Behavior Research Methods
2006, 38 (2), 218-228

NUANCE 3.0: Using genetic programming to model variable relationships

GEOFF HOLLIS and CHRIS F. WESTBURY
University of Alberta, Edmonton, Alberta, Canada

and

JORDAN B. PETERSON
University of Toronto, Toronto, Ontario, Canada

Previously, we introduced a new computational tool for nonlinear curve fitting and data set exploration: the Naturalistic University of Alberta Nonlinear Correlation Explorer (NUANCE) (Hollis & Westbury, 2006). We demonstrated that NUANCE was capable of providing useful descriptions of data for two toy problems. Since then, we have extended the functionality of NUANCE in a new release (NUANCE 3.0) and fruitfully applied the tool to real psychological problems. Here, we discuss the results of two studies carried out with the aid of NUANCE 3.0. We demonstrate that NUANCE can be a useful tool to aid research in psychology in at least two ways: It can be harnessed to simplify complex models of human behavior, and it is capable of highlighting useful knowledge that might be overlooked by more traditional analytical and factorial approaches. NUANCE 3.0 can be downloaded from the Psychonomic Society Archive of Norms, Stimuli, and Data at www.psychonomic.org/archive.

Genetic programming (GP) is a paradigm for automating the process of computer programming. It works in a fashion analogous to selective breeding in biology. The user provides two elements: an operational definition of the goal, and a set of operators and operands that can be used to achieve that goal. In selective breeding, the goal may be to come up with a smaller dog or a cow that produces more milk. In GP, the goal can be the optimization of any well-defined function, from maximizing food collected by virtual creatures to minimizing error in a regression equation. The important point in all these cases is that the *fitness* of a candidate solution is quantifiable. As long as the dog or error is getting smaller, a solution is getting better.

In selective breeding, the operators are genetically specified and are often (until recently, always) only implicit from the breeder's point of view. A dog breeder can create a smaller dog by selective breeding without ever knowing which genes his directed mating is affecting. Relatives of GP such as genetic algorithms (Holland, 1992) are analogous, because they use arcane problem-specific binary representations for a solution. In GP, however, the operators are explicitly specified, consisting of well-defined, general computational operations such as addition, subtraction, square root, and log.

This work was made possible by a National Engineering and Science Research Council grant from the Government of Canada to C.F.W. Correspondence concerning this article should be addressed to C. F. Westbury, Department of Psychology, University of Alberta, P220 Biological Sciences Building, Edmonton, AB, T6G 2E9 Canada (e-mail: chrisw@ualberta.ca).

When the goal and operators are defined, GP proceeds by creating a large set of computer programs (agents) that combine the operators in random ways. Each agent in the population attempts to solve the problem. The agents that perform best at this task are selected out and mated (duplicated, broken apart, and recombined with each other in random ways) to form new agents. These new agents and the best agents from the previous generation are used to create a new population pool. This process—test, select, and mate—is repeated until a completion criterion specified by the user is met (e.g., until a certain amount of time or number of generations has elapsed, or until one agent gets close enough to the goal).

Recently, we have developed the Naturalistic University of Alberta Nonlinear Correlation Explorer (NUANCE 2.0) (Hollis & Westbury, 2006), a platform-independent program written in Java that uses GP to model nonlinear variable relationships. When NUANCE was introduced, it was demonstrated to work on two toy problems. The focus of the present studies is to demonstrate that NUANCE can be applied to real problems in psychology. In this article, we introduce and use a new version of the program, NUANCE 3.0. This tool and a manual that includes a description of parameters and new features added since the previous version are available as a free download from the Psychonomic Society archive at psychonomic.org/archive. Our aim here is to demonstrate that NUANCE can be applied to real psychological problems, revealing new findings with both utilitarian and theoretical value. To this end, we apply NUANCE to very different data sets. The first study is a short example having to do with pharmacists' prescription errors. This serves as an example of how GP can be used to simplify predictive models. The

second, larger study has to do with predicting lexical decision reaction times (LDRTs). It illustrates some of the advantages of nonlinear regression and provides several examples of how GP can enhance understanding of complex sets of data involving many dependent variables.

STUDY 1

The ability to predict human performance can be useful in applied psychology. Peterson (2005), for instance, looked at the domain of the pharmacist. Pharmacists can make errors in the prescriptions that they give to customers. As an example of how common such errors are, Peterson cites a survey showing that 34% of Texan pharmacists have an error rate of greater than one prescription error per week. This is extremely undesirable, of course, because peoples' health depends on the accuracy of such prescriptions. Most attempts to correct the problem of misprescription errors have focused on refining the process of dispensing prescriptions in general, by using better labeling of pharmaceutical products and developing methods to automate the process. Very little attention has been focused on studying how individual differences among pharmacists might relate to prescription errors. Such research might reveal new methods for dealing with substandard job performance.

Peterson (2005) undertook a study to discover how individual differences might play a role in errors for dispensing drug prescriptions. He assessed pharmacists with a battery of cognitive tests sensitive to frontal lobe functioning, assessing decision making, error monitoring, planning, problem framing, and novelty analysis. Using the results from this battery, Peterson was able to correctly classify 77.4% of pharmacists as having been or not having been reprimanded for making prescription errors (60% correct for reprimanded; 85.7% correct for unreprimanded). Such information is useful, because it may suggest methods for identifying pharmacists at risk for making errors, as well as intervention methods to reduce their error rates.

We were interested in trying to improve on Peterson's results, using NUANCE, by increasing the classification accuracy and/or by developing a simpler classification strategy. Peterson's classifications were based on a logistic regression analysis incorporating performance ratings on seven different tasks sensitive to prefrontally mediated cognitive ability. Finding a simpler strategy should make identifying and dealing with pharmacists at risk for making errors more feasible in practice.

Method

Stimuli. This study used Peterson's (2005), with two of the original entries removed because of missing values. This left us with performance measures for 60 pharmacists across 7 cognitive tasks, 19 of whom had been reprimanded for misprescriptions, and 41 who had not. The data were turned into *z* scores before being used. To prevent the uneven group sizes from allowing base rates to influence how classifications were made, we broke these data into two sets. The first contained the 19 reprimanded pharmacists and 19 randomly selected unreprimanded pharmacists. The second set contained entries for the remaining 22 pharmacists who were unreprimanded. The first set was used as a *training set*; we supplied it to NUANCE for building a classification model. The second set was used as a *validation set*. After NUANCE had created a classification model, we tested the model on the validation set to ensure that it would generalize to unseen data.

Procedure. NUANCE was run with default settings on the training set, with the exception of three parameters: parsimony pressure, minimum constant, and maximum constant. The default settings are outlined in the NUANCE manual, which is available for download from the Psychonomic Society website (www.psychonomic.org/archive), and are discussed in detail in Hollis and Westbury (2006). Here we discuss only the three parameters that we adjusted.

Parsimony pressure is a parameter that addresses one of GP's major limitations: the fact that functions may get so large that they are completely incomprehensible, intractable to run, or a combination of the two. The parsimony pressure parameter imposes a user-specifiable fitness penalty on large solutions, which is a percent value equal to the parsimony pressure times the number of nodes (operators and arguments) in the function (Hollis & Westbury, 2006). The default parsimony pressure is a 0.2% reduction in fitness for every node in a classification tree. For this problem, parsimony pressure was increased to 1.5%. This was done to encourage the development of simple models.

The minimum and maximum constants are parameters that allow the user to constrain any randomly generated variables used in evolved equations to a specified range. The default range of constants that NUANCE allows is 0 to 1, which, through division, can emulate all real numbers greater than zero. For this problem, the minimum and maximum constants were set to -3 and 3, respectively, because this was roughly the range over which our predictors varied (-2.7 to 3).

The operator set is the set of operators allowed within evolved functions. For this problem, the operator set was limited to the "equals" operator and the "less than" operator. This was motivated by the comment that "a linear combination of prefrontal tests was able to correctly classify approximately three quarters of pharmacists into the correct groups, reprimanded and unreprimanded. Such classification was particularly accurate in the case of the unreprimanded group, suggesting that executive or prefrontal function scores above a particular cutoff are very infrequently associated with serious performance error" (Peterson, 2005, p. 20). Because of the infrequency of association between high prefrontal functioning scores and performance error, the implication appears to be that one can derive a very simple and accurate model of performance by classifying pharmacists on the basis of whether they fall above or below a threshold on some combination of prefrontal functioning tests.

Results

The best equation evolved by NUANCE correctly classified 76% of the pharmacists in the training set: 58% accuracy on reprimanded pharmacists ($p = .14$ by exact binomial probability; 2% less than the linear regression and thereby equal within the rounding error due to the smaller n) and 95% accuracy on unreprimanded pharmacists ($p < 0.001$; 9.3% better than the linear regression). It correctly classified 91% of the 22 unreprimanded pharmacists composing the validation set ($p < .001$). The pooled accuracy across all 60 pharmacists was 82% ($p < .001$), which is a 5% improvement in accuracy over the classifier developed by Peterson (2005).

This difference is not statistically reliable ($p = .66$, by Fisher's exact test). Both classification models performed equally. However, the solution found by NUANCE is much simpler than the previous solution: After simplifying the model created by NUANCE by removing tautological and contradictory statements, we were left with a single conditional statement incorporating the result of a single test. The final model is as follows:

If random letter span task score is less than -1.26 z scores,
  group = unreprimanded
else
  group = reprimanded.

The random letter span task requires subjects to input a random sequence of letters for a given letter span: for example, L to O. The participant indicates a letter, using the mouse to cycle through all letters in the given span so that one letter shows up on the screen at a time. A mouse click selects the currently visible letter. When the subject produces an acceptable sequence—a randomized sequence that uses all the letters in the span—a random span one letter longer than the previous one is presented. The task terminates when there are two failures in a row, or if the participant successfully completes two spans of 14 letters (Peterson, 2005).

A person's standardized score on a random letter span task can classify pharmacists as well as a strategy using a linear combination of all seven cognitive performance tasks outlined by Peterson (2005) can. The importance of this task to the problem was suggested by Peterson's analysis: The effect size (Cohen's d) of the random letter span task was .90, the second largest effect size of the seven predictors considered. Pharmacist scores on the random letter span test were reliably correlated ($r = .46$; $p < .001$) with scores for the predictor with the largest Cohen's d (.92), the acquired nonspatial association task.

Like Peterson's linear combination, NUANCE's solution correctly classifies unreprimanded pharmacists much better than reprimanded pharmacists, suggesting that high scores mean that prescription errors are unlikely but low scores do not necessarily mean that prescription errors are likely. Unlike the linear combination, NUANCE's solution is not able to capitalize on unequal base rates favoring unreprimanded pharmacists in the input or on data set specific variance. Its solution was developed on a data set with equal numbers of reprimanded and unreprimanded pharmacists, and that solution was shown to generalize very well to an unseen validation data set. Since the validation data set consisted only of unreprimanded pharmacists' test scores, the strength of the conclusions that we can draw is of course limited. However, even such a weakly cross-validated solution is likely to be more reliable than a linear regression solution that is never cross-validated at all.

Discussion

Although NUANCE was not able to improve on raw accuracy of classification, it did produce a highly simplified classification model, which is a great improvement, practically speaking. Peterson (2005) suggested that practicing pharmacists be screened with a battery of tests measuring prefrontal cognitive ability. The model derived by NUANCE suggests that results from a single test may be sufficient for recognizing pharmacists who are unlikely to make errors. The fact that a very specific type of task can predict the probability of pharmacist error so well also gives us insight into why these pharmacists tend to make errors. The random letter span task taxes working memory, and it requires a modest amount of planning based on the contents of working memory. Dispensing errors on a pharmacist's part may be due to a below average capacity of one or both of these faculties. Intervention for reprimanded pharmacists may want to focus on honing a pharmacist's capacity for these aspects of cognition, or on using external aids (such as written notes) for retaining information rather than relying on working memory.

When NUANCE was introduced, it was framed as a tool for modeling nonlinear variable relationships (Hollis & Westbury, 2006). The results of this study suggest that it is not limited to this single type of task. It is demonstrably suitable for simplifying preexisting models. Users have a great deal of control over how important parsimonious solutions are by manipulating the parameters of NUANCE at runtime. In addition to providing us with a great deal of predictive power, NUANCE can reduce complex models to a level of simplicity that gives the models practical utility.

STUDY 2

Lexical access is a complex process influenced by many factors. To complicate matters, these factors often contribute to the process in complex ways and interact with other factors in equally complex ways. Empirical research on lexical access typically follows an analytic approach, with factorial manipulation as the main tool of choice for understanding how factors contribute to the process of lexical access. This approach has almost single-handedly taken research on lexical access (and psychology in general) to its current standing. But it is not without its shortcomings, which we discuss in more detail in the conclusion to this study. Here we instead take a synthetic approach to studying lexical access, focusing on mathematical modeling rather than factorial manipulation. At most, our aim is to

demonstrate how this approach can reveal useful information that would otherwise be overlooked by an analytic, factorial approach to studying psychological phenomena. At the very least, we hope to demonstrate that a synthetic approach can supplement and inform an analytic approach.

We used NUANCE to model the relationship between 16 variables of potential relevance to the process of lexical access and the behavioral measure of lexical decision reaction time (LDRT). The 16 variables and their abbreviations are listed in Table 1. LDRTs are a common measure of how long it takes subjects to decide whether a presented string is a legal word. We focus in this study on the individual effects of each variable on LDRTs, as well as examine all pairwise interactions among our sixteen predictors. There were three reasons for this. First, it can be very difficult to understand nonlinear interactions of more than two variables. Second, we wanted to allow the possibility of conducting follow-up experiments (not reported here) on any interesting interactions, and it is difficult to design experiments that factorially manipulate more than two variables. Third (as we discuss in more detail below), it can be very difficult to separate effects attributable to the individual predictors from effects attributable to their interactions in nonlinear equations, which is necessary in order to understand how each variable contributed to variance in the dependent measure. Our focus on singletons and pairwise interactions gave us a grand total of 136 "experiments" in this study—a thorough search of the research space that analytic researchers have been exploring experimentally for decades.

We performed this research without entertaining any specific hypotheses. For a discussion of the merits, limitations, and dangers of using GP for such "fishing expeditions," see Westbury, Buchanan, Anderson, Rhemtulla, and Phillips (2003).

Method

Stimuli. One advantage of the synthetic approach that NUANCE enabled us to adopt in this study was that we could study many more stimuli than would be realistically possible in a single experiment. We used behavioral measures taken from the English Lexicon Project (Balota et al., 2002), an online database of over 40,000 words and behavioral data collected on participants' response capacities for the words. We used a total of 4,778 words. For a word to be included, it had to be 4–6 letters long and have an entry in each of the three repositories from which we drew our predictor and dependent variable values, described below.

Predictors. Among the sixteen predictors used in this study were measures of frequency, neighborhood size, average neighborhood frequency, position-controlled bigram/biphone frequencies, and position-uncontrolled bigram/biphone frequencies, on both phonological and orthographic dimensions. Also included were the first and last trigram frequencies for the words. Estimates of these values were calculated directly from the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995). In addition to these 14 predictors, 2 predictors derived from word co-occurrence frequencies were used: the number of semantic neighbors and the average radius of co-occurence (Shaoul & Westbury, 2006). A brief description of each predictor is provided in Table 1.

Procedure. First, each of the predictors was taken alone and used to model LDRTs for half of the 4,778 words—the *training set*. The other half were defined as the *validation*

Table 1
Descriptions of the 16 Predictors Used in Study 2

Variable	Description
LETTERS	Word (letters)
PHONEMES	Word length (phonemes)
OFREQ	Orthographic frequency (per million)
ON	Number of orthographic neighbors
ONFREQ	Average OFREQ of orthographic neighbors
PFREQ	Phonological frequency (per million)
PN	Number of phonological neighbors
PNFREQ	PFREQ of phonological neighbors
CONBG	Summed frequency for any letter pairs in the word in the place they are in for the current word (counted across words of the same length)
UNBG	Summed frequency for any letter pairs in the word (position in word and word length do not matter)
CONBP	Summed frequency for all phoneme pairs occurring together in the place they are in for the current word (only in words with an equal number of phonemes)
UNBP	Summed frequency for any two-pair in the word (position in word and phoneme count do not matter)
FIRSTTRI	Frequency of first three letters of the word as first three letters for all words of same length
LASTTRI	Frequency of last three letters of the word as last three letters for all words of same length
ARC	Average distance between a word and all of its semantic neighbors
NN	Number of semantic neighbors

set. Modeling was performed with NUANCE 3.0. We had three goals: to understand how much variance in LDRTs these predictors accounted for; to discover and test hypotheses about the shape of the relationship between these predictors and LDRTs; and in so doing to demonstrate by example one way in which NUANCE could be used in investigations with many predictor variables.

We were also interested in studying how well the *interaction* between any two predictors accounted for variance in LDRTs, and in understanding the nature of these interactions. To do this, our 16 predictors were taken 2 at a time and used by NUANCE to predict LDRTs as in the first portion of the study.

To maximize the probability of discovering that we had the most predictive functions in both the individual and the pairwise cases, we ran NUANCE on each problem 20 times. The best-fitting equation across all runs was selected for analysis.

Results

The amount of variance in LDRTs accounted for by each individual predictor is displayed in Table 2. All significant interactions are displayed in Table 3. All reported values are from performance on the 2,389-item validation set, to which NUANCE was not exposed while modeling LDRTs. With these data, we will address three questions: "Which predictors account for the most variance?" "Which predictors are the most interactive?" and "What is the shape of the relationships between predictors and LDRTs?"

Which variables account for the most variance? It should be noted that the summed variance accounted for by all of the predictors when run individually (Table 2) exceeds 1. This reflects the fact that there is much overlap among our predictors insofar as how they relate to the process of lexical access. For instance, we should expect phonological and orthographic frequency to relate to LDRTs in roughly the same manner, since they were strongly correlated ($R = .70$ across all 4,778 words; $p < .001$). To understand which predictors account for unique variance, we performed a linear stepwise, backward regression on LDRTs with the NUANCE-derived functions of our 16 predictors as terms in the regression equation. This was appropriate because the relationship between these transformed variables and RTs is indeed as close to linear as NUANCE was able to make them; the fitness function is the linear correlation. The validation set was used to perform the regression. The predictors left in after the backward, stepwise regression are presented in Table 4. The predictors removed during the model simplification included PFREQ, CONBP, PN, PNFREQ, and ONFREQ—mostly phonological variables whose orthographic counterparts remained in the model.

Of the remaining 11 predictors, the 4 that account for the most variance in LDRTs (OFREQ, LETTERS, ON, and LASTTRI) combine to account for 41% of the total variance in LDRTs. This is 96% of the variance accounted for by all 16 predictors together. Frequency, length, orthographic neighborhood size, and body frequency (which is approximated by LASTTRI) are all well-studied variables in lexical access. It did not come as a surprise that orthographic frequency accounts for far more of the variance in LDRTs than any other predictor used in this study. Fre-

Table 2
Variance in LDRTs Accounted for by Each Predictor, Its Log Transformation, and Its Best-Fit NUANCE Transformation

Variable	Untransformed	Log Transformed	NUANCE
OFREQ	0.015***	0.331	0.363††
PFREQ	0.002**	0.121	0.141†
LASTTRI	0.003***	0.072	0.131†††
FIRSTTRI	0.004***	0.092	0.115††
ON	0.078***	0.093	0.093
NN	0.065***	0.096	0.085
ONFREQ	0.000	0.045	0.076†††
PN	0.054***	0.072	0.066
ARC	0.039***	0.047	0.059
LETTERS	0.053***	0.048	0.059
PHONEMES	0.042***	0.039	0.048
PNFREQ	0.001*	0.027	0.048†††
CONBG	0.004***	0.008	0.025†††
UNBP	0.011***	0.011	0.018†
UNBG	0.006***	0.007	0.006
CONBP	0.001*	0.005	0.003

Note—All values are for performance on the validation set. All log and NUANCE-transformed effects significant at $p < .001$. For untransformed variables, $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$. Differences in predictive power between NUANCE-derived fits and best maximum of the other two fits are marked: $^{\dagger}p < .05$; $^{\dagger\dagger}p < .01$; and $^{\dagger\dagger\dagger}p < .001$. For the methodology used to determine significance values for correlational differences, see Blalock (1972).

Table 3
Significant Pairwise Interactions

Variable 1	Variable 2	R^2
LETTERS	PFREQ	.015***
FIRSTTRI	LASTTRI	.010***
ON	PFREQ	.009***
CONBP	UNBP	.008***
LETTERS	OFREQ	.007***
CONBG	UNBG	.006***
ONFREQ	PFREQ	.006***
PHONEMES	ONFREQ	.006***
ONFREQ	UNBP	.005***
ONFREQ	PN	.005***
LETTERS	ONFREQ	.005***
OFREQ	ON	.005***
PFREQ	NN	.004**
ONFREQ	UNBG	.004**
PHONEMES	PFREQ	.004**
ON	ONFREQ	.004**
OFREQ	PN	.003**
PN	PNFREQ	.003*
UNBG	UNBP	.003*
PHONEMES	ON	.003*
LETTERS	CONBP	.002*
PN	UNBP	.002*
PN	UNBG	.002*
ON	NN	.002*
PFREQ	UNBP	.002*
PNFREQ	UNBP	.002*
PFREQ	PNFREQ	.002*

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. $df = 120$.

Table 4
Variables Left in After Stepwise, Backward Regression of the 16 Individual Variables

Variable	R^2
OFREQ	.321***
LETTERS	.059***
ON	.018***
LASTTRI	.008***
FIRSTTRI	.004***
PHONEMES	.004***
ARC	.004***
UNBG	.003***
NN	.002**
CONBG	.001*
UNBP	.001*

*$p < .05$. **$p < .01$. ***$p < .001$.

quency is an important factor in just about every psychological task, including lexical access. What may come as a surprise is how much variance in LDRTs is accounted for by only 4 variables.

Table 2 also enables one to compare the NUANCE-transformed variables, the untransformed variables, and their natural logarithms in terms of their ability to predict RTs. Using methodology described in Blalock (1972), we compared the differences in predictive power statistically to see whether the NUANCE-transformed variables were reliably better at predicting RTs than the raw variables or their logs. Eight of the 16 transformed predictors are reliably better ($p < .05$).

Particularly noteworthy are the variables (such as ONFREQ, PNFREQ, PFREQ, and CONBG) with correlations very close to 0 when untransformed, but much higher when transformed. The importance of these variables could easily be neglected in traditional linear correlational studies. The average of the untransformed correlations of the four variables listed above is .002. The average of their transformed correlations is over 41 times larger, .07. Log transformation of the four variables reduces this difference substantially. However, the average of the NUANCE-transformed variables is still 1.4 times larger than the average of the log-transformed variables (.05). Although these differences of course decrease when the focus is not on the variables with the largest differences, the average correlation across all 16 transformed variables (.08) is still 3.53 larger than the average correlation across all 16 untransformed variables (.02).

Which predictors are most interactive? As stated earlier, we know that language processing is a complex task involving many factors that can interact in complex ways. One cannot understand the mechanics of language processing completely in terms of single causes (Van Orden & Paap, 1997); to understand the mechanics of language processing, one must understand how different pieces of a language processing system interact. Many factorial experiments are designed to look at how two or more variables may interact. NUANCE allows one to search for interactions on a large scale, possibly suggesting variables worthy of closer experimental study.

Deciding which variables are the most interactive is not as straightforward as deciding which variables account for the most variance. The best-fit functions provided by NUANCE may contain effects attributable to the individual predictors, in addition to effects attributable to their interactions. Decomposing each function into its contributing parts can be extremely difficult, because it is not always obvious where the interactions are and where the main effects are in the complex functions provided by NUANCE. We worked around this problem by performing two multiple linear regressions with the output of the functions supplied by NUANCE, for each predictor pair. Since NUANCE tries to predict the dependent measure by linear correlation, these function outputs are guaranteed to be roughly linearly related to that dependent measure, justifying the use of linear regression. The first regression contained terms only for the functions derived when each variable was run alone, as follows:

$$\text{LDRT} = \beta_0 + \beta_1 f_1(a) + \beta_2 f_2(b) + \text{error}.$$

The second regression contained terms for the same functions, plus the function derived when both predictors were used together to predict LDRTs, as follows:

$$\text{LDRT} = \beta_0 + \beta_1 f_1(a) + \beta_2 f_2(b) + \beta_3 f_3(a, b) + \text{error}.$$

By subtracting the variance accounted for by the first regression equation from the variance accounted for by the second regression equation, one can obtain an estimate for the strength of the interaction between any two predictors.

This method is not without its flaws. There is no guarantee that some better fit for each predictor is not embedded within the interaction function of any two predictors—that is, no guarantee that some of the variance that our method attributes to the interaction should not properly be attributed to one or the other of the predictors. Insofar as this is the case, our method will incorrectly attribute too much accounting for variance to the pair's interaction. However, no better option for deducing the strength of any predictor pair's interaction presents itself. Decomposing each pairwise equation by hand is impractical, given how many variable pairs we have and how complex the interactions might be.

After deriving estimates for all interactions using the method above, we can get an estimate of how interactive any single variable is by summing across the R^2 values for all significant interactions in which the predictor is involved with all 15 other predictors in the study. The reliable interaction values are presented in Table 3. The results of summing across all pairwise interactions are presented in Table 5.

The four predictors whose interactions account for the most variance are PFREQ, ONFREQ, UNBP, and LETTERS. Even though ONFREQ was pushed out of the linear stepwise backward regression of the solitary variables, it is the second most interactive variable out of the 16 that we considered. UNBP is the third most interactive variable, but accounts for the third least amount of variance in LDRTs by itself. These findings suggest that there may

Table 5
Results of Summing Across All Pairwise Interactions

Variable	R^2
PFREQ	.030
ONFREQ	.028
LETTERS	.027
UNBP	.013
ON	.013
OFREQ	.011
FIRSTTRI	.010
LASTTRI	.010
CONBP	.008
CONBG	.006
UNBG	.006
PHONEMES	.006
PN	.005
PNFREQ	0
ARC	0
NN	0

Note—All values significant at $p < .05$. $df = 120$.

be some factors in lexical access that make little or no individual contribution to lexical access but are, instead, purely mediating factors.

It is conceivable that ONFREQ appears to be so interactive because of its similarity to ON, which is itself the fifth most interactive variable. The correlation between the best-fit transformations for ON and ONFREQ for predicting LDRTs is very high [$R(2387) = .74$, $p < .0001$]. Further evidence of their relation is provided by the fact that ON remained in the stepwise backward regression and ONFREQ did not. ONFREQ may simply be getting at the same aspects of lexical access as ON does. However, if we look at the significant interactions, we have good reason to suspect this is not the case. ONFREQ has significant interactions with five variables (PFREQ, UNBP, PN, PHONEMES, LETTERS), while ON has interactions with just two variables (PFREQ, OFREQ). There is only one variable with which both ON and ONFREQ interact, PFREQ. The interaction between ONFREQ and ON is marginally reliable at a Bonferroni-adjusted α of .05/120 [$R(2387) = .06, p = .003$]. For these reasons, the two variables do not seem to be getting at the same relationships, and ONFREQ appears to function as a strictly mediating factor with no individual contribution to lexical access.

Another striking result is that interactions with phonological frequency account for approximately three times more variance in LDRTs than do interactions with its orthographic counterpart (Table 3). When the two variables are looked at alone, the ratio flips: phonological frequency accounts for approximately 2.5 times less variance in LDRTs than does orthographic frequency (Table 2). This does not run counter to the general knowledge that frequency mediates almost every other effect in lexical decision tasks (Cutler, 1981), but it does add an extra layer of complexity to this fact.

This summary of the findings emphasizes that many main effects and interactions of potential interest may be overlooked with purely linear methods or with standard transformations such as the logarithm. When we are interested in accounting for as much variance as possible in a dependent measure, we may be on a wild goose chase if we use only linear methods, because some of the variance will be invisible to such methods if the relation between predictors and the dependent measure is not linear. By using NUANCE as we did in the example above, one can select the predictor variables and interactions that are most promising for explaining variance. Such findings might be followed up with more traditional scientific methods such as factorial manipulation of the selected variables, in order to contain convergent evidence of any findings suggested by NUANCE.

What is the shape of the relationships between predictors and LDRTs? We showed earlier that taking the logarithm of most variables increases their correlation with LDRTs. By convention, psycholinguistic researchers take the logarithms of variables that have a large range before considering them as predictors of behavioral measures of lexical access (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2005; Colombo & Burani, 2002; Morrison & Ellis, 2000). This is advisable because such variables have a much larger range than the range of RTs. NUANCE allows us to address a question rarely asked: Is taking a logarithm the best transformation for these variables?

NUANCE's transformations are almost always better than the log transformation (Table 2). Examination of these transformations reveals a general pattern between the relationship of frequency variables and LDRT: The best fit for all frequency measures (excluding uncontrolled bigram/biphone frequency) is not a log transformation, but a reciprocal relationship. A reciprocal function seems more applicable in terms of a simple transformation that maximizes the predictive value of most lexical measures with a large range. Table 2 shows how much of a difference taking the reciprocal of a frequency variable makes (NUANCE transformation) in comparison with logging the measure. In general, the use of NUANCE may allow us to spot general transformations that apply to a class of predictors, and thereby to gain some understanding into how those predictors have an effect.

Another useful piece of information that NUANCE can provide is a principled answer to another question of direct practical importance to experimental psychologists: How large does a variable have to be to be considered high? We know, for example, that word frequency mediates most other variable effects (Cutler, 1981), including the orthographic neighborhood effects seen only in low-frequency words (Andrews, 1989). In the past, this relationship has been characterized with genetic programming (Westbury et al., 2003). When designing factorial experiments to study the effects of orthographic neighborhood, we must use only low-frequency words. But how low is low? Plotting predicted LDRTs by orthographic frequency, we can see how frequency and LDRTs relate to each other, and we can thereby get a principled estimate across a large word set of how low "low frequency" is. Figure 1 suggests that

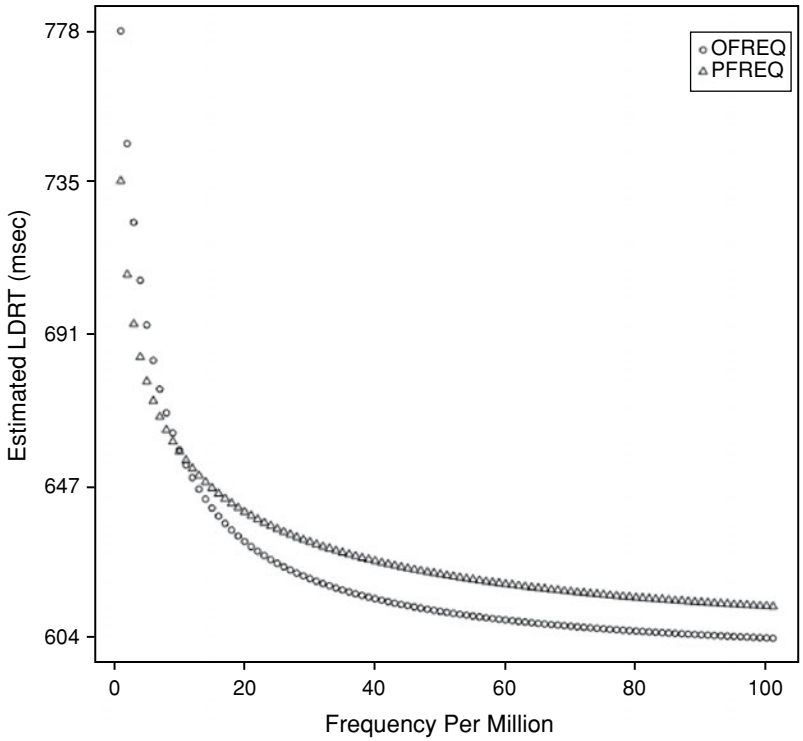

Figure 1. Estimated lexical decision reaction times as a function of orthographic and phonological frequency.

there will be a very small effect (about 20 msec across the entire range of frequency) for words with an orthographic or phonological frequency above 20 occurrences per million. Figure 2, which shows the equivalent curves for orthographic and phonological *N*, suggests that the effect is high at about 8 orthographic neighbors and at about 13 phonological neighbors.

As we have mentioned, our results suggest that phonological frequency may be more important with respect to mediating effects. However, the shape of the relationship between phonological frequency and reaction times was very similar to that between orthographic frequency and reaction times (Figure 2).

Discussion

A few points from the analysis above bear further discussion. Our data seem to suggest that interactions involving phonological frequency account for more variance than do interactions involving orthographic frequency (Table 4). This may be an artifact of the different corpora used to derive these two measures. It may also have more important implications for understanding frequency effects in lexical access, and it may be worthy of further scrutiny as other phonological frequency values become available.

Our observation that some variables appear to have interactions but account for little or no variance in LDRTs individually (most notably, ONFREQ and UNBP) seems in line with an account of psychological systems as reciprocally causal, as laid out by Van Orden and Paap (1997). However, we also note that almost all of the variance accounted for in LDRTs derives from four main effects.

Fifteen of our 16 predictors enter into significant nonlinear relations with lexical decision reaction times. Furthermore, all of these relations are simple (as far as nonlinear relations are concerned), being monotonically increasing or decreasing functions. On average, our untransformed predictors account for just 35% of the variance in LDRTs that our transformed predictors account for (Table 2). As we have noted above, the remaining 65% of the variance that is accounted for by nonlinear transformation of the predictors will be invisible if only linear methods are used.

Inasmuch as one goal of investigations such as ours is to maximize our ability to predict some dependent measure, our finding that most variables measuring the frequency of some event have an inverse relationship with LDRTs is important. Previous research that has looked at frequency as a continuous variable has employed log frequency (Balota et al., 2005; Colombo & Burani, 2002; Morrison & Ellis, 2000, for example). However, NUANCE's fits suggest that a logarithmic transformation is not the best transformation for frequency measures. At least for orthographic frequency, a reciprocal function of frequency accounts for 4% more variance in LDRTs than does a log function. This is a substantial gain in our ability to predict LDRTs when compared with the amount of unique variance accounted for in LDRTs by most predictors (Table 4), constituting as it does 36% of the total variance accounted for.

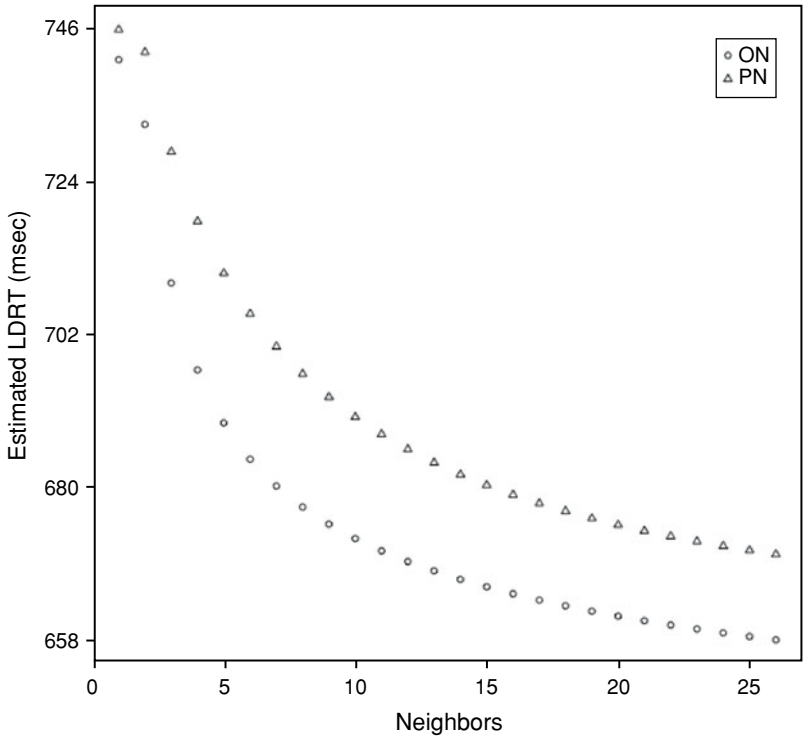

Figure 2. Estimated lexical decision reaction times as a function of orthographic and phonological neighborhood size.

Why a reciprocal transformation of frequency measures is a better fit for LDRTs than a logarithmic transformation may be explained by physiological constraints. We can respond only so fast to a stimulus. At some point, it becomes a physiological impossibility for us to respond any faster. This real-world constraint is captured by the asymptotic nature of a reciprocal function, but not by the continually increasing nature of a logarithmic function. An appeal to physiological constraints would seem to suggest that any predictor with a large range should have a reciprocal-like relationship with measures of human performance. Generally this was true in our study. All but two of our frequency-related variables have reciprocal relationships with LDRTs. It is curious, then, that the relationship between our measures of uncontrolled bigram and biphone frequencies and LDRTs is not best described by a reciprocal function, yet both account for unique variance in LDRTs (Table 5).

A synthetic approach to studying effects such as that which we have used here has many advantages over the more common method of studying effects by using factorial manipulation, which has flaws, especially in the study of lexical access. Balota et al. (2005) have provided five reasons why factorial manipulation is a limited technique. Briefly, the points are the following:

1. It is difficult to find stimuli that vary only along one categorical dimension.

2. Researchers may have implicit knowledge that biases item selection.

3. Stimulus sets often contain words from opposite ends of a dimension of interest, which may change a participant's sensitivity to the factors of interest.

4. Most variables that we study are continuous, and treating them as categorical in factorial manipulations decreases reliability and statistical power.

5. We run into problems concerning whether significant effects are a reflection of lexical processing in general, or an artifact of the selected stimuli. In some cases, it may be hard to differentiate because of Point 1.

There is a sixth reason why studying effects factorially should be expected to gloss over critical information. Analytical tools such as ANOVA treat independent variables as if their underlying relationship with dependent variables is *linear*. This is a gross oversimplification. For example, Baayen (2005) examined the relationship between LDRTs and 13 predictors. Eleven predictors had significant relationships with LDRTs. Of the 11, 6 had nonlinear relationships with LDRTs. Furthermore, 4 of these relationships were nonmonotonic. Nonlinearity is potentially interesting information that is glossed over by factorial manipulation, and nonmonotonicity is completely missed.

Nonlinear relations between stimulus and action (in our case, word properties and LDRT) are a fundamental requirement for behavior that is sufficiently complex to be worth psychological scrutiny. Consider the history of the artificial neural network. Until the late 1960s, a specific class of neural networks received much interest from

psychologists: perceptrons. Perceptrons have two input nodes chained to a third node, an output node. Banks of perceptrons can do tasks such as like pattern recognition and classification. However, Marvin Minsky and Seymour Papert (1968) proved that traditional perceptrons were unable to solve a certain class of problems: linearly nonseparable problems. This proof rendered perceptrons uninteresting in the context of complex psychological behavior. Since Minsky and Papert's proof it has been realized that although perceptrons are of limited interest to psychology, neural networks in general are powerful enough to offer insights to psychology. Whole perceptrons can be chained together to provide more complex behavior. However, this is contingent on the nodes in each perceptron's having *nonlinear* activation functions. Chains of perceptrons with nodes employing only linear activation functions can be reduced to a single bank of perceptrons (Dawson, 2004, pp. 170–173) and are thus uninteresting by Minsky and Papert's proof.

The lesson to be drawn from the history of neural networks is that computational power does not necessarily increase with structural complexity in systems that only perform linear transformations on their inputs. If a system is to be psychologically interesting—if it is to be more than merely the sum of its environment—the system *requires* nonlinear dynamics. As such, psychologists need to pay attention to nonlinearity to get a complete grasp on psychologically interesting behavior. Furthermore, the specific *shapes* of nonlinear relationships are equally important. Minsky and Papert's (1968) demonstration that perceptrons are unable to solve linearly nonseparable problems is not true when nonmonotonic activation functions (such as a Gaussian activation function) are used (Dawson & Schopflocher, 1992).

Factorial manipulation does not adequately capture these formal constraints on complex systems. The analytic approach, which is often coupled with factorial manipulation in psychological research, is not without its own shortcomings. This approach—and Popper's hypothetico-deductive approach to science more generally—is theory driven (Popper, 1959). Research is conducted either to compare the merits of one or more theories or because a theory makes an unexpected prediction and we are interested in verifying it. The Popperian approach to science is not without its detractors (Feyerabend, 1975; Neisser, 1997). One problem with adopting a strict hypothetico-deductive approach to science is that many topics of psychological scrutiny are complex, with many interacting forces directing how they work. This makes building a complete theory of a psychological topic through a strictly analytical approach difficult. We simply do not have the disposition for thinking in terms of complex, nonlinear interactions. Eventually, we will have to incorporate new methods of analysis into our research programs.

Van Orden and Paap (1997) give an account of human behavior that—if true—is even more worrisome for investigators who rely on analytic, factorial methods to study lexical access. Their argument suggests that reductive (analytic) approaches to psychology will eventually need to be replaced because human behavior has *reciprocal causality*: "Reciprocal causality implies that each and every component of a system contributes to every behavior of the whole system" (Van Orden & Paap, 1997, p. 92). When a system is reciprocally causal, the functioning of its components is context dependent, and those components are highly interactive. Reciprocal causality calls into question the applicability of an analytic, reductive approach to studying human behavior. Context dependence implies that a static explanation of the system in question (what a reductive approach aims to provide) will miss critical details. An analytic approach assumes that the system under question can be broken down into basic components that constitute the core of what functionally matters. This is at odds with what we would expect in a highly interactive system. In an interactive system, we would expect that individual components mean very little in comparison with the coordination of those components. Isolating a component may not yield any useful information, since it will ultimately be how that component is related to every other component that matters.

CONCLUSION

We have presented two case studies intended to illustrate that NUANCE is helpful for making sense of real problems in psychology. These studies elucidate two ways in which NUANCE can aid research in psychology. First, it can help simplify complex models by pruning factors that do not matter. Second, it can discover new relationships that were not previously thought to exist. These two abilities can aid in theory development as well as theory simplification, and can both supplement and inspire more traditional experimental investigations. They can also be of utility in applied situations where human behavior is a critical factor. The importance of such tools is accentuated by our earlier assertion that nonlinearity is of fundamental importance to psychological behavior and by our inability to easily reason in terms of complex, nonlinear relationships.

We hope that these results will encourage researchers to employ the use of NUANCE in their own work. Automating the discovery of new knowledge in the manner that we have described here has very little overhead in terms of resources, and it may bring to light information that would otherwise be overlooked by a traditional, analytic approach to psychology.

REFERENCES

Andrews, S. (1989). Frequency and neighorhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 802-814.

Baayen, R. H. (2005). Data mining at the intersection of psychology and linguistics. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 69-83). Hillsdale, NJ: Erlbaum.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (Release 2) (CD-ROM). Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Balota, D., Cortese, M., Hutchison, K., Neely, J., Nelson, D., Simpson, G., et al. (2002). *The English lexicon project: A Web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords*. Available at elexicon.wustl.edu.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D., & Yap, M. (2005). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, **133**, 283-316.

Blalock, H. (1972). *Social statistics*. New York: McGraw-Hill.

Colombo, L., & Burani, C. (2002). The influence of age of acquisition, root frequency, and context availability in processing nouns and verbs. *Brain & Language*, **81**, 398-411.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, **10**, 65-70.

Dawson, M. (2004). *Minds and machines*. Oxford: Blackwell.

Dawson, M., & Schopflocher, D. (1992). Modifying the generalized delta rule to train networks of nonmonotonic processors for pattern classification. *Connection Science*, **4**, 19-31.

Feyerabend, P. (1975). *Against method*. London: New Left Books.

Holland, J. (1992). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Cambridge, MA: MIT Press.

Hollis, G., & Westbury, C. (2006). NUANCE: Naturalistic University of Alberta Nonlinear Correlation Explorer. *Behavior Research Methods*, **38**, 8-23.

Minsky, M., & Papert, S. (1968). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

Morrison, C. M., & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology*, **91**, 167-180.

Neisser, U. (1997). The future of cognitive science: An ecological analysis. In D. M. Johnson & C. E. Erneling (Eds.), *The future of the cognitive revolution* (pp. 247-260). New York: Oxford University Press.

Peterson, J. (2005). *To err is human; to predict, divine: Neuropsychological–cognitive profiles of error-prone pharmacists*. Unpublished manuscript.

Popper, K. (1959). *The logic of scientific discovery*. New York: Harper & Row.

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurence models: A new approach. *Behavior Research Methods*, **38**, 190-195.

Van Orden, G., & Paap, K. (1997). Functional neuroimaging fails to discover pieces of mind in the parts of the brain. *Philosophy of Science*, **64**, 85-94.

Westbury, C., Buchanan, L., Anderson, M., Rhemtulla, M., & Phillips, L. (2003). Using genetic programming to discover nonlinear variable interactions. *Behavior Research Methods, Instruments, & Computers*, **35**, 202-216.

ARCHIVED MATERIALS

The following materials associated with this article may be accessed through the Psychonomic Society's Norms, Stimuli, and Data archive, www.psychonomic.org/archive.

To access this file, search the archive for this article using the journal name (*Behavior Research Methods*), the first author's name (Hollis), and the publication year (2006).

File: Hollis-BRM-2006.zip

Description: The compressed archive file contains one file:

NUANCE-3.zip, a new version of a Java program that uses genetic programming (computation by natural selection) to find nonlinear relations between any number of predictors and a dependent value to be predicted.

Corresponding author's e-mail address:chrisw@ualberta.ca

Corresponding author's Web site: www.ualberta.ca/~chrisw

(Manuscript received November 16, 2005;
revision accepted for publication March 21, 2006.)