gershman, S. J. (2015). Assessing the Perceived Predictability of Functions. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Vul, E., Goodman, N., Griffiths, T. L., Tenenbaum, J. B. (2014). One and Done? Optimal Decisions From Very Few Samples. *Cognitive Science*, 38, 599–637.