SOFTWARE FOR MEASURING THE INTELLECTUAL DEVELOPMENT OF STUDENTS: ADVANTAGES AND LIMITATIONS

Michael J. Pavelich, Ronald L. Miller and Barbara M. Olds

Colorado School of Mines

Abstract

Most methods currently available to measure intellectual development in college students are either marginally reliable or are expensive and time consuming. In an attempt to circumvent these difficulties, we have developed Cogito©, a software package which uses a neural network to find patterns in "noisy" paper-and-pencil data and relate them to the Perry or Reflective Judgment models of intellectual development. The project was supported by a grant from FIPSE. We will report the results of testing this software on 88 students and faculty from two colleges and students from a high school. Data from standard Reflective Judgment interviews and from Cogito© have been analyzed in a variety of ways using neural-net software. The better fits show correlation coefficients between Cogito© and interview ratings of 0.5-0.8. Most other fits show correlations below 0.4. These results are slightly to significantly better than previous paper-and-pencil instruments for measuring intellectual development. We will discuss what our results mean for effective assessment. Are R values in the 0.5-0.6 range good enough? Why is there an apparent ceiling on R values for paper-and-pencil instruments?

Keywords — intellectual development, assessment, neural network, Cogito©

Introduction and Objective

Most engineering programs expect that their students will develop intellectually in addition to acquiring knowledge and skills in a specific engineering discipline. However, nearly all measures of student achievement are focused on content knowledge, process ability (e.g. design), or communication skills; students are assumed to be developing intellectually, especially in their ability to think critically, but rarely are meaningful data collected and reported which support such an assumption. However, the recent movement towards outcomes assessment now requires reliable measures of students' abilities to make reasoned decisions as they solve complex problems. For example, in the U. S. the Accreditation Board for Engineering and Technology (ABET) requires institutions to develop assessment processes which can demonstrate "that the outcomes important to the mission of the institution and the objectives of the program are being measured".[1]

Perhaps the most recognized and valid method to quantify maturation of college students' intellectual abilities relies on developmental process models such as Perry's Model of

Intellectual and Ethical Development [2] and King and Kitchener's Reflective Judgment (RJ) Model [3]. These models measure students' positions along a hierarchical construct of stages representing increasingly more sophisticated ways of understanding knowledge and solving complex, open-ended problems. The standard method for evaluating students' stage of development is the structured, hour-long interview, conducted by an expert. The interview is transcribed and then studied and rated by a second expert. Thus gathering data is time consuming, requires experts in the models, and is costly.

In attempts to circumvent these disadvantages, several researchers over the years have developed paper-and-pencil (P&P) instruments to assess a person's position on the RJ or Perry models. These instruments have been disappointing because of the low correlation between results and standard interview results. The objective of the current research was to determine the shortcomings of these previous P&P instruments and use current computer technology to overcome them. We hoped to develop an inexpensive, easy to use assessment instrument that had a correlation coefficient to interview results of 0.8 or better.

This paper briefly reviews the RJ and Perry Models, as well as previous P&P instruments. It then describes how we used computer technology, especially neural-net fitting, to improve on previous P&P efforts. Our results, although better than previous P&P instruments, did not meet our expectations. We conclude by discussing why there is an apparent ceiling to the correlation of P&P instruments to interviews and where P&P instruments might be used in college level assessment.

The RJ and Perry Models of Intellectual Development

The Reflective Judgment [3] and Perry [2] Models of Intellectual Development describe an important aspect of the intellectual maturation we would wish all students to go through before entering their professions. Thus the models speak directly to a universal goal in higher education: to improve students' higher level thinking abilities. These models also allow assessment of that goal.

The models, portions of which are summarized in Table 1, describe the stages people pass through as they mature in their understanding of the nature of knowledge, use of evidence, and open-ended problem solving. For example, people at RJ/Perry stage 2 believe that all questions have single right answers and, thus, no problem is truly "open-ended." Students with some hint of this dualism still in their thinking often view professors who admit to not knowing an answer as incompetent. People at stage (or position or level) 4 understand that there are legitimate unknowns and uncertainties, even in science and engineering, and they do use evidence reasonably well. However, they feel that there are no legitimate ways to weigh alternative possibilities, and, thus, all solutions are equally valid and "everyone has a right to their own opinion." Therefore, students at position 4 see no reason to explore alternatives before reaching a decision, because one well-argued possibility is sufficient. At stage 6, the individual understands the need to use evidence and explore alternatives when solving an open-ended problem, the need for judgments based on personal and articulated standards, and the need to be open to changing circumstances.

William G. Perry, Jr. developed his model from clinical studies of Harvard students in the 1960's. As he interviewed a group of students at the end of each academic year, probing their views of their university experiences, he observed patterns of thinking that were hierarchical and chronological. These patterns Perry translated into his nine-stage model of development that he validated by a second, more extensive, longitudinal study.

The Reflective Judgment (RJ) model was developed beginning in the late 1970's by Patricia M. King and Karen S. Kitchener from their graduate research on student intellectual development. They also used probing interviews of students as their primary data source and were able to identify hierarchical patterns of thought within those data. King and Kitchener have each spent the decades since refining their model, gathering extensive reliability and validation data on it and teaching it to others. The RJ Model has seven stages.

Table 1: *The RJ and Perry Models -- descriptions*

Stage 2:

- **dualist; things seen as right or wrong**
- **authority has all the answers**
- **use of evidence is not understood**
- **ambiguity is a shortcoming or game to get the answer**

Stage 4:

- **ambiguity legitimate, but vexing**
- **uses evidence, but without trust**
- **sees no need to consider alternatives**
- **"all opinions are equally valid"**

Stage 6:

- **ambiguity common to most questions**
- **evidence used to explore alternatives**
- **finds better or best answer in context**
- **commitment using own, considered value system**

The two models are essentially identical for the purposes of the current research. They agree through position 4 and differ only slightly at positions 5, 6, 7. In assessing this higher, more complex thinking, the Perry Model searches for commitment to action based on articulated values, while the RJ Model searches for integration of reasoning between disparate domains of thought. The RJ Model has the advantage of a more substantial research history and more precisely articulated and documented interview/rating protocols; thus, it was chosen as the primary model around which the present research was conducted. However, we made substantial use of Perry Model expert through William S. Moore's work as both a consultant and a rater of interview transcripts.

Data from both models show that undergraduate college students start just above level 3 as freshmen, but have progressed, on average, only about 1/3 of a position by the time they graduate as seniors.[3] Other research shows that how we teach may make a noticeable improvement in students' progress up the levels.[4] Further, there are some data indicating that progress above level 4 is affected by education and is not simply a result of aging.[3] Clearly these models describe a development in students we should work for in higher education, and their interview databases give us a method of assessing that development.

Previous Measurement Methods

The structured, hour-long interview [3,4] is the universally accepted method for determining a student's position on the RJ or Perry developmental scales. These are conducted by trained interviewers, transcribed and then evaluated by trained raters. Interviews are the most likely method to reveal a person's thinking patterns because the interviewer is listening to the person discuss open-ended questions, is drawing that person out and requesting specific clarifications and elaborations. Structured interviews were the measurement instrument used by Perry and by King and Kitchener to establish and refine their models.

The downside of the interview method is that it is costly and requires highly trained experts; each data point consumes 6-8 person hours and costs about $150. In attempts to overcome these disadvantages, experts, particularly using the Perry Model, have developed a variety of paper-and-pencil (P&P) methods to measure intellectual development. These are summarized in Table 2. The Measure of Intellectual Development (MID) and the Measure of Epistemological Reflection (MER) present a series of open-ended questions that the respondent answers in essay form. Topics include decision-making, preferred classroom environment, role of the learner. The essays are then read and rated by trained raters using established criteria. The Parker Cognitive Development Inventory (PCDI), the Learning Environment Preferences (LEP) and the Reflective Thinking Appraisal (RTA) replace the essay with a series of statements that the respondent shows his/her relative agreement with using a Lickert scale. (See Table 4 for the Lickert format in Cogito.) The PCDI, LEP and RTA differ in the issues used and the formats around which the Likert scaled statements are presented.

Although these P&P methods have been thoughtfully and inventively constructed, none has achieved wide acceptance in the educational community. A search of the literature showed seven studies of intellectual development in college students over the last decade. Four of these used interviews [4, 12, 13, 14] while three used P&P methods [15, 16, 17]. Also, all the data compiled in the 1980's using the RJ Model (see reference 3) were obtained using interviews.

Probably the biggest reason P&P instruments are not widely used is their frustratingly low correlations to traditional interview results. The few for which such data exist are indicated in the last two columns of Table 2. The correlation coefficients of 0.4 and 0.3 correspond to covariances between the two measures of only about 16 and 9%, respectively. The very high R = 0.93 reported for the MER is an exciting results on its face value; however, the high result may just be an artifact of the experimental design. The concern is with the interview used. As described in the paper [6], it was not the traditional Perry interview, but one that modeled the MER itself very closely. The traditional Perry interview was considered too free form to ensure

discussion of the MER topics. Why that consistency of topic between paper-and-pencil and interview was important was not explained. Furthermore, the interviewers seemed restricted in the follow-up questions and tangents that could be used. As described, the interview seems to have been almost a verbal rendition of the paper and pencil instrument. Thus, it is not surprising that the raters saw strong similarities as they interpreted the two.

The purpose of the current research was to determine the apparent drawbacks of previous P&P methods and to use up-to-date computer technology to overcome these. The goal was to develop and validate an intellectual development assessment instrument that would give correlation coefficients greater than 0.8 versus the traditional interview measure.

Table 2: *Previous Paper-and-Pencil Instruments*

Name	Year	Authors	Format	Ref	Validation Method	R vs. interview	Ref
MID	1975	Knefelkamp and Widick	essay	5	rater reliability, scores vs. grade level.	0.41	10
MER	1987	Baxter-Magolda	essay	6	rater reliability, scores vs. grade level.	0.93 ??	6
PCDI	1984	Parker	Lickert	7	experts judge statements, internal consistency, scores vs. grade level	NA	
LEP	1989	Moore	Lickert	8	experts judge statements, internal consistency, scores vs. grade level	0.32	11
RTA	1994	Kitchener, et. al	Lickert	9	experts judge statements, internal consistency, scores vs. grade level	NA	

P&P methods will suffer in comparison to interviews because there is no interaction with a listening, probing expert. Thus any P&P method should produce very "noisy" data at best, since we must infer the reasoning process of the subject from the unexplored, first response data given by that subject. That the data are "noisy" is shown in the low correlation coefficients between P&P results and interview results given in Table 2. When dealing with necessarily noisy data, one can traditionally improve the results in two ways: gather more diverse kinds of data and use more sophisticated pattern finding methods. We have attempted to do both of in the current research, as elaborated below.

Cogito© - Its Advantages

Cogito is an interactive computer program we created that uses web and animation software to carry the subject through four open-ended scenarios, seamlessly following up on their decisions and gathering thirty-three numeric data results in the 30-40 minutes it takes to complete the tasks. In effect, Cogito is four P&P instruments on a computer. Cogito was tested by having 88 individuals, from high school students to college faculty, work through it and recording their data. These individuals also volunteered for traditional RJ Interviews that were transcribed and rated by experts. Neural-net software was then used to determine what patterns existed between various sets of Cogito data and the RJ ratings obtained from interviews. Thus Cogito itself gathers a large amount of diverse data without frustrating the subject, while the neural-net supplies the sophisticated pattern finding. These two uses of computer technology were expected to overcome the shortcomings of previous P&P methods.

We wrote the four scenarios and imbedded subject response fields after studying previous P&P instruments, studying interview transcripts and discussing ideas with Perry and RJ experts. The four scenarios deal with controversial issues (see Table 3) as do the traditional interviews. As with the interviews, we are not interested in the subject's stance on the issue, but rather, the thinking behind that stance. One way to get to that is to allow the subject to pick a stance and then rate the reasoning of several fictional people. As an example, Table 4 shows excerpts from Cogito screens in the Overpopulation scenario given to subjects who see the problem as serious. A companion set of statements is given to those who make the alternative decision. The computer's ability to seamlessly branch as it follows the subject's decisions is a real advantage. If Cogito were on-paper, it would run 100 pages and branching would be a frustrating paper shuffling problem for the subject.

Table 3. Scenario topics included in Cogito software

Topic	Dilemma or Controversy
Overpopulation	Is overpopulation a significant problem in the world?
College education	Describe what a college education should do for a student – educate for life or train for a job?
tax rebates	Who should get tax cuts – rich or poor people?
nitrate contamination	How might nitrates in groundwater be controlled?

Three of the Cogito scenarios are simple problem-opinion-reasoning formats while the Nitrate scenario is a complex, step-wise investigation of a problem. It presents subjects with various amounts of information, asks him/her to select a remedy and then show his/her reasoning concerning that remedy in several formats. Several data collection formats were built into Cogito to test which gives more useful data. These are the Lickert scale approach (used three times), a compare answer format (used twice), a time on task monitor, an amount of information accessed counter and an "advantages/disadvantages" list in the Nitrate problem. Cogito, using current programming advances, allows us to take the subject through this complexity of scenarios without losing focus or causing frustration. "Think aloud" testing and anecdotal feedback from

our subjects strongly indicates that they are highly engaged throughout Cogito and see the controversies as intended. Thus, with Cogito, we are able to collect much more data and more diverse data than any on-paper instrument could hope to.

Table 4: Excerpts from the Overpopulation Scenario

Some people believe that overpopulation is one of the greatest dangers facing humans today. They argue that if population growth rates are not substantially reduced within the next few years, the earth faces widespread starvation, resource depletion, and massive environmental degradation by the year 2020. Other people contend that the problem has been exaggerated. They say that humans are distinguished by their resourcefulness and that we will be able to contend with population growth just as we have dealt with other challenges. Supporters of this view point out that past "Doomsday" predictions have been unfounded and argue that this one is likely to be also.

(Statements given to those who checked that overpopulation is a problem we must deal with)

Listed below are statements made by other people who believe that overpopulation is a problem that we must address now. Please check the spaces on the left to indicate how closely each person's statement reflects your thinking.

Strongly Agree		Neutral		Strongly Disagree	
____	____	____	____	____	1. I don't think we will know the answer until we get to the year 2020. You can't really prove that there is a problem. Either side could be right. I personally feel one way; others are entitled to their opinions.
____	____	____	____	____	2. It is a problem we must deal with. I see population growing in places like India and China. People who ignore the population problem just don't care about the future.
____	____	____	____	____	3. Even though we can't be certain because there is evidence on both sides, we have to take action. You make a commitment and then monitor how the situation plays out and adjust your response accordingly. Better safe than sorry.
____	____	____	____	____	4. You need to keep a balanced view, looking at the claims from both groups. It is a potentially serious problem that experts need to keep taking data about. We need to keep tabs on their studies and see where they lead us.
____	____	____	____	____	5. If the experts on this topic say there is a population problem, I believe them. They are, after all, authorities on the subject.

Given that we are not formally trained in educational psychology, one might ask if scenarios and response fields in Cogito are state-of-the-art for such P&P instruments. We feel that our development methods and comparisons of statements with other P&P instruments testify to Cogito being state-of-the-art. We chose new scenarios so that Cogito and the interviews topics would not be confounded. As mentioned above, "think aloud" studies showed that the subjects saw the sides of each controversy as expected. The Lickert [7, 8, 9] and compare answer [9] formats were much like those used in previous P&P instruments. We studied the models in detail, attended an RJ workshop run by Kitchener and Lynch and became certified as RJ interviewers. We had lengthy discussions with Perry and RJ Model experts. We constructed statements by studying rated interviews and extracting the language and reasoning of people at different levels. Most importantly, comparisons of Cogito statements (see Table 4 for examples) with those used in previous P&P instruments, show a strong similarity in concept and language.

The second advantage of this work is that we used neural-net software to search for patterns in the "noisy" P&P data. Neural nets, a computerized attempt to emulate human thought processes and decision-making, are particularly effective at recognizing and analyzing complex patterns with subtle features.[18, 19] In effect, we do not have to assign an RJ/Perry level to a P&P statement as previous researchers had to do (see references in Table 2). The neural-net software takes an unprejudiced look at the P&P statements and discovers which combinations of statements people at the higher RJ/Perry levels prefer and which people at lower levels prefer.

The key to successful neural net fitting is obtaining a comprehensive and valid data set from a large number of subjects whose RJ/Perry ratings vary widely. We currently have comprehensive data from 88 subjects ranging from high school juniors through college undergraduates to graduate students and faculty. Students and faculty from two colleges were used: an engineering school with traditionally aged students and a city commuter college having older, working students. The high school students were from a social sciences class in a suburban school. Each subject worked through Cogito giving us his/her individual thirty-three answers. Each also did a standard RJ interview of three incidences. Their interviews were transcribed and then rated by an RJ expert. Thwenty four were also rated by a Perry Model expert. Thus we have the Cogito data and the "true" RJ/Perry rating for each subject.

Rater reliability in this study was good. The blind RJ ratings of three incidences per subject showed an average spread of 0.6 of a position. The average difference between an RJ and Perry rating of a subject was also 0.6 of a position. The correlation coefficient between RJ and Perry ratings for 24 subjects was 0.84. These figures match those in reference 3.

In a typical neural-net fitting run, 70 of the 88 data sets were inputted to the program and it sought a pattern between the Cogito data and the RJ/Perry ratings. With this number of subjects we only inputted a maximum of six Cogito data pieces per subject. When the best fit was obtained (5,000-30,000 iterations taking a few minutes on a Pentium PC) the 18 data points held back were inputted without their RJ ratings as a test of the fit. We then compared the computer's estimate of the RJ/Perry ratings of these 18 with what we knew them to be from the interviews. Goodness of fit was measured by the correlation coefficient, R, between the neural-net's estimate of the subjects RJ/Perry rating and that obtained from the interviews.

Cogito - Results

We systematically ran neural-nets fits looking for those Cogito data that best correlated with interview results. We found that six Lickert statements in the Overpopulation and Education scenarios (op2, op4, op5, ed2, ed3, ed5) gave by far the better fits. We also experimented with data input approaches and determined that numeric inputs of RJ ratings data gave far better fits than did symbolic inputs. These conditions gave individual fits whose testing correlation coefficients ranged from 0.2 to 0.8 with an average R = 0.5. Thus our best Cogito results to date are only slightly better than those obtained using previous P&P instruments (see Table 2). Details of the data fitting are given below.

Earlier testing with the first twenty and then the first forty data sets collected showed that the Lickert answer fields (see Table 4) gave the much better fits than the alternative formats we tried (compare-answer, advantages-disadvantages, information accessed and time on task). Thus we concentrated our fitting on the sixteen Lickert statement inputs from each subject.

The variations and results of Lickert statement fitting with the pool of data from 88 subjects are given in Table 5. The first column numbers the fits, the second is our code name for the fit and its resulting neural net algebra. The next four columns give the conditions of the fit. "Statements Fitted" are the Cogito Lickert statements whose data was used. The A, B, C, D sets of statements were different combinations from the three scenarios as identified at the bottom of the Table. These statements were identified by three different people analyzing the raw Lickert statement data for those that seemed to differentiate among subjects best. The "test set" column identifies which group of 18 subjects was held back to test the net. Three different test groups were used. The "RJ data" column tells how the interview rating was inputted, as a numeric value or as a symbol. We converted the ratings to symbols by clustering them in four groups: Low = RJ $\leq$ 4.0, Medium = 4.1 $\leq$ RJ $\leq$ 4.8, High = 4.9 $\leq$ RJ $\leq$ 5.9, Expert = 6.0 $\leq$ RJ $\leq$ 7.0. The "Cogito data" column tells if the data from the statements was inputted as numeric or symbolic. In numeric form 1= strongly agree to 5 = strongly disagree. In symbolic form, we just assigned a different symbol to each possible response to each different statement. The neural-net sees numeric data as being on a continuous number line, but sees symbolic data as several unconnected inputs.

The last four columns give the results of these fits. The column "R train" gives the correlation coefficient for the 70 data sets used to train the net. The "R test" column gives the correlation coefficient for the 18 data sets held back to test the net. To give a visual idea of how the R values and data relate, we have plotted the RJ level from interviews versus computer fitting for fits # 19 and 26 in Figures 1 and 2. When the RJ rating was put in symbolically, the net gave a symbol as a result. Here we documented the results by counting how many subjects were missassigned a symbolic RJ level. These results are given in the last two columns.

The "R train" and "# bad train" outputs are very good as expected; with these the software knows the RJ ratings and is trying to reproduce those values. The "R test" and the "# bad test" values are the true look at the goodness of fits. Here we use the algebra developed in the training to predict the RJ level of the 18 data sets held back. These show what one could expect using Cogito and that fit as an assessment measure for a group of people.

Table 5: Cogito Lickert Statement Fits (representative ones)

#	name	statements fitted	test set	RJ data	Cogito data	R train	R test	# bad train	# bad test
1	OP41	overpop	4	num	num	0.81	-0.10		
2	OP42	overpop	4	num	num	0.79	0.23		
3	OP43	overpop	4	num	num	0.70	0.21		
4	OPsym41	overpop	4	num	symbolic	0.86	-0.19		
7	A41	A	4	num	num	0.95	0.21		
9	A61	A	6	num	num	0.98	0.12		
10	Ax41	A	4	symbolic	num			15 of 70	12 of 18
11	Axy41	A	4	symbolic	symbolic			2 of 70	10 of 18
16	B41	B	4	num	num	0.94	0.60		
17	B42	B	4	num	num	0.94	0.42		
18	B43	B	4	num	num	0.93	0.34		
19	B51	B	5	num	num	0.95	0.78		
20	B52	B	5	num	num	0.95	0.63		
21	B61	B	6	num	num	0.95	0.20		
22	B62	B	6	num	num	0.92	0.38		
23	Bxy41	B	4	symbolic	symbolic			2 of 71	12 of 17
24	Bxy51	B	5	symbolic	symbolic			1 of 70	10 of 18
25	Bxy61	B	6	symbolic	symbolic			0 of 71	12 of 17
26	By41	B	4	num	symbolic	0.96	0.60		
27	By51	B	5	num	symbolic	0.96	0.54		
28	By52	B	5	num	symbolic	0.97	0.49		
29	By53	B	5	num	symbolic	0.96	0.57		
30	By54	B	5	num	symbolic	0.96	0.28		
31	By61	B	6	num	symbolic	0.98	0.60		
32	By611	B	6	cont above	cont above	0.98	0.32		
33	C41	C	4	num	num	0.71	0.30		
34	C51	C	5	num	num	0.67	0.18		
35	C61	C	6	num	num	0.82	0.27		
40	D41	D	4	num	num	0.93	0.53		
41	D411	D	4	num	num	0.93	0.29		
42	D51	D	5	num	num	0.88	0.71		
43	D61	D	6	num	num	0.95	0.04		
46	Dy51	D	5	num	symbolic	0.95	0.31		
	A=op4, ed1, ed3, nit2, nit4, nit5					C=op4, ed3, nit2, nit5			
	B=op2, op4, op5, ed2, ed3, ed5					D=op2, op4, ed1, ed3, ed5			

Fits # 1-4 show that using the 5 Lickert statements from the Overpopulation scenario gives poor results (R ranging from -.2 to +.2). This was also the case with the Lickert statements from the Education and Nitrate Scenarios. We then looked at the raw data from the 88 subjects for all Lickert statements and determined which seemed to discriminate better among the subjects. We came up with 10 of the 16 standing out. These were grouped in four combinations and fitted. Of the groups, B statements gave good fits most consistently. Note that fits with all other statement grouping were poor. D statements look promising, but do not give R values as high as do B statements. Also note that any fits where the RJ data was entered symbolically gave poor fits; having 2/3 of the test subjects miscategorized is not a good result.

B statements do give encouragingly high R test results. However, even here there is room for concern that the pattern found in each fit is not robust. If the patterns were robust, we should get nearly identical results each time the fit is independently run. We sometimes don't see that. Fits # 16-18 are the same fit run at different times with the training data entered in a different order. All other parameters were the same. We get three fairly different R test values. Fits # 27-30 are again fits using the same parameters. Three of the four match, but the fourth R is well off from the others. Robust patterns should also return nearly the same results when the test set of subjects is changed. We do get nearly identical results with test set 4 versus 5, but test set 6 consistently gives much lower R test values. It is very possible that the problem lies in the number of data points available. We may need to double or triple our data base before the neural nets can find robust patterns and return more consistent R values.

In its current state of validation, Cogito could be used to determine the average RJ level of a population. One would have the subjects do Cogito, collect their outputs and then run them through the 14 trained neural-nets we have for the B statements with numeric RJ input. The outputs from these 14 could then be analyzed statistically to determine the best estimate of each subject's RJ level. This approach is described in reference 20. The estimates of individual RJ levels could then be combined to determine the population's average. This approach could be expected to have an R = 0.5 versus what would find from standard interview measures. (If we average the R test values for B statements in Table 5, we get R = 0.5.)

We continue working on more complex analyses of the Cogito versus interview data seeing if the data can yield better correlations. A current approach called Group Count Probability Analysis is outlined in the Future Research section below.

Why do P&P measures, including Cogito©, not work better?

Cogito© works better than, or at least as well as, other paper-and-pencil (P&P) instruments in assessing the RJ/Perry intellectual development of a student. We see correlation coefficients of Cogito© data to interview data with B Statements at 0.5 or better with some consistency. These beat or match cited results with other P&P instruments. Furthermore, the current database is much more extensive than in any previous studies lending higher transfer reliability to the results. However, the 0.5 or slightly higher range of R values is still disappointing, as we had anticipated much higher and more consistently higher correlations.

Even with more data, the R test values for Cogito© may not improve; we may simply get less variation from fit to fit. It may very well be that R = 0.5-0.6 is all that even Cogito© can achieve.

This FIPSE Project was built around learning from previous P&P investigations and overcoming their apparent shortcomings. We feel we did this well. The set-ups and questions in Cogito© are state-of-the-art, subjects get through many data gathering exercises with energy and interest, and we used a sophisticated pattern searching procedure. However, our correlation coefficients are still disappointing. The R range of 0.5-0.6 corresponds to 25-36% covariance between Cogito© and interviews. Thus less than 36% of the factors that influence a person's interview, are affecting their Cogito© answers. We had hoped for much better than this, somewhere in the 65% covariance range, or correlation coefficients of 0.8 or higher.

Why has Cogito© not worked better? We feel that it is because we may have reached the limit of what one can do with such P&P instruments. With P&P instruments, we are asking people to convey their very complex and subtle differences in thinking through a series of check marks on a scale (or by writing a short essay). No matter how sophisticated and well worded the statements, no matter how many scales they mark, no matter the sophistication of the pattern searching, the gulf between how they think and what they mark is probably vast. Why one marks a series of statements as he or she does depends not only on RJ level, but also on word interpretations, specific thoughts on the problem and many other variables. This is understood in the interview and is why the certified interviewers must be carefully trained to probe behind the immediate answers. The success of the interview depends on probing specific to that subject's response. In P&P instruments, we have only the subject's immediate answers. This is obviously true with Lickert scale instruments like Cogito, LEP, PCDI and RTA. On reflection, it is also true of essay instruments such as the MID and MER.

Our conclusion is that when we must rely only on the immediate answers provided by the subject, we can only see 25-36% of the factors needed to assess his/her intellectual development accurately. It takes the interchange with a trained interviewer, probing for the meaning, the thinking, behind these immediate answers, to be able to see 90-100% of the factors. Thus P&P instruments perhaps cannot give us better than a 0.5-0.6 correlation with what we are trying to measure.

What is the value of P&P instruments?

Paper-and-pencil instruments, including Cogito©, can be valuable in certain circumstances, but these circumstances are limited. One should judge the value of a P&P instrument by deciding if the low correlation coefficients it has with interviews is acceptable to the measurement purposes. Using Cogito, for example, one can expect results of R = 0.5-0.6. What this means can be seen by viewing the scatter in the plots of P&P measured levels versus interview measured levels. Such plots are shown in Figures 1 and 2. Figure 1 shows data where the "R test" value is 0.78. Note that there is good agreement between P&P and interview for most of the 18 test points, but for several the difference is still greater than 1.0 RJ/Perry level. Also note that neither Cogito nor any other P&P can certify results at this high R test value. The more realistic view is shown in Figure 2 where the R test = 0.54. In particular look at the data points near RJ interview level 4.0. The neural-net fit (RJ computer) gives values for these

subjects ranging from 3.2 to 6.2. That is quite a large range of scatter. However, the P&P results do correlate to interviews; the R = 0.5 comes from the large fraction of points that do correlate well.

If one is looking for relative changes over long times with large populations, P&P instruments may give useful data. In fact they may be the only viable assessment instrument, given costs and time. Looking for relative changes and in large populations makes P&P data distributions more meaningful and over relatively long times may allow the "noisy" measure to see some real change. A prime example of reasonable P&P use is the Alverno College use of the MID as one measure of total curriculum effects.[21]

If the curriculum question involves smaller populations over shorter times, traditional interviews probably should be used. An example would be looking at the effect of some experimental courses over a few semesters. Here the populations are smaller and the changes may be small; thus the "best" measure, the traditional interview where valid data per person is obtained, is the assessment measure of choice. In circumstances where we wanted solid data to convince faculty about curricular changes,[4] we deemed it worth the time and money to use interviews.

Further Research and Reports

We continue working on more complex analyses of the Cogito© data versus interview data to see if better correlations can be found. A current approach called Group Count Probability Analysis looks promising, but needs more investigation before we can publish results. Here we again assign RJ/Perry interview scores to four categories (see symbolic RJ inputs above), but we also convert the Cogito© data to probability group counts per individual. The RJ category versus group count is then fitted by the neural-net. The approach and a preliminary pass at using it are reported in our final report to FIPSE.[22] If repeated testing of the approach using various train/test groups succeeds, we will publish the results.

The data gathered in this FIPSE project can be mined for more insights than computer predictions of RJ/Perry levels. These will be the subject of other reports and papers. The topics include RJ level of seniors at two engineering schools, RJ levels versus education, correlations of interview ratings between RJ and Perry experts, analysis of why Statements B worked best and comparisons from the two different campus cultures.

We also invite other researchers to use our interview data if it fits their needs. Theo Dawson of Berkeley has used of our interview data as part of the validation study of her LAAS assessment system. Interested users can contact one of us or obtain our final FIPSE report.[22]

Figure 1: Results of Neural-Net Fit #19, R test = 0.78

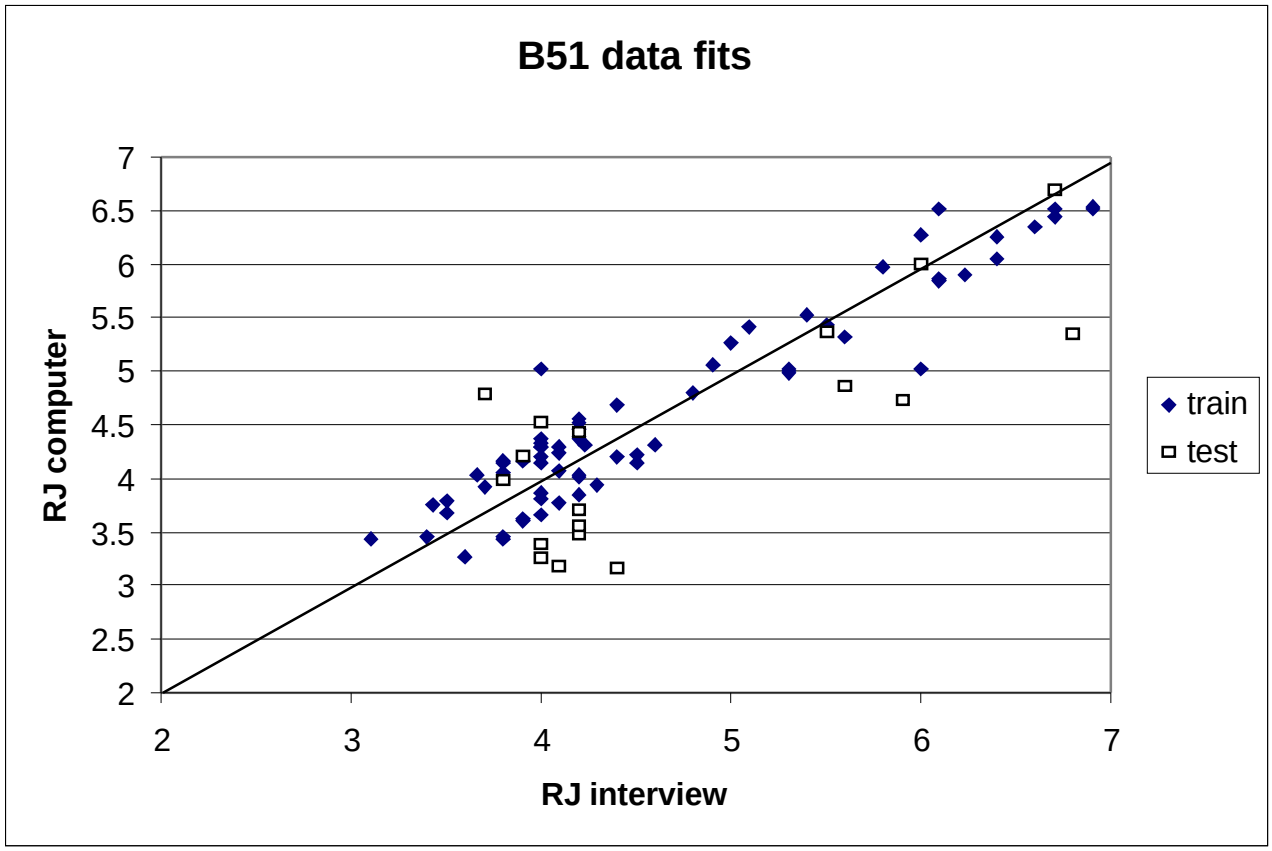
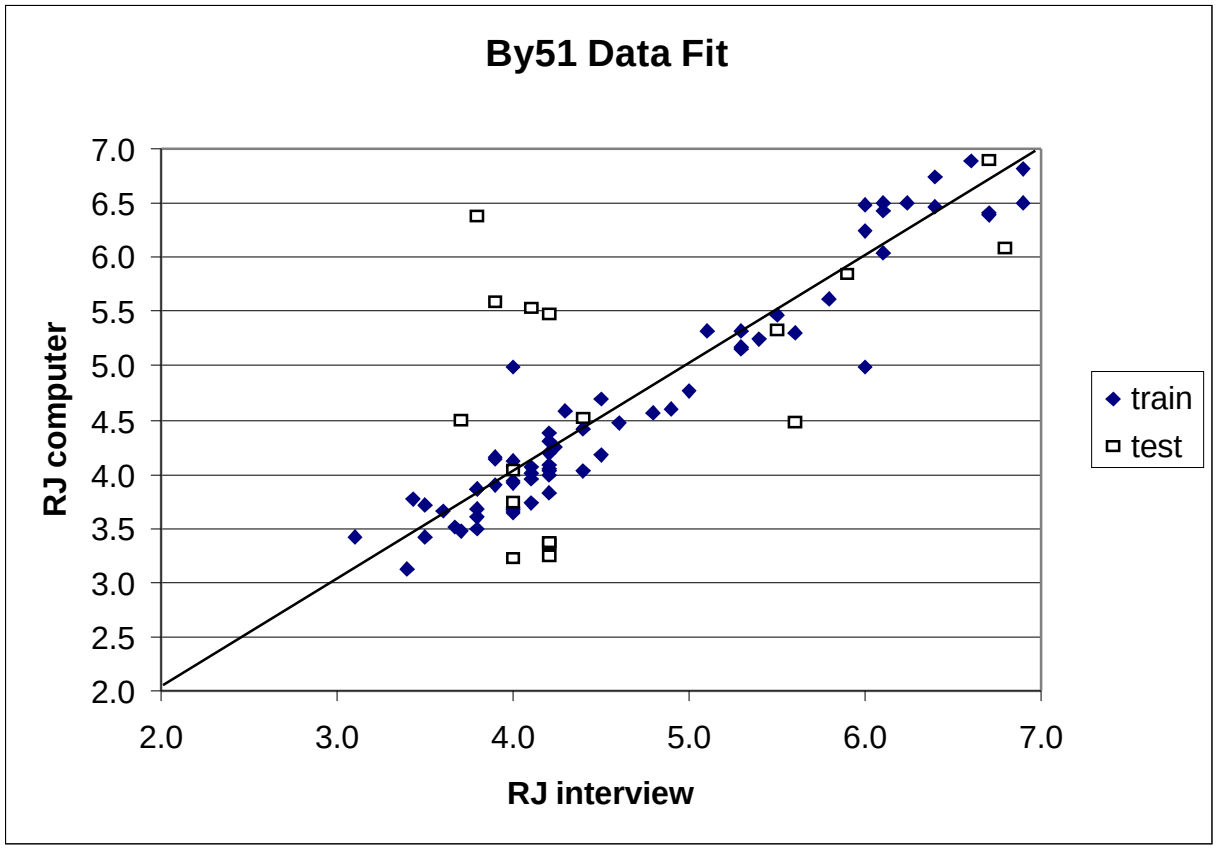


Figure 2: Results of Neural-Net Fit #27, R test = 0.54



Acknowledgment

We wish to acknowledge the Fund for the Improvement of Postsecondary Education of the U.S. Department of Education for financial support of this project. We also wish to thank our consultant William S. Moore for his insights and for evaluating interview transcripts.

In Memory

We also acknowledge the fine work of Cindy Lynch on this project as consultant and principal rater of RJ interviews. We deeply mourn her passing in an auto accident in February, 2002. Cindy will be sorely missed within the circle of folks working with the intellectual development models. She brought a wonderful energy, precision and wellspring of ideas to that group. She will also be missed as a friend. We dedicate this publication to her memory.

References Cited

1. "Criteria for Accrediting Programs in Engineering," Accreditation Board for Engineering and Technology, Baltimore, MA, 1998 (available on ABET WWW homepage: www.abet.org).

2. Perry, W.G., Jr., *Forms of Intellectual and Ethical Development in the College Years*, Holt, Rinehart and Winston, Inc., New York, 1970.

3. King, P.M. and K.S. Kitchener, *Developing Reflective Judgment*, Jossey-Bass Publishers, San Francisco, 1994.

4. Pavelich, M.J., and W.S. Moore, "Measuring the Effect of Experiential Education Using the Perry Model," *Journal of Engineering Education*, vol. 85, pp. 287-292, 1996.

5. Moore, W. S.. "The Measure of Intellectual Development: A Brief Review", 1982, a monograph available through CSID & The Perry Network, 1505 Farwell Ct. NW Olympia, WA 98502, 360-786-5094 (voice) 503-212-8082 (efax), wsmoore51@home.com

6. Baxter-Magolda, M.B., "Comparing Open-Ended Interview and Standardized Measures of Intellectual Development," *Journal of College Student Development*, vol. 28, pp. 443-448, 1987.

7. Parker, J.C. "The Preliminary Investigation Of the Validity and Reliability of the Parker Cognitive Development Inventory" unpublished doctoral thesis, 1984, University of Iowa, Iowa City. Also see Parker, J.C. and Hood, A.B., "The Parker Cognitive Development Inventory" URL: http://web.indstate.edu/dragon/ia-pcdi.html, 1997.

8. Moore, W.S. "The Learning Environment Preferences: Exploring the construct Validity of an Objective Measure of the Perry Scheme of Intellectual Development" Journal of College Student development, vol 30, pgs 504-514, 1989.

9. Kitchener, K.S. 1994. "Assessing Reflective Thinking within Curricular Contexts" Report to Fund for the Improvement of Post-Secondary Education (FIPSE), Washington, D.C., 1994, Contract: P116B00926. ERIC Clearing House # ED 415 751.

10. Pavelich, M.J. and P. Fitch, "Measuring Students' Development Using the Perry Model," *Proceedings of the American Society for Engineering Education Annual Conference*, Washington, DC, pp. 668-672, 1988.

11. Culver, R.S., P. Cox, J. Sharp, and A. Fitzgibbon, "Student Learning Profiles in Two Innovative Honors Degree Engineering Programmes," *International Journal of Technology and Design Education*, vol. 4, pp. 257-287, 1994.

12. Kitchener, K.S., Lynch, C.L., Fischer, K.W. and Wood, P.K., "Developmental range of Reflective Judgment: The Effect of Contextual Support and Practice on Developmental Stage", *Developmental Psychology*, Vol. 29, pp.893-906, 1993.

13. Kronholm, M.M., "The Impact of Developmental Instruction on Reflective Judgment" , *The Review of Higher Education*, Vol. 19, pp.199-225, 1996.

14. Marra, R.M., Palmer, B., Litzinger, T.A., "The Effects of a First-Year Engineering Design Course on Student Intellectual Development as Measured by the Perry Scheme", *Journal of Engineering Education*, Vol. 89, pp. 39-45, 2000.

15. Durham, R.L., Hays, J. and Martinez, R., "Socio-cognitive Development Among Chicano and Anglo College Students", *Journal of College Student Development*, Vol. 35, pp.172-182, 1994.

16. Wilson, B.A., "A Descriptive Study: The Intellectual Development of Business Administration Students", *The Delta Pi Epsilon Journal*, Vol. 38, pp. 209-221, 1996.

17. Zhang, L. and Watkins, D., "Cognitive Development and Student approaches to Learning: An Investigation of Perry's Theory with Chinese and U.S. University Students", *Higher Education*, Vol. 41, pp.239-261, 2001.

18. Mehrotra, K., C.K. Mohan, and S. Ranka, *Elements of Artificial Neural Networks*, MIT Press, Cambridge, Massachusetts, 1997.

19. Eberhart, R., P. Simpson, and R. Dobbins, *Computational Intelligence PC Tools*, Academic Press, Inc., Boston, 1996.

20. Olds, B.M., Miller, R.L. and Pavelich, M.J., "Measuring the Intellectual Development of Students Using Intelligent Assessment Software", *Proceedings of the Frontiers in Education*, Kansas City, MO, November, 2000, (electronic), IEEE, Washington, D.C.

21. Marcia Mentkowski, M, *Learning That Lasts: Integrating Learning, Development and Performance in College and Beyond*, Jossey-Bass Publishers, San Francisco, 2000.

22. Miller. R.L., Olds, B.M. and Pavelich, M.J. " A Computer-Based Expert System to Measure Intellectual Development in College Students", Report to Fund for the Improvement of Post-Secondary Education (FIPSE), Washington, D.C., 2001, Contract : P116B70050.

MICHAEL J. PAVELICH

Michael J. Pavelich, a Professor of Chemistry at the Colorado School of Mines, has been active in engineering education and chemical education circles for the last 30 years. He counts as mayor accomplishments the creation and continued success of the freshmen/sophomore design program at CSM and an inquiry formatted lab program for freshmen chemistry that is used by schools across the country. He has over forty publications in the college education literature and is currently working on applying intellectual development theories to the teaching and assessment of design courses. He has presented numerous workshops on college teaching at campuses across the country and was an ASEE/NSF Visiting Scholar in 2000-2001. He has been recognized with several teaching awards at CSM.

RONALD L. MILLER

Ronald L. Miller is Professor of Chemical Engineering and Petroleum Refining at the Colorado School of Mines where he has taught chemical engineering and interdisciplinary courses and conducted research in educational methods for over seventeen years. He has received three university-wide teaching awards and has held a Jenni teaching fellowship at CSM. His paper entitled "Using Portfolios to Assess a ChE Program" (co-authored with

Barbara Olds) won the Corcoran Award from the chemical engineering division of ASEE for best paper published in Chemical Engineering Education during 1999. His paper entitled "Connections: A Longitudinal Study of an Integrated Freshman Program" (co-authored with Barbara Olds) won the award for best paper in the Educational and Research Methods Division of ASEE during the 2001 annual conference.

BARBARA M. OLDS
Barbara M. Olds is Associate Vice President for Academic Affairs and Professor of Liberal Arts and International Studies at the Colorado School of Mines where she has taught for the past eighteen years. She has participated in a number of curriculum innovation projects and has been active in the engineering education and assessment communities. Dr. Olds has received the Brown Innovative Teaching Grant and Amoco Outstanding Teaching Award at CSM and was the CSM Faculty Senate Distinguished Lecturer for 1993-94. She was a Fulbright lecturer/researcher in Sweden in 1999. She has received grant awards for educational research from the National Science Foundation, the U.S. Department of Education (FIPSE), the National Endowment for the Humanities, and the Colorado Commission on Higher Education.