Using evolutionary programming to design psychometric tests

Michael Sanderson, Mijke Rhemtulla, Leah Phillips, & Chris Westbury, Department of Psychology, University of Alberta

Abstract:: Two difficult problems facing designers of psychometric tests are item selection and item weighting. A test designer must select items for inclusion in the psychometric instrument from a larger pool of items. After the items are selected, the designer must decide whether any subsets of items should be differentially weighted. We have used evolutionary programming techniques to automate these two steps. We evolved predictor equations across item sets for two psychometric instruments independently developed as an assignment in Psych 431: Psychometrics. One had already been subject to traditional item analysis. Values of the evolved non-linear equations were more highly correlated with the validation items than the original item set, as good or better at predicting validation scores on unseen tests, and used only a small subset of weighted items. We conclude that automated selection of questions holds promise to be a useful tool for addressing these difficult psychometric problems.

How does GP work?

Any mathematical equation can be expressed as a tree. For example, consider the tree at the top left in Figure 1. It expresses the equation: w + ((y * z) / x). The one beside it expresses the equation: (a / b) * log(w). We can mate any two equations by randomly swapping subtrees that compose them to produce children: equations that have the same elements as their parents. The two trees at the bottom are children of the two at the top. GP ensures that only the best parents are allowed to mate: in this case, the ones that best predict the validity scores. This selectivity ensures that produced children will contain elements that may be useful for the problem at hand. Across many generations of selective breeding, average and best fitness increase. Since fitness is determined here by utility for solving the problem, increases in fitness = better solutions to the problem of interest. The process is formally identical to selective breeding in biology, where the breeder decides who is good enough to be allowed to breed. Following repeated breeding sessions, we select the best solution that has evolved.

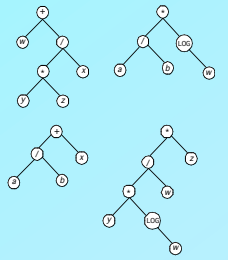

Figure 1: Some equations as trees.

Method: Genetic programming (GP) is a means of programming computers automatically using natural selection. In this context, it may be thought of as a method of automatic model generation and model testing, which makes no pre-assumptions about linearity, independence, or variable distribution. We used genetic programming to try to predict a validity score on two psychometric tests (described at right). The computer was given the question answers, the validity measure, and wide assortment of mathematical functions. It randomly generated 2500 equations that combined the answers using the functions, and selected a small subset that correlated most highly with the predictors. It then 'bred' these equations by random tree swapping (see 'How does GP work?', above right). By repeating this process over 75 generations, increasingly good predictor equations emerge. We used a 'trick' called 'averaged multi-test fitness' to maximize the probability that any evolved equation would be good at predicting data from datasets other than the one used to evolve the equation.

The evolved equations are compared to estimator equations derived using multiple regression.

Both tests were well-designed (according to the 431 prof!) to satisfy the basic requirements for a psychometric instrument. They had clear questions with good face value, they included validation scores, and they had a consistent method of scoring which showed some variability within the population and which allowed summing of question scores as a predictor of the validation scores. They were administered to many subjects. The validation scores are treated as the 'true' value of the construct. We split the dataset into two subsets. The larger set was *the development set*, used to derive the regression equation and evolve the predictor equation. A smaller subset- *the test set*- was set aside so that we could test the ability of the two equations to predict validation scale scores.

Test #1 Results: The first test was designed to measure the construct of 'geekiness': the extent to which a person is a geek. This test was validated against a self-rating on a Likert scale. The test consisted of 76 questions. The validation set contained 59 subjects. The test set contained 30 subjects.

The correlations of the three methods of interest with the validation scores of the development and test sets are shown the the following table:

	Development Set	Test Set
Summed score	0.54	0.59
Multiple regression	0.70	0.20
GP	0.89	0.56

The estimate produced by GP is about as good at predicting scores on unseen tests as using the summed score. However, the GP equation used a non-linear combination of responses to only 12 of the 76 test questions in its prediction.

Test #2 Results: The second test was designed to measure the construct of 'test anxiety'. This test was validated against a published anxiety-rating instrument. The test consisted of 17 questions, following item analysis. The validation set contained 57 subjects. The test set contained 25 subjects.

The correlations of the three methods of interest with the validation scores of the development and test sets are shown the the following table:

	Development Set	Test Set
Summed score	0.77	0.54
Multiple regression	0.85	0.47
GP	0.93	0.49

The estimate produced by GP is almost as good at predicting scores on unseen tests as using the summed score. The GP equation used a non-linear combination of responses to only 9 of the 17 test questions in its prediction.

Conclusions: Our purpose in undertaking this work was exploratory. We wished to examine the possibility of using GP as a method of selecting and weighting questions on psychometric tests. We take the results to be encouraging. The equations evolved using GP are extraordinarily good at 'summarizing' the dataset on which they were evolved, by using a small subset of the questions in that dataset to predict the validation scores with a high degree of accuracy. They are about as good at predicting validation scores on unseen datasets as using the tests in the manner for which they were designed, by summing scores. However, they use many fewer questions than the entire test to achieve that level of predictive accuracy.

The tests used here were 'toy' datasets. However, by combining the power of GP to evolve non-linear predictor equations with much larger preliminary question sets, it seems plausible that we might be able to design more accurate psychometric tests than we could without using GP. Moreover, a careful analysis of the evolved predictor equation (not carried out here) may provide insights into weighting of questions and into relations between questions, and thereby into the formal structure of the constructs we wish to measure.