Dynamic Testing in Selection for an Educational Programme: Assessing South African Performance on the Raven Progressive Matrices

Hermien Zaaiman, Henk van der Flier* and Gerard D. Thijs

This article describes the research results of an investigation into the use of dynamic testing for the selection of candidates for an educational programme. The selection of students from educationally disadvantaged backgrounds for mathematics-, science-, and technology-based programmes is a problem for which most South African higher education institutions still have to find adequate solutions. Dynamic testing procedures are often seen as more fair to use for selection than single-session tests in situations of unequal educational opportunity. The possibility that a selection instrument using a dynamic testing process could add significantly to the selection effectiveness already achieved in a South African mathematics and science foundation year using single-session tests was investigated. The performance of a group of educationally disadvantaged black South African students on the Raven Progressive Matrices is compared to that of other groups reported in research literature. Considering the disadvantaged nature of the South African group, the group performed well on the Raven test when compared with data from other countries.

In South Africa, the educational backgrounds of applicants for higher education vary widely in terms of quality of primary and secondary school education received. The majority of South African school graduates are not adequately prepared when leaving school to succeed in higher education. The mainly race-based differences in educational opportunities complicate the development of fair and effective selection mechanisms. An important issue is the identification of students with the potential to succeed in higher education study despite previous educational disadvantage. This is a problem for which most South African higher education institutions still have to find adequate solutions. The question addressed in this paper is whether dynamic testing procedures can be expected to contribute in this respect.

Of course the problem of finding ways of reducing adverse impact against members of classes that have traditionally been unfairly discriminated against is not unique for South Africa. Selectors in the USA and Europe are confronted with similar issues (see, for example, Wood *et al*. 1997). The issue has both legal and political importance where universities could be accused of discrimination by virtue of their selection procedures.

*Address for correspondence: Henk van der Flier, Department of Industrial and Organizational Psychology, Vrije Universiteit, v.d. Boechorststraat 1, 1081 BT Amsterdam, The Netherlands

In South Africa, as elsewhere, black students are underrepresented in natural science-, engineering- and technology-based programmes (DACST 1996). Inadequate numbers of South African black students from disadvantaged backgrounds achieve the school-leaving (Grade 12) results required for direct enrollment into science-based programmes. Students with the potential to succeed despite weak school results therefore have to be identified and selected for suitable higher education programmes.

The high failure rates and low numbers of students in the science-based faculties at the University of the North in South Africa led to the UNIFY programme. The University of the North is one of South Africa's historically disadvantaged institutions. It is situated in the mainly rural Northern Province and was established during the apartheid era to cater to the higher education needs of black South African school graduates. In the post-apartheid era, the majority of students who apply to study at this university still come from educationally disadvantaged backgrounds and are not adequately prepared to succeed in higher education study programmes. UNIFY is a mathematics and science foundation year which aims to give disadvantaged South African students a chance to enter and succeed in the

University of the North's science-based faculties. It is a one-year programme that uses a student-centred approach to teaching in which staff act as facilitators. Emphasis is placed on problem-solving skills rather than content knowledge in the subjects biology, chemistry, English and study skills, mathematics and physics.

Each year 150 students have to be selected from between 450 and 700 applicants. These students have to be identified from an applicant pool consisting mainly of applicants with weak school-leaving results. A comprehensive research project on the selection of students to UNIFY started in 1994. This project was done at the University of the North in cooperation with Vrije Universiteit Amsterdam in the Netherlands. Zaaiman (1998) gives the full research report.

The UNIFY selection mechanism consisted of single-session selection tests that test mathematics-, science- and English skills. The tests focus mainly on subject-related skills while requiring as little content knowledge as possible. During the research the question arose of whether a selection instrument based on dynamic testing principles could add to the predictive validity achieved by the single-session tests. This article summarizes the results of an investigation into this possibility and compares the findings of the dynamic testing with data of other groups reported in the literature.

Dynamic Testing

Dynamic testing procedures are often seen as more fair to use for selection than single-session tests in situations of unequal educational opportunity. Dynamic testing is intended to assess a student's learning potential or ability to benefit from instruction and is usually done using a test-teach-test type of selection mechanism. This kind of testing procedure involves an initial test without much instruction. The pretest provides data on the student's current level of functioning on the required task. Instruction, called an intervention, is then given to the student with the aim of teaching the student the necessary problem solving skills. The student's new level of functioning is tested during a post-test session. An improvement in the post-test results is seen as an indication that learning has occurred. The intervention phase should give students who did not have an adequate opportunity to develop their academic potential, a greater chance to achieve a fair test result (Hamers and Resing 1993).

On the other hand, indications in the existing literature are that although the concept of dynamic assessment may seem promising in a selection situation in which disadvantaged applicants have to be given an adequate chance to prove their ability to succeed, research results thus far have not convincingly supported this promise (Ruijssenaars *et al*. 1993).

There are a number of statistical issues involved in the use of pretest-post-test scores in the prediction of performance or in the analysis of the amount of learning during the intervention. In a test-teach-test situation the pretraining score reflects the student's currently achieved ability or performance level on a task. The posttraining score includes three major components, namely the initial performance level, the effect of exposure or practice due to repeated administration of the test and the effect of the intervention or training. Although an improvement in post-test score versus the initial pretest score can be taken as an indication of learning, the interpretation of the change is not psychometrically simple.

One problem is that the difference score is dependent on the pretest score. For example, students who scored highly initially may not be able to improve as much on the test as students who scored in the lower ranges due to statistical regression effects or a built-in ceiling effect in the test. Another problem with comparing the pre- and post-test scores, is that the instruction may be designed to modify the student's problem solving strategies. It is thus possible that the pre- and post-tests may be measuring two different constructs. The measurement of two different constructs complicates the interpretation of the pretest-post-test score differences in terms of a general learning potential construct. A simple post-test-pretest change score is usually not seen as a satisfactory measure of the amount of learning that has taken place. Alternative approaches are using the pretest and post-test scores as independent predictors in a regression analysis design or using the post-test scores only. For more detailed discussions of the psychometric problems associated with dynamic testing see Cascio and Kurtines (1977), Embretson (1987, 1992), Budoff (1987), Campbell and Reichardt (1991) and the contributions in Hamers *et al*. (1993).

In a selection context the predictive validity of selection instruments using a dynamic testing process is important. This predictive validity is usually analysed by using the pretest and post-test scores as predictors in regression analyses (Campbell and Reichardt 1991; Guthke 1993; Hamers and Sijtsma 1993). Little positive data exist on the predictive validity of dynamic assessments for further academic achievement. Nevertheless, post-test predictive validities are usually found to be marginally higher than pretest validities (Embretson 1992; Guthke 1993; Hamers and Sijtsma 1993). Reported research results have not shown a significantly higher

predictive validity of dynamic tests compared to that of traditional single-session academic selection tests. However, it has been found that dynamic tests are less sensitive to environmental factors, such as support given by parents, than the single-session tests (Guthke 1993).

There are also a number of logistical problems connected to dynamic assessment. A dynamic selection mechanism is not practical when large numbers of applicants have to be considered as it is resource intensive and time consuming. The cost effectiveness of such testing can be questioned even for small numbers of applicants (Guthke 1993; Embretson 1992).

However, based on the hope that dynamic testing could be more fair than other selection options, other South African entrance programmes adopted a dynamic approach to the admission of underprepared students, most notably at the University of Cape Town and the University of Natal. The aim of these programmes is to give underprepared applicants an optimal chance to prove that they have the ability to succeed at further study (University of Natal 1993; Haeck *et al*. 1997; Yeld and Haeck 1997). Predictive validity coefficients reported by these programs have not shown a significant correlation between the dynamic selection tests and the academic performance of educationally disadvantaged students (Haeck *et al*. 1997).

In the UNIFY case, the possibility that a selection instrument using a dynamic testing process could be more fair for assessing the learning potential of disadvantaged students than a single testing session led to an investigation based on the Raven Progressive Matrices (RPM). The results of the Raven investigation were expected to give an indication of whether further work should be done on developing a selection strategy for UNIFY which includes a test-teach-test type of selection instrument.

Raven Progressive Matrices

The RPM was chosen as a starting point for the development of a dynamic testing procedure instrument because of its status as one of the most culture-reduced tests available and its position of having been extensively used and researched (Jensen 1980; Arthur and Day 1994). The RPM tests consist of pattern (matrix) completion problems. Each problem builds on the previous problems and becomes steadily (progressively) more difficult.

According to Raven, Court and Raven (1988) and Raven, Raven and Court (1991) the matrices were developed to assess, as simply and unambiguously as possible, the eductive ability of a person. Eductive mental ability involves making meaning out of confusion. The ability to form comparisons and reason by analogy is assessed by the matrices. Effective eductive behaviour requires problem identification, re-conceptualization of the whole field, and the monitoring of tentative solutions for consistency with all the available information. Solving the Raven matrix problems involves the following: counting, identifying size changes, form changes, orientation changes, movement, combinations of patterns (repetitions, shading, pattern overlap) and the adding and subtracting of parts of patterns.

Two of the RPM tests were used in the present research. The first was the Raven Standard Progressive Matrices test (SPM) and the second was the Advanced Progressive Matrices test (APM). The SPM is intended for all age groups. The APM is intended as a test of intellectual efficiency for people of more than average intellectual capacity and is intended to differentiate between the top 10% of persons (Raven 1989). The Raven APM test consists of two test booklets. Set I consists of twelve problems covering all the intellectual processes and the full range of difficulty sampled by the SPM. Set II consists of 36 problems designed to discriminate between those who can solve all (or almost all) of the problems of the SPM. The types of problem in Set II are identical to those in Set I, but they increase in difficulty more steadily and become considerably more complex. Raven, Court and Raven (1988) state that the APM is particularly useful as an aid in selecting students for advanced science or technical studies.

The non-verbal content of the RPM was expected to give a fair assessment of ability to benefit from instruction by minimising the home and educational background-dependent influence of English proficiency and subject knowledge on the test scores. However, some students would have been better prepared to solve these kinds of problems than others due to differences in educational opportunity and home backgrounds. Nevertheless, all the UNIFY students came from disadvantaged backgrounds as shown by Zaaiman (1998). Solving the matrix problems was expected to be a new experience for most of them as only about three to four UNIFY students per class of thirty indicated that they had seen tests like the RPM before.

Research Questions and Methods

Two research questions were investigated, namely:

1. Will an instrument using a dynamic testing procedure contribute significantly to the prediction of performance in UNIFY?

2. How does the UNIFY student performance in the Raven tests compare to that of other groups reported in the (I/O) research literature?

Standard Progressive Matrices

The SPM test was used in the exploratory phase of the UNIFY dynamic testing investigation following Owen (1992) who found that South African black pupils do not do well on the SPM. He found an average score of 27.7 (out of 60 items) and a standard deviation of 10.7 for a sample of 1 093 Grade 9 black pupils from 28 schools (three in Gauteng and 25 in KwaZulu Natal). He did not give an indication of the ages of these pupils but, taking into account the diversity in the ages of disadvantaged pupils in South Africa, one could assume that their ages would have ranged widely around 14 years. Even though the UNIFY students had completed Grade 12, it was assumed that they would not do too well on the SPM given their lack of quality educational opportunity. This would allow room for an increase in scores after intervention.

The SPM was administered to the 1995 UNIFY students ($N = 147$) at the start of the academic year. Their performance on the SPM was used to evaluate the suitability of the SPM for a dynamic testing selection procedure for UNIFY. Interviews were held with a selected number of students to establish the kind of problems the students experienced in solving the matrices. The predictive validity of the SPM results for student performance in UNIFY was investigated using as a criterion the UNIFY final average, that is, the average of the final percentages attained in the five UNIFY subjects (biology, chemistry, English, mathematics and physics). The examination is directed at educational achievement with an emphasis on conceptual understanding, application of knowledge and problem solving. As such it can be seen as a good proxy for later job performance. The internal homogeneity of the UNIFY final average score was estimated by calculating the Cronbach's alpha reliability coefficient for the scores on the five subjects and amounted to .89. The mean of the final average score was 58.2% with a standard deviation of 7.8%.

The SPM data showed a significant correlation of 0.33 (significant at $\alpha = 0.01$) with the UNIFY final average score. This correlation compared well with the correlation of single-session selection tests with the final average performance in UNIFY. The correlation of the UNIFY mathematics selection test scores with final average performance was 0.39, the science selection test's correlation coefficient was 0.38 and the English selection test's correlation was 0.27.

The data showed that the opportunity for improvement in scores on the SPM was marginal. The average score for the whole group was 52.3 out of a possible maximum of 60 with a standard deviation 4.2. The distribution of scores was skewed to the left. This implied that the SPM was too easy for the UNIFY group to be used in a test-teach-test investigation.

Student interviews

Individual interviews were held with 14 of the students who did the SPM in 1995. The aim of the interviews was to assess the kinds of difficulty UNIFY students could be expected to experience when trying to solve the matrix problems. Students were invited to the interviews on the grounds of their scores in the SPM. Although the whole range of scores was covered, more students in the lower-scoring regions were invited. Attendance to the interviews was voluntary. All of the invited students chose to attend.

The invited students were presented with selected problems based on their test answers and asked to talk their way through solving each problem. The discussion started with problems they managed to solve in the SPM and progressed to problems they got wrong. Most of the students could work out the easier ways of solving the matrices and progressively build on the insights obtained during the solving of the problems. When the interviewed students got stuck, suggestions often helped them gain their own insight into solving a problem. None of the problems was solved for them. From these interviews the following summary of 'typical' UNIFY student problem areas in solving the SPM items was made:

- Seeing and interpreting overlap of shading (e.g. problem C12).
- Handling matrices containing more than one kind of change; e.g. both repetition of patterns, and form and orientation changes (e.g. problems D11, D12 and E7). Raven, Court and Raven (1988) called this type of error an 'incomplete solution' or 'incomplete correlate', referring to errors due to the failure to grasp all the variables which determine the nature of the correct figure required to complete a test item.
- Grasping the concept of adding or subtracting parts of a pattern (e.g. problems C11, E4, E6, E8 to E12).
- Focusing too much on details, rather than on the whole and not being able to distinguish between relevant and irrelevant information. Raven, Court and Raven (1988) called these types of error 'over-determined choices due to a 'confluence of ideas'. An illustration of

such an error is when the student would choose a figure which combined as many as possible of the individual characteristics shown in the matrix problem.
- Not considering all the rows or columns in the matrix problems as part of the matrix or leaving out one of the given figures in the evaluation of the matrix, again showing an inability to view the problem in a holistic manner.
- Attempting to use a previously successful technique without adapting to a new problem situation. This could connect to the 'arbitrary lines of reasoning' or 'wrong principle' error when the person uses a principle of reasoning qualitatively different from that demanded by the problem, compare Raven, Court and Raven (1988).

Raven, Court and Raven (1988) reported that incomplete correlate errors were found to be the most frequent in solving the APM problems. Wrong principle errors were more likely to be found at the lower scoring levels. Errors due to a confluence of ideas and repetition (simply selecting a figure identical with one of the three figures immediately next to the space to be filled) were not found by them to be common at the higher scoring levels.

The difficulties identified during the UNIFY interviews guided the design of the intervention patterns. One recommendation for improvement of the design of the intervention that became apparent was that the language used in the intervention should be kept as simple as possible. During the intervention every symbol should be given an explicit name, since it appeared that the students often did not know the English names for the shapes being used in the SPM. It should also be made clear that different solution strategies can exist; either working along rows, or working along columns.

Advanced Progressive Matrices

The exploratory work indicated that the SPM test was too easy for the UNIFY students and offered too little opportunity for improvement. It was therefore decided to use the APM for the development of the dynamic testing procedure. The APM was divided into two parallel versions, one to be used as a pretest and the other as a post-test. The test-teach-test phase of the investigation was run on the 1996 UNIFY student group. The first version of the test was applied to the selected UNIFY student group at the start of the academic year in February 1996. It was administered per group of about 30 students (total N = 126) with standard instructions. The students were asked to look at the first problem and decide what they thought the correct answer was. They then gave the answer and the group was asked whether everyone agreed. When general consensus seemed to exist, they were told to do the other problems on their own. They were told to expect the test to be difficult and not to worry if they could not do some of the problems as the group would get back to this type of test later in the year. They were also told that the test results would only be used for research purposes and would not count for any part of their UNIFY results. Nevertheless, most students seemed to take the test seriously and put a great deal of effort into completing the questions. They had about 45 minutes to complete the 20 questions, which proved to be adequate time for most of the students.

An intervention was developed by UNIFY teaching staff. The aim of the intervention was to investigate whether a student was able to benefit from UNIFY instruction. It was based on the essential nature of a UNIFY learning experience, that is focused on the gaining of both insight into the problem and problem-solving skills through experimentation, rather than being taught a recipe to solve the matrix problems. The intervention was designed using the UNIFY mathematics teaching approach, which involves allowing the students to gain insight into a problem without drilling them in specific methods. The students covered a topic called 'Patterns' for a week during the mathematics periods.

The intervention was standardized, as far as possible, by giving each student group the same tasks. The students were initially given simple pattern sequences which they had to complete. They also had to describe the pattern changes. Although these patterns were based on the principles of the Raven Progressive Matrices, none of the matrix problems were repeated. The patterns became progressively more difficult with changes in more than one characteristic and direction being included. After this, the students were asked to design their own pattern sequences and give them to a neighbouring student to solve.

The intervention was implemented in combination with the second version of the APM test in June 1996. This was later than expected, due to practical problems relating to the availability of students and instructors. The students wrote the test on the Monday after the intervention week. The predictive validity of the pre- and post-tests for final performance in UNIFY was determined. If significant predictive validity was found, further development of the dynamic testing procedure was envisaged. This would have included an investigation of whether the intervention actually made a difference and whether a significant improvement would have

occurred by simple test-retesting, without intervention by using a non-intervention control group.

This work also gave an opportunity to compare the performance on the matrices of the UNIFY student group consisting of disadvantaged black South African students to that of other groups published in research literature.

Results

Creation of the Parallel Test versions

The APM test was divided into two parallel versions to avoid repetition by the students of specific items in the pre- and post-tests. This was expected to minimise a possible artificial learning effect because of previous exposure to the test. In splitting the test it was attempted to create identical levels of the indicators of item difficulty and item total correlation for the two subtests, and an equal coverage of problem characteristics in each version.

The APM Manual reports upon two sets of empirical data (Raven *et al*. 1988). The first data set was used to compile the APM in 1962, and was based on data for adults and technical and university students in the UK. Data were given only for Set II, and no r_{iT} data was given. This set is called the Raven 1962 data set in this discussion. The other data set, reported in 1982, gives information for both APM sets for a group of 300 German candidates for officer posts and 241 university-entrance students. The mean age of the sample was 19.3 years with a standard deviation of 1.4 years. This German group corresponds with the UNIFY student group on age and educational level. This data set is called the German 1982 data set. More recent comparable data on Set II was found in Arthur and Day (1994). Their data referred to 202 students and other volunteers who were recruited from two university communities in the United States. The mean age of the sample was 21.4 years, with a standard deviation of 4.4 years. This data set is called the USA 1994 data set. These three data sets were used to make two parallel versions of the APM.

The characteristics of each problem were evaluated in an attempt to achieve a balanced spread of item characteristics. Most of the matrix problems consisted of a combination of characteristics. The identified characteristics were:

1. Number-counting
2. Linestyle
3. Size-changes
4. Symbols/shapes-repetition
5. Symbols/shapes-changes in form
6. Symbols/shapes-changes in orientation
7. Shading-repetition/changes
8. Shading-overlapping
9. Pattern-adding
10. Pattern-subtracting
11. Pattern-movement.

The resulting two parallel versions of the APM are given in Table 1. Each version consisted of 21 items with the first item intended as a practice item.

Pretest

The test was scored out of 20, ignoring the first practice problem. A group of 126 students wrote both the pre- and post-tests. The group on the pretest achieved a mean score of 12.2 with a standard deviation of 3.1. The Cronbach's α internal reliability coefficient for version 1 was 0.64. The test scores were more normally distributed, allowing more room for improvement through an intervention than the previous scores on the SPM. An evaluation of the items the students struggled with, pointed to the same matrix solution problems identified during the SPM interviews.

Post-test

After the intervention, the mean score on the post-test was 14.0 out of a possible maximum of 20, with a standard deviation of 2.5. The Cronbach's α internal reliability coefficient for the post-test was 0.59.

The correlation between the two Raven applications was 0.43. This is lower than what would have been expected from the Cronbach's α coefficients of the two tests if they measured the same construct (about the average of the two reliability coefficients, i.e. 0.62). The correlation could be smaller for two possible reasons.

The first possible reason is that the post- and pretests do not have a linear relationship because of a ceiling effect on the improvement of scores for the high-scoring pretest group. However, the inclusion of a quadratic pretest term in the regression of the post-test on the pretest was not statistically significant (increase of R^2 from 0.1833 to 0.1949, $F = 1.77$ with degrees of freedom 1,123 and significance of 0.19). The post-test scores therefore showed a mainly linear relationship with the pretest scores.

The second possibility is that the smaller correlation coefficient is due to the time lapse between the pre- and post-tests. The five-month time difference probably meant that factors such as motivation, interest and adjustment to UNIFY teaching and campus life had started to play a more significant role. The results of the Raven investigation included not only the student's reaction to the intervention but also

Table 1: The two parallel versions of the Raven APM

Version 1			Item characteristics												% answering correctly			r_{iT}	
It	APM Set	APM Item	1	2	3	4	5	6	7	8	9	10	11		Raven 1962	German 1982	USA 1994	German 1982	USA 1994
1	1	1	x	x												99		0.60	
2	1	3			x											100		0.55	
3	2	1	x			x		x							98	96	93	0.55	0.37
4	2	3											x		93	97	91	0.55	0.25
5	2	5										x			90	93	93	0.50	0.31
6	1	6	x													100		0.45	
7	2	9									x				88	99	89	0.61	0.40
8	2	10											x		88	94	88	0.60	0.39
9	2	11									x				84	95	93	0.60	0.49
10	2	13	x			x		x							79	83	66	0.44	0.23
11	2	15							x	x	x				79	88	81	0.51	0.36
12	2	19			x						x				64	76	75	0.43	0.24
13	2	22									x	x			50	71	48	0.45	0.40
14	2	23									x	x			49	72	62	0.43	0.42
15	2	25						x	x	x					45	52	45	0.39	0.37
16	2	26				x	x						x		38	51	53	0.37	0.28
17	2	27				x	x								38	53	43	0.30	0.31
18	2	28	x			x									32	34	35	0.32	0.29
19	2	31	x						x						24	38	42	0.25	0.36
20	2	33	x								x				24	29	37	0.28	0.26
21	2	35		x							x	x			14	33	38	0.22	0.33
		Σ	7	2	2	5	2	3	3	2	8	4	3	Mn	59.8	74.0	65.1	0.45	0.38

Table 1: Continued

Version 2			Item characteristics																
															% answering correctly			r_{iT}	
It	APM Set	APM Item	1	2	3	4	5	6	7	8	9	10	11		Raven 1962	German 1982	USA 1994	German 1982	USA 1994
1	1	2	x													100		0.59	
2	1	4			x											100		0.57	
3	1	8	x			x										99		0.52	
4	2	2	x		x		x								96	100	96	0.58	0.26
5	2	4										x			91	96	91	0.57	0.34
6	2	6	x												90	99	94	0.50	0.26
7	2	7									x				90	95	89	0.48	0.25
8	2	8							x						87	91	84	0.52	0.40
9	2	12										x			84	97	87	0.53	0.23
10	2	14											x		80	92	86	0.56	0.28
11	2	16										x			73	87	81	0.51	0.36
12	2	17		x			x								69	72	80	0.44	0.28
13	2	18					x								65	75	72	0.45	0.36
14	2	20							x	x	x				59	77	75	0.42	0.34
15	2	21					x	x	x						58	82	70	0.44	0.54
16	2	24								x		x			49	50	50	0.29	0.38
17	2	29			x	x		x							29	35	28	0.24	0.25
18	2	30							x						26	34	45	0.30	0.42
19	2	32							x				x		17	24	32	0.22	0.35
20	2	34	x				x								13	24	37	0.19	0.32
21	2	36										x			5	5	7	0.11	0.18
		Σ	5	1	3	2	5	2	5	2	2	5	2	Mn	60.1	73.0	66.9	0.43	0.32

Notes; Numbers 1 to 11 are the item characteristics
Σ gives the total number of items per characteristic
r_{iT} gives the correlation of the item with the total score; Mn gives the mean % answering correctly and r_{iT}

Table 2: The correlation (r) between pre- and post-test scores and UNIFY final average performance for groups with differing levels of initial ability

Maximum score on pretest	Group size	Pretest – Final (r)	Probability Pretest – Final r	Post-test – Final (r)	Probability Post-test– Final r	Pretest – post-test correlation (r)
18	126	0.15	0.09	0.13	0.14	0.43**
14	97	0.07	0.48	0.15	0.13	0.41**
13	79	0.12	0.28	0.18	0.11	0.45**
12	63	0.09	0.48	0.24	0.06	0.46**
11	45	0.17	0.28	0.29	0.05	0.43**
10	31	0.29	0.12	0.26	0.16	0.32

Notes: r = correlation.
Prob r = probability that $r = 0$.
** - two-tailed significance less or equal to 0.01.

development that occurred in the five month pre-post-test period.

Predictive Validity

The main aim of the Raven investigation was to see whether an instrument using a dynamic testing process would add significantly to the prediction of UNIFY final average. Neither of the applications of the APM correlated significantly with the UNIFY final average score (see Table 2).

Dynamic assessment tests are expected to give the most useful results for students with low initial achievement, due to a lack of exposure to the desired skill. Whether the intervention improved the predictive validity of the post-test for students with low pretest scores, was therefore checked. An inspection of the UNIFY final average scores versus the pretest scores showed that students who scored above 14 on the pretest were almost guaranteed to pass UNIFY. The correlation between the pre- and post-test performance and UNIFY average for student subgroups who scored 14 or below 14 on the pretest is shown in Table 2. Although the correlation coefficients are still not significant, indications are that the post-test is a better predictor than the pretest for the students who initially did not do well on the pretest. This does not hold for the group with a maximum score of 10, probably because of the small sample size involved and regression to the mean effects. The correlation between the pre- and post-tests is not significant for this group.

Although the correlation coefficients were not significant and sampling error cannot be ruled out as an explanation, the data support the idea that a post-intervention test can be more valid for the prediction of the performance of the initially low-scoring students. This is in line with Embretson (1987) who suggests that if the goal of a dynamic test is to improve ability estimates, the initial test score can be regarded as misleading or irrelevant to the goals of testing. The post-test can then be used on its own, while the pretest and intervention are regarded as part of the testing procedure.

Comparison with Other Groups Reported in the Literature

The performance of the UNIFY students in the pre- and post-tests was compared to that of the Raven 1962, the German 1982 and the USA 1994 groups described earlier in this article. Figure 1 gives the percentage of each group answering each item correctly, per item, for the UNIFY pre- and post-tests.

Table 3 gives the average of the percentages correct, per group, for the items for which data were available for all four groups (i.e. the APM set 2 items, see Table 1). The data between groups can be compared for each of the two versions. For the UNIFY group, version 1 gives the pre-intervention data and version 2 the post-intervention data.

Figure 1 and Table 3 show that the UNIFY students fared well in the pre-intervention test when compared with the other data groups. All three comparison data sets included significant numbers of university and other higher education students. This result is remarkable, considering the fact that the UNIFY students came from extremely disadvantaged backgrounds.

The overall UNIFY student performance improved after the intervention. This corresponds with Guthke (1993) who reports of increased performance on all intelligence tests after training. The UNIFY mean performance improved by $(14.0 - 12.2)/2.8 = 0.64$ (pooled) standard deviations after the intervention. The other groups' data were based on a single test session and one would also have expected their performance to improve after receiving the same intervention. It is difficult to

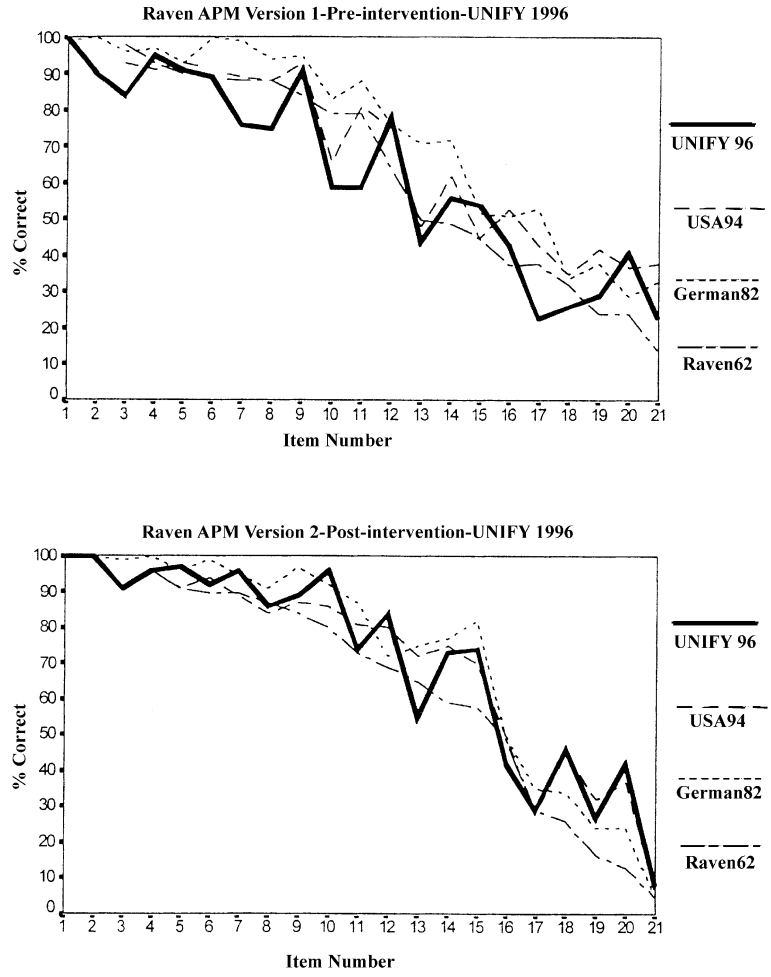

Figure 1: The percentage of a group answering each matrix problem correctly for the UNIFY 1996 (UNIFY96), USA 1994 (USA94), German 1982 (German82) and Raven 1962 (Raven62) groups

predict how much the scores of the other groups would have improved after intervention. Van der Molen *et al.* (1995) reported an improvement of 0.9 standard deviations on a number-completion test for 156 Dutch higher education students after relevant training. They regarded this to be a considerable improvement in performance when compared to that reported in other studies. The post-intervention improvement in the UNIFY scores is probably in the region to be expected for most groups.

Discussion

Although the post-test scores for the lower-scoring pretest group show the most promise as a valid predictor for UNIFY performance, the Raven investigation did not succeed in its aim of developing a sufficiently valid dynamic selection procedure for UNIFY selection. It was decided not to develop the idea of a dynamic selection procedure for UNIFY further, because of the non-significant results, the existing uncertainty in the predictive validity of dynamic assessment

Table 3: Group averages of the percentage answering an item correctly, taking only the APM set 2 items in each test version (18 items per test)

Group	Average percentage correct APM version 1	Average percentage correct APM version 2
Raven 1962	60	60
German 1982	70	69
USA 1994	65	67
UNIFY 1996	63	67

instruments, and the logistical problems involved in implementing a dynamic assessment method for large numbers of applicants. The high predictive validity and lack of predictive bias in the existing UNIFY selection tests led to a decision to devote the available time and resources to the further development of the existing tests. These tests can be described as aptitude-styled tests, aiming at testing subject-related problem-solving skills and insight, with as little as possible content knowledge required. Specifications of incoming student requirements, based on critical incident interviews with UNIFY staff members formed the basis for further improvement of these tests. A detailed description is given in Zaaiman (1998).

The Raven investigation could be criticised on a number of aspects. One problem could have been the long time lag between the first and second sessions. On the other hand, this time lag should have increased the influence of actual UNIFY teaching on student development. A reason for the lack of predictive validity for UNIFY performance could have been the lack of correspondence between the content of the RPM and the UNIFY programme content. Using subject matter that is more directed to the UNIFY programme itself might have led to a higher predictive validity. On the other hand, such improvement will not necessarily be ensured by using subject-related test content. Nevertheless, the basic dynamic testing principle of optimally supporting an underprepared student to show his/her full potential during testing should be taken into account during the evaluation and development of higher education selection instruments, even for single-session instruments.

Two internal validity issues should be mentioned as limitations of this study. The first one is that we did not have a non-intervention control group. This implies that it was not possible to separate the simple test-retest effect from the intervention effect. As a research strategy however, it was thought to be more efficient to study the combined effect first and should this effect be significant, try to separate its two components. Moreover the test-retest effect could be expected to be small in view of the fact that the retest was done with a parallel version of the test and the long time lag between the two sessions. The second internal validity issue has to do with the creation of parallel versions of the APM by splitting it in half. This procedure reduced the reliability in each version to relatively low levels for a cognitive ability instrument and obviously negatively influenced the predictive validities of the two halves. However, this was not considered to be a major problem in this Raven experiment since the unreliability could be expected to have the same effect on the predictive validity of the pretest and post-test. Moreover, a correction for attenuation would not lead to very different outcomes.

As to the second research question, the South African educationally disadvantaged students selected for UNIFY performed reasonably well on the Raven tests compared to the other groups reported in the literature. This apparent lack of cultural bias is certainly noteworthy. The UNIFY selection research results can make a contribution to current discussions on the validity of tests such as the Raven in different cultural contexts.

References

Arthur, W. and Day, D.V. (1994) Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement*, **54,** 394–403.

Budoff, M. (1987) The validity of learning potential assessment. In C.S. Lidz (ed.) *Dynamic Assessment: An Interactional Approach to Evaluating Learning Potential* (pp. 52–81). New York: The Guilford Press.

Campbell, D.T. and Reichardt, C.S. (1991) Problems in assuming the comparability of pretest and posttest in autoregressive and growth models. In R.E. Snow and D.E. Wiley (eds) *Improving Inquiry in Social Science. A Volume in Honor of Lee J. Cronbach* (pp. 201–220). New Jersey: Lawrence Erlbaum Associates Publishers.

Cascio, W.F. and Kurtines, W.M. (1977) A practical method for identifying significant change scores. *Educational and Psychological Measurement*, **37,** 889–895.

DACST (1996) *South Africa's Green Paper on Science and Technology. Preparing for the 21st Century.* The Department of Arts, Culture, Science and Technology, Pretoria, January 1996.

Embretson, S.E. (1987) Toward development of a psychometric approach. In C.S. Lidz (ed.) *Dynamic Assessment: An Interactional Approach to Evaluating Learning Potential* (pp. 141–170). New York: The Guilford Press.

Embretson, S.E. (1992) Measuring and validating cognitive modifiability as an ability: a study in the spatial domain. *Journal of Educational Measurement*, **29,** 25–50.

Guthke, J. (1993) Developments in learning potential assessment. In J.H.M. Hamers, K. Sijtsma and A.J.J.M. Ruijssenaars (eds.), *Learning Potential: Assessment. Theoretical, Methodological and Practical Issues* (pp. 43–67). Amsterdam/Lisse: Swets and Zeitlinger B.V.

Haeck, W., Yeld, N., Conradie, J., Robertson, N. and Shall, A. (1997) A developmental approach to mathematics testing for university admissions and course placement. *Educational Studies in Mathematics*, **33,** 71–91.

Hamers, J.H.M. and Resing, W.C.M. (1993) Learning potential assessment: Introduction. In J.H.M. Hamers, K. Sijtsma and A.J.J.M. Ruijssenaars (eds.), *Learning Potential Assessment. Theoretical,*

Methodological and Practical Issues (pp. 23–41). Amsterdam/Lisse: Swets and Zeitlinger B.V.

Hamers, J.H.M. and Sijtsma, K. (1993) Learning potential assessment: epilogue. In J.H.M. Hamers, K. Sijtsma and A.J.J.M. Ruijssenaars (eds.) *Learning Potential Assessment. Theoretical, Methodological and Practical Issues* (pp. 356–376). Amsterdam/Lisse: Swets and Zeitlinger B.V.

Hamers, J.H.M., Sijtsma, K. and Ruijssenaars, A.J.J.M. (eds.) (1993) *Learning Potential Assessment. Theoretical, Methodological and Practical Issues.* Amsterdam/Lisse: Swets and Zeitlinger B.V.

Jensen, A.R. (1980) *Bias in Mental Testing.* London: Methuen and Co. Ltd.

Owen, K. (1992) The suitability of Raven's Standard Progressive Matrices for various groups in South Africa. *Personality and Individual Differences*, **13,** 149–159.

Raven, J. (1989) The Raven Progressive Matrices: a review of national norming studies and ethnic and socioeconomic variation within the United States. *Journal of Educational Measurement*, **26,** 1–16.

Raven, J.C., Court, J.H. and Raven, J. (1988) *Advanced Progressive Matrices. Raven Manual: Section 4.* Oxford: Oxford Psychologists Press.

Raven, J., Raven, J.C. and Court, J.H. (1991) *Raven's Progressive Matrices and Vocabulary Scales. Section 1. General Overview* (1991 edition). Oxford: Oxford Psychologists Press.

Ruijssenaars, A.J.J.M., Castelijns, J.H.M. and Hamers, J.H.M. (1993) The validity of Learning Potential Tests. In J.H.M. Hamers, K. Sijtsma and A.J.J.M. Ruijssenaars (eds.) *Learning Potential Assessment: Theoretical, Methodological and Practical Issues* (pp. 69–82). Amsterdam/Lisse: Swets and Zeitlinger B.V.

University of Natal (1993) Access, selection and educational development: A case study. Supporting documentation prepared for the Exemptions Committee of the Matriculation Board, Committee of University Principals. The TTT Programme, University of Natal, 12 October 1993.

Van der Molen, H.T., Te Nijenhuis, J. and Keen, G. (1995) The effects of intelligence preparation. *European Journal of Personality*, **9**, 43–56.

Wood, R., Hamer, G., Johnson, C. and Payne, T. (1997) Selection for a profession: a case Study. In N. Anderson and P. Herriot (eds.) *International Handbook of Selection and Assessment.* Chichester: John Wiley and Sons Ltd.

Yeld, N. and Haeck, W. (1997) Educational histories and academic potential: can tests deliver? *Assessment and Evaluation in Higher Education*, **22**, 5–16.

Zaaiman, H. (1998) *Selecting Students for Mathematics and Science: The Challenge Facing Higher Education in South Africa.* Pretoria: HSRC Publishers.