

An Empirical and Strategic Analysis of the Utility of the Xtend–GeneSys guidance process for ELTEC

Author: Paul Barrett
Specialist Consultant (spcon@cs.com)

Date: 15th January, 2000

This report is divided into two sections:

Section 1: Executive Summaryp. 2–3

Bullet–point summary of the main conclusions, referenced to analyses in section 2. Also, four key comments are made.

Section 2: Empirical Analysesp. 4–27

There are six hypotheses of interest tested in this section:

1. Can GeneSys variables discriminate between students who succeed or fail their NVQ training?
2. Can GeneSys variables predict the attained maximum NVQ level for a student?
3. Can GCSE indicator variables discriminate between students who succeed or fail their NVQ training?
4. Can GCSE indicator variables predict the attained maximum NVQ level for a student?
5. What are the relationships between GCSE indicator variables and GeneSys variables?
6. Are starting SOC code areas associated with GeneSys Occupational Interest variable profiles?
7. Is Success or Failure with NVQ training related to the overall level of Interest in Any Occupational categories?

Appendix ... p.24

A.1. Statistica Fieldname Specification for ELTECFIN.sta

Section 1: Executive Summary

The following 7 conclusions can be drawn from this report. Each conclusion is that following from a specific hypothesis test, referenced by page number.

1. The GeneSys variables (either singly or in a joint prediction function) do not predict the likelihood of failure or success to complete an NVQ qualification training course (with grade > 0) at a level which has substantive value to ELTEC (*p.5*)
2. Given a linear, additive-unit prediction model, **four** GeneSys variables predict NVQ maximum attainment within the success group with 72% accuracy. GRT2 – Verbal Reasoning and Abstract Reasoning, and OIP Assertiveness and Scientific Interest (*p.8*)
3. With only 2% increased accuracy over the base rate classification, it is clear that there is no substantive predictive function (predicting successful NVQ course completion from GCSE indicator variables) available from GCSE data. However, given 70% of students with at least one GCSE succeed in gaining an NVQ qualification, this is a key predictor as it stands (*p.11*)
4. Average Pass Rate, computed over all GCSEs taken by an individual, correlates 0.44 with NVQ level attained (*p.14*)
5. There is a substantive relationship between GCSE average pass rate, the number of GCSEs attained at grade C and above, and GeneSys GRT2 Verbal and Abstract reasoning ability (*p.16*)
6. GeneSys variables are substantively associated with SOC broad occupational categories, showing good construct validity (*p.19*)
7. Interest level, indexed by the count of the number of GeneSys interests for a student that are greater than 1 standard deviation from each group mean interest, is not associated with success or failure to complete an NVQ qualification course (*p.21*)

Some Comments

1. The primary aim of the analyses above is to examine the Xtend–GeneSys scales in terms of their predictive utility for NVQ course completion and level. This analysis was extended into the utility of GCSE scores solely as a comparison with GeneSys prediction functions. Finally, some limited analysis of Occupational Interest scale validity and utility was also undertaken.
2. There is much of potential information value within the dataset, but the analysis of these kinds of data is outside the scope of this particular analysis (e.g. Gender and Race bias, sharpened definitions of failure, optimised prediction strategies for success/failure identification). Each of these items requires a strategic analysis plan and a firm goal. Without these, much expensive time and effort can be wasted on analyses that are later considered to be irrelevant or obsolete.
3. I have been careful to distinguish between those analyses that might be called GeneSys Evaluations vs those that might be called Program Evaluation. Given the number of variables available, and the "drilling down" that can take place amongst and within them, I have confined my analyses to those that are specific to GeneSys Evaluations only. To explore student groups who "Fail" or "Succeed" in detail, requires analysis of other kinds of information, as well as a comprehensive examination of the constituent samples currently composing the broad failure sample we have used to date. The problem is, we have too few students in the 1998 cohort to permit much drilling down beyond a group or two, especially when using joint variable functions.
4. If ELTEC are considering production of an optimised prediction system, then I would recommend that they reconsider their database and information acquisition in light of this, and set firm and achievable targets for program evaluation. However, it might be informative to engage in a cost–benefit model analysis before undertaking such a major strategic exercise – in order to ensure that likely savings outweigh the expertise and operational costs of implementing such a strategy. I would also recommend that they consider a multi–site trial/strategy – partly to offset costs, and perhaps just as important, in order to cross–validate any results for regional, cultural, environmental differences.

Section 2: Empirical Analyses

<p>Initial Client Database from Psytech International: Contains 96 variables, 5653 cases</p>	ELTEC99.sta
<p>Augmented file on which many analyses are undertaken: Contains 112 variables, 5653 cases</p>	ELTECFIN.sta
<p>Augmented file on which some analyses are undertaken: Contains 113 variables, 5653 cases This file is simply ELTECFIN.sta with one extra variable (SOCATS) which is used to hold the SOC categorisation field value (Hypothesis #6).</p>	ELTECSOC.sta
<p>Augmented file on which an analysis is undertaken: Contains 113 variables, 5653 cases This file is simply ELTECFIN.sta with one extra variable (SOCATS) which is used to hold the SOC categorisation field value (Hypothesis #6). However, this file also has all the GeneSys scale scores for GRT2 and OIP standardized (with mean 0.0 and SD of 1.0). This was used specifically to create Figure 1 in Hypothesis #6 analysis.</p>	ELTECSTD.sta
<p>Augmented file on which an analysis is undertaken: Contains 114 variables, 5653 cases This file is simply ELTECSTD.sta but with an extra variable added – INTLEVEL. This variable holds the count of the number of OIP interests (0–7) that are 1 standard deviation above their respective mean, for each student.</p>	ELTECINT.sta
<p>Subset file using Jan. 98 through to Dec.98 (incl) cohort: Contains 112 variables, 1039 cases This subset file was created by selecting cases from ELTECFIN which possessed a START DATE between January 1st 1998 through to 31st December 1998 inclusive.</p>	ELTEC98.sta

All variables names used in this section are jointly referred with their variable number within the Statistica file used for a particular analysis. **Appendix A.1** contains the variable listing for files ELTECFIN.sta and ELTEC98.sta.

Hypothesis Tests

1. Can GeneSys variables discriminate between students who succeed or fail their NVQ training?

Analysis File: Uses ELTEC98.sta

Here, we first have to define the constituent properties of the "success" and "fail" groups.

The **Success** group is defined as consisting of those students who possess a START DATE for training, and who have an entry > 0 in at least one of the QUALnLEV (qualification attained at a particular level) fields.

The **Fail** group is defined as consisting of those students who possess a START DATE for training, but who do not have an entry in any of the QUALnLEV fields, and are flagged as "Left Training" on the LEAVERCO (v.31) variable. This is the broadest definition of "failure" – encapsulating the totality of cost incurred by any student not successfully completing (greater than LEVEL 1, 2, or 3) a NVQ training qualification.

A new group variable (denoted SUCCESS (v.112)) was created, with two values: Fail, Success, keyed for each individual who met the filter criteria.

Table 1: Summary frequency of Fail and Success Groups in the Jan98–Dec98 Cohort

SUCCESS: 0 = Left, Non-Achievers, 1 = NVQ Achievers Group				
BASIC STATS	Count	Cumul. Count	Percent	Cumul. Percent
Fail	344	344	33.10876	33.1088
Success	361	705	34.74495	67.8537
Missing	334	1039	32.14629	100.0000

We have 344 in the Fail group, and 361 in our success group, out of a total sample size of 1039 students in this particular cohort. Percentages in the table are percentages of the total sample size.

Returning to our hypothesis: "Can GeneSys variables discriminate between students who succeed or fail their NVQ training?" we now have a criterion variable against which we can estimate the predictive utility of the GeneSys variables.

Table 2: The GeneSys variables used as predictors**Graduate Reasoning Test #2 (GRT2)**

33	_GRT2_VR	8.0	-9999	Test Score - Verbal Reasoning
34	_GRT2_NR	8.0	-9999	Test Score - Numerical Reasoning
35	_GRT2_AR	8.0	-9999	Test Score - Abstract Reasoning

Occupational Interest Profile (OIP)

40	_OIP_VEN	8.0	-9999	OIP - Venturesome
41	_OIP_PHL	8.0	-9999	OIP - Phlegmatic
42	_OIP_RAD	8.0	-9999	OIP - Flexible
43	_OIP_GRE	8.0	-9999	OIP - Gregarious
44	_OIP_ASS	8.0	-9999	OIP - Assertive
45	_OIP_PER	8.0	-9999	OIP - Persuasive
46	_OIP_SCI	8.0	-9999	OIP - Scientific
47	_OIP_PRA	8.0	-9999	OIP - Practical
48	_OIP_ADM	8.0	-9999	OIP - Administrative

Occupational Interest Profile (OIP) (cont).

49	_OIP_NUR	8.0	-9999	OIP - Nurturing
50	_OIP_ART	8.0	-9999	OIP - Artistic
51	_OIP_LOG	8.0	-9999	OIP - Logical

***Note:** Mechanical Reasoning was not used in this or any other analysis as there were only 102 cases with test scores amongst the "Success" and "Fail" groups. This is in contrast to all 703 cases with a success/fail code also possessing GRT2 and OIP test scores. The statistics used in this hypothesis analysis require that predictor variable scores exist for every case, thus, if the MRT2 test scores are utilised, this would mean the loss of 602 cases. This was considered an unacceptable loss of data.

The particular statistical analyses used for this hypothesis test were linear discriminant function analysis (in this particular binary response variable case, equivalent to a multiple linear regression), and logistic regression. Both these methods are suited to the prediction scenario, differing only in their assumptions concerning the distributions of the criterion variable (normal vs logistic) and those of the predictor variables. Essentially, both methods permit us to define our prediction in terms of a "classification equation", which enables us to assign a student into either a success or fail group, just using their optimally weighted test scores on the GeneSys variables. The essential result from these analyses is a classification table. This kind of table tabulates the actual outcome (success or failure) against the predicted outcome that has been computed using the optimally weighted GeneSys variables.

For the linear discriminant function analysis, we have:

Table 3: Linear Discriminant Function Classification Table

Classification Matrix (eltec98.sta)			
DISCRIM. ANALYSIS	Rows: Observed classifications Columns: Predicted classifications		
Group	Percent Correct	Fail p=.48791	Success p=.51209
Fail	55.39359	190	153
Success	61.11111	140	220
Total	58.32148	330	373

The multiple (canonical) R for this function that expresses the correlation between all the GeneSys variables jointly and the criterion is 0.22 ($p < 0.005$). We have a 58% classification accuracy, and a false positive rate of 45% (i.e. we predicted 153 successes who in actual fact failed, out of a total *failure* sample of 343 cases). The false negative rate is 39% (i.e. we predicted 140 cases to fail, who actually went on to succeed, out of a total *success* sample of 360 cases). For the logistic regression analysis, we were unable to achieve any greater accuracy. In fact we equalled the performance of the linear discriminant function.

If we examine the correlations between each of the GeneSys variables and the prediction criterion, we observe the figures in Table 4 below. I have used Pearson correlations as a reasonable estimate of the size of relationship (equivalent to point-biserial correlation).

Table 4: Pearson Correlations between the criterion variable and each GeneSys variable.

Variable	SUCCESS
_GRT2_VR	.00
_GRT2_NR	-.00
_GRT2_AR	.06
_OIP_VEN	-.01
_OIP_PHL	.04
_OIP_RAD	-.03
_OIP_GRE	-.03
_OIP_ASS	-.02
_OIP_PER	-.03
_OIP_SCI	.04
OIP_PRA	.15
_OIP_ADM	-.02
_OIP_NUR	-.10
_OIP_ART	-.01
_OIP_LOG	.09

As can be seen from this table, the OIP variable "Practical" is the best individual predictor. However, this only accounts for 2.25% of the variation in the criterion variable.

Result: Using the maximum number of available GeneSys variables, we only achieve 58% classification accuracy (50% represents classification accuracy achieved by chance alone). The maximum explanatory power of any individual GeneSys variable is only 2.25%.

Conclusion: The GeneSys variables (either singly or in a joint prediction function) do not predict the likelihood of failure or success at a level which has substantive value to ELTEC.

Caveat: The definition of the criterion variable is extremely broad. This may account for the almost chance-level predictions as the criterion is confounded by a plethora of perhaps quite unrelated variables.

2. Can GeneSys variables predict the attained maximum NVQ level for a student?

Analysis File: Uses ELTEC98.sta

Here, we focus on the group of students who do attain an NVQ qualification. The indicator variable INDIV (v.108) indexes the maximum level of NVQ attained, regardless of how many were attained. So, here we examine whether GENESYS variables can predict NVQ attainment level, measured using a 1, 2, or 3 level value.

The statistical analysis adopted here, for simplicity, is a linear multiple regression model. We might have used nonlinear polychotomous ordinal logistic regression, but, the computational complexity of the technique and the explanation of its parameterisation are probably not cost-effective as yet. Linear function methods are known to slightly underestimate the classification accuracy of these more suitable (from a statistical viewpoint) methods.

So, given our criterion of maximum attained NVQ level, and the 15 GeneSys predictor variables noted in Table 2 above, we observe the following beta regression weights for our variables (the regression weights are those weights applied to each of the GeneSys variables, such that when we add up the values of these multiplications, we achieve a result that should predict a level value of 1, 2, or 3 for every case). Basically, the larger the weight, the more important that variable is to the overall prediction of NVQ level.

Table 5: Multiple Regression Weights, predicting NVQ Maximum Attainment from 15 GeneSys variables.

Regression Summary for Dependent Variable: INDIV						
MULTIPLE REGRESS.	R= .53423569 R ² = .28540778 Adjusted R ² = .25424823 F(15,344)=9.1596 p<.00000 Std.Error of estimate: .45501					
N=360	BETA	St. Err. of BETA	B	St. Err. of B	t(344)	p-level
Intercept			.338108	.339925	.99466	.320603
_GRT2_VR	.196236	.066178	.017465	.005890	2.96526	.003235
_GRT2_NR	.004630	.071446	.000447	.006892	.06480	.948370
_GRT2_AR	.272156	.069374	.026632	.006789	3.92303	.000106
_OIP_VEN	-.067929	.060743	-.005889	.005266	-1.11830	.264220
_OIP_PHL	.053737	.054253	.004331	.004372	.99049	.322633
_OIP_RAD	-.044630	.055099	-.004646	.005735	-.81000	.418502
_OIP_GRE	.067600	.059202	.005792	.005073	1.14186	.254308
_OIP_ASS	.156415	.062676	.012242	.004906	2.49562	.013042
_OIP_PER	-.093683	.071371	-.007425	.005657	-1.31262	.190187
_OIP_SCI	.200958	.055217	.015059	.004138	3.63942	.000315
_OIP_PRA	-.006299	.066226	-.000435	.004576	-.09511	.924285
_OIP_ADM	.000362	.068292	.000023	.004431	.00530	.995778
_OIP_NUR	.029751	.059626	.002094	.004197	.49896	.618124
_OIP_ART	.012865	.059336	.000894	.004125	.21681	.828482
_OIP_LOG	-.042335	.070803	-.003704	.006195	-.59792	.550285

From this table, we can see that our predicted values correlate **0.504** with our actual criterion values (adjusted for the number of predictors). If we select out just those predictors that are statistically significant, we are left with GRT2 – Verbal Reasoning and Abstract Reasoning, and OIP Assertiveness and Scientific Interest. If we now re–run the regression analysis using just these predictors, we obtain:

Table 6: Multiple Regression Weights, predicting NVQ Maximum Attainment from a reduced subset of GeneSys Variables.

Regression Summary for Dependent Variable: INDIV						
MULTIPLE REGRESS.	R= .51823907 R ² = .26857173 Adjusted R ² = .26033029 F(4,355)=32.588 p<.00000 Std.Error of estimate: .45315					
N=360	BETA	St. Err. of BETA	B	St. Err. of B	t(355)	p-level
Intercept			.259255	.138506	1.871798	.062057
_GRT2_VR	.195985	.060234	.017443	.005361	3.253711	.001248
_GRT2_AR	.266408	.060111	.026069	.005882	4.431960	.000012
_OIP_ASS	.120216	.046379	.009409	.003630	2.592022	.009935
_OIP_SCI	.189384	.046311	.014192	.003470	4.089428	.000054

All predictors are significant. The Adjusted Multiple correlation is **0.51**.

The prediction equation is of the form:

$$NVQ'_L = a + b_1 \cdot GRT2_VR + b_2 \cdot GRT2_AR + b_3 \cdot OIP_ASS + b_4 \cdot OIP_SCI$$

where

NVQ'_L = predicted NVQ maximum attainment level (nearest integer)

a = Intercept constant

$b_1...b_4$ = unstandardized regression weights

if we insert our regression weights into this equation, and compute a value for case 3 in this data file, we have ...

$$NVQ'_L = 0.2593 + 0.0174 \cdot GRT2_VR + 0.026 \cdot GRT2_AR + 0.0094 \cdot OIP_ASS + 0.0142 \cdot OIP_SCI$$

Given ... $GRT2_VR = 7$, $GRT2_AR = 7$, $OIP_ASS = 33$, $OIP_SCI = 26$... we have ...

$$NVQ'_L = 0.2593 + 0.0174 \cdot 7 + 0.026 \cdot 7 + 0.0094 \cdot 33 + 0.0142 \cdot 26$$

$$NVQ'_L = 1.24 \quad \dots \text{then rounded to the nearest whole number} \dots = 1$$

The actual level achieved for this individual was 1 also – thus our prediction equation was accurate for this individual.

These four GeneSys variables predict NVQ outcome substantially better than chance. The classification table computed by applying the prediction weights to each of the 4 variables is:

Table 7: NVQ Predicted Maximum Attainment: Classification Table using optimal GeneSys variable weights

Summary Frequency Table (tempregr.sta)			
BASIC STATS	Marked cells have counts > 10 (Marginal summaries are not marked)		
INDIV	Predicted Level 1	Predicted Level 2	Row Totals
Level 1	103	54	157
Level 2	41	156	197
Level 3	0	6	6
Column Totals	144	216	360

Classification Accuracy is 72%. To put this into context, if we were to make predictions at random for our 360 cases, given that we preserve the outcome proportions of 0.436 cases at Level 1, 0.547 at Level 2, and 0.017 at Level 3, our classification table would look like:

Table 8: NVQ Predicted Maximum Attainment: Classification Table composed using entirely random assignment of cases

Summary Frequency Table (tempregr.sta)			
BASIC STATS	Marked cells have counts > 10 (Marginal summaries are not marked)		
INDIV	Predicted Level 1	Predicted Level 2	Predicted Level 3
Level 1	68	86	3
Level 2	86	108	3
Level 3	3	3	0

It is clear that GeneSys prediction of NVQ maximum attainment is substantively better than chance.

Result: Given a linear, additive unit prediction model, **four** GeneSys variables predict NVQ maximum attainment within the success group with 72% accuracy. GRT2 – Verbal Reasoning and Abstract Reasoning, and OIP Assertiveness and Scientific Interest.

Conclusion: With 72% classification accuracy, it might be productive to build in the prediction equation into GeneSys, and use this as part of a modified report system.

3. Can GCSE indicator variables discriminate between students who succeed or fail their NVQ training?

Analysis File: Uses ELTECFIN.sta

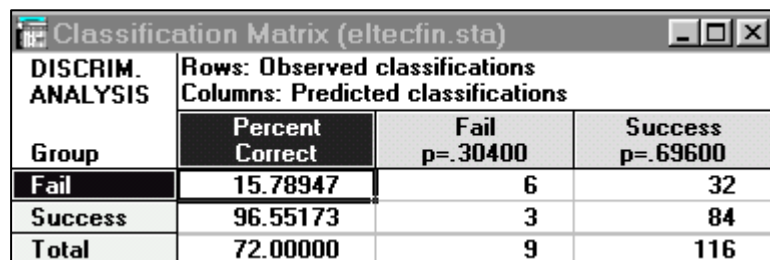
For this analysis, we revert to the larger ELTECFIN.sta, as we require as much data as possible on GCSE variables. In the ELTEC98 subset file, we only have 160 cases with GCSE results entered. Crosstabbing this with the criterion variable SUCCESS yields only 113 valid cases. In the main file, we have 356 cases. Crosstabbing these data with the criterion variable SUCCESS yields 125 valid cases – slightly better but still lower than optimal.

For the purpose of the GCSE variable analyses (both this one and the next two), 6 new indicator variables were created via a small program. These are:

88	GCSE_ENT	8.0	-9999	Number of GCSE's Entered
89	GCSEFAIL	8.0	-9999	Number of GCSE's Failed (Grade 1)
90	GCSE_F	8.0	-9999	Number of GCSEs with grade F and above
91	GCSE_C	8.0	-9999	Number of GCSEs with grade C and above
92	AVG_PASS	8.2	-9999	Average Pass Mark over all GCSEs taken
93	MAX_PASS	8.0	-9999	Maximum Passmark across all GCSEs taken

These six variables were entered into a linear discrimination function, as per hypothesis test #1 above (p.5). The resultant classification table is...

Table 9: Linear Discriminant Function Classification Table (GCSE variable predictors)



DISCRIM. ANALYSIS	Rows: Observed classifications Columns: Predicted classifications		
	Percent Correct	Fail p=.30400	Success p=.69600
Fail	15.78947	6	32
Success	96.55173	3	84
Total	72.00000	9	116

The multiple (canonical) R for this function that expresses the correlation between all the GeneSys variables jointly and the criterion is 0.31 ($p = 0.06$). We have a 72% classification accuracy, and a false positive rate of 84% (i.e. we predicted 32 successes who in actual fact failed, out of a total *failure* sample of 38 cases). The false negative rate is just 3% (i.e. we predicted 3 cases to fail, who actually went on to succeed, out of a total *success* sample of 87 cases). In effect, given the high 70% base rate of success, the function is over-predicting success, and dramatically under-predicting failure (only 16% correct failures).

What is happening here, unlike with our use of the GeneSys variables in Hypothesis #1, is that there is a confound between a student having a GCSE entry and eventual success on an NVQ training course. That is, there is already a 70% chance that a student with at least one GCSE will attain an NVQ qualification, before we begin any optimised prediction. In this context, it can be seen that the optimised weights add 2% more accuracy to this 70% "expected" base rate.

A closer examination of the six predictor variables demonstrated that the GCSEFAIL (v.89 Number of GCSEs Failed) was badly skewed, in that only 5.3% of those who had an entry for at least one GCSE failed any of them. This variable was subsequently dropped and the whole analysis rerun. However, as might be expected, there was no change to the classification results.

It was also considered important to examine the GCSE indicator variable correlation matrix, for possible collinearity (two variables correlate very highly with one another) amongst the predictor variables. This correlation matrix is given below..

Table 10: Correlations between GCSE indicator variables

Correlations (eltecfin.sta)					
BASIC STATS					
Marked correlations are significant at $p < .05000$					
N=356 (Casewise deletion of missing data)					
Variable	GCSE_ENT	GCSE_F	GCSE_C	AVG_PASS	MAX_PASS
GCSE_ENT	1.00	.95	.50	.22	.33
GCSE_F	.95	1.00	.57	.39	.43
GCSE_C	.50	.57	1.00	.85	.78
AVG_PASS	.22	.39	.85	1.00	.83
MAX_PASS	.33	.43	.78	.83	1.00

It can be seen that collinearity is in fact present amongst all variables. This fact contraindicates the use of any regression equation (which assumes independence between predictor variables).

As a final simple test, it was decided to examine the relationship between each GCSE indicator variable and the criterion variable of SUCCESS. These results are given below ...

Table 11: Correlations between the Criterion variable SUCCESS and the GCSE indicator variables.

Correlations (eltecfin.sta)	
BASIC STATS	
Marked correlations are significant at $p < .05000$	
N=125 (Casewise deletion of missing data)	
Variable	SUCCESS
GCSE_ENT	.14
GCSE_F	.18
GCSE_C	.08
AVG_PASS	.08
MAX_PASS	-.05

This table indicates that there are no substantive relationships between any indicator GCSE variable and the criterion variable SUCCESS.

Result: Given a linear discriminant function analysis composed of 6 GCSE predictor variables, discriminating between successful and unsuccessful NVQ students, a classification accuracy of 72% was obtained. However, it was shown that this is a spurious result, caused by an excessively high base rate of 70% of students with at least one GCSE completing at least one successful NVQ qualification. Given this fact, it can be seen that the optimised weights within the prediction equation account for just 2% extra prediction above expected chance levels. Further analysis indicated that none of the variables individually correlated substantively with the criterion, a result in line with that of Hypothesis #1, using the GeneSys variables.

Conclusion: With only 2% increased accuracy over the base rate classification accuracy, it is clear that there is no substantive predictive function available from GCSE data. However, given 70% of students with at least one GCSE succeed in gaining an NVQ qualification, this is a key predictor as it stands.

4. Can GCSE indicator variables predict the attained maximum NVQ level for a student?

Analysis File: Uses ELTECFIN.sta

Bearing in mind the conclusions concerning the multicollinearity of the 5 GCSE indicator variables used in Hypothesis #4 test above, and also given the exclusion of the GCSEFAIL (v.89) variable due to very low numbers of failures, it was decided to approach this particular analysis with some caution. Although we might proceed with a multiple linear regression (or the polychotomous ordinal logistic model), it is unlikely that we will achieve anything other than that we might have achieved with standard bivariate correlation analysis. However, in order that we might at least see the results, a multiple linear regression was implemented, using the 5 predictor variables as in the previous analysis, against the criterion variable INDIV (v.108) – which is indexing the maximum attained NVQ level for a student. The results were:

Table 12: Multiple Regression Weights, predicting NVQ Maximum Attainment from 5 GCSE variables.

Regression Summary for Dependent Variable: INDIV						
MULTIPLE REGRESS.	R= .54144953 R ² = .29316760 Adjusted R ² = .24953597 F(5,81)=6.7192 p<.00003 Std.Error of estimate: .32453					
N=87	BETA	St. Err. of BETA	B	St. Err. of B	t(81)	p-level
Intercept			-.101806	.468767	-.21718	.828616
GCSE_ENT	.535014	.405664	.107919	.081828	1.31886	.190932
GCSE_F	-.148362	.381775	-.027401	.070511	-.38861	.698584
GCSE_C	-.802221	.256370	-.111291	.035566	-3.12915	.002437
AVG_PASS	1.132054	.315849	.379042	.105755	3.58416	.000576
MAX_PASS	-.092508	.211262	-.029118	.066497	-.43788	.662638

From this table, we can see that our predicted values correlate **0.50** with our actual criterion values (adjusted for the number of predictors). If we select out just those predictors that are statistically significant, we are left with GCSE_C (the number of C-level and greater passes) and AVG_PASS (the average of all GCSEs taken). If we now re-run the regression analysis using just these predictors, we obtain:

Table 13: Multiple Regression Weights, predicting NVQ Maximum Attainment from a subset of GCSE variables.

Regression Summary for Dependent Variable: INDIV						
MULTIPLE REGRESS.	R= .46481580 R ² = .21605373 Adjusted R ² = .19738834 F(2,84)=11.575 p<.00004 Std.Error of estimate: .33561					
N=87	BETA	St. Err. of BETA	B	St. Err. of B	t(84)	p-level
Intercept			.829961	.251613	3.29856	.001426
GCSE_C	-.310090	.183565	-.043018	.025466	-1.68926	.094878
AVG_PASS	.698900	.183565	.234011	.061463	3.80736	.000266

Now, only one of the predictors is significant, AVG_PASS. This is what one might have expected from the arguments above. Note also that the number of cases available for this analysis (N=87) has dropped substantially. So, in order to gauge the predictive capability of each GCSE variable uniquely, we move onto standard Pearson correlations between each of the predictor variables and the criterion variable.

Table 14: Correlations between the 5 GCSE predictor variables and the Maximum NVQ level qualification attained

Correlations (eltecfin.sta)	
BASIC STATS	Marked correlations are significant at $p < .05000$ N=87 (Casewise deletion of missing data)
Variable	INDIV
GCSE_ENT	.20
GCSE_F	.28
GCSE_C	.28
AVG_PASS	.44
MAX_PASS	.38

What we see here is that Average Pass Rate (computed over all GCSEs taken) is the best predictor of NVQ attainment level, with a correlation of **0.44**. This value can be contrasted with that of **0.51**, using the multiple prediction equation computed using four GeneSys variables (in Hypothesis #2 analysis above).

Result: Due to the multicollinearity (variables correlating very highly with one another) within the 5 GCSE predictor variables, only one of them seems to be relevant for prediction purposes. This is the Average Pass Rate computed over all GCSEs taken by an individual (AVG_PASS). This alone correlates 0.44 with NVQ level attained.

Conclusion: GCSE average pass rate correlates 0.44 with NVQ level attained.

5. What are the relationships between GCSE indicator variables and GeneSys variables?

Analysis File: Uses ELTECFIN.sta

Given the similarity of the GeneSys results in Hypothesis #2 and those in Hypothesis #4, using the GCSE variables as predictors of NVQ level attained, it is of interest to determine whether a common "cause" is at work. This cause is assumed to be general intellectual ability. To this end, the relationships between our 5 GCSE indicator variables and the 15 available GeneSys variables were computed...

Table 15: The Relationships between GeneSys variables and 5 GCSE indicator variables

Correlations (eltecfin.sta)					
BASIC STATS					
Marked correlations are significant at $p < .05000$					
N=355 (Casewise deletion of missing data)					
Variable	GCSE_ENT	GCSE_F	GCSE_C	AVG_PASS	MAX_PASS
_GRT2_VR	.25	.31	.45	.44	.37
_GRT2_NR	.17	.23	.31	.34	.32
_GRT2_AR	.30	.35	.36	.37	.32
_OIP_VEN	.04	.08	.08	.12	.10
_OIP_PHL	-.02	-.02	-.05	-.02	-.00
_OIP_RAD	-.06	-.09	-.06	-.09	-.07
_OIP_GRE	.04	.06	.07	.10	.13
_OIP_ASS	.05	.07	.17	.18	.24
_OIP_PER	.05	.05	.15	.14	.20
_OIP_SCI	.03	.05	.16	.17	.20
_OIP_PRA	-.08	-.07	-.09	-.07	-.10
_OIP_ADM	.12	.12	.24	.21	.26
_OIP_NUR	-.04	-.02	.09	.11	.14
_OIP_ART	-.04	-.03	.10	.12	.17
_OIP_LOG	.19	.23	.27	.29	.26

What we see here is that the substantive correlations between these two kinds of variables are occurring with the Verbal and Abstract Reasoning ability variables in GeneSys, and GCSE_C and AVG_PASS GCSE variables (GCSE_C indexes the number of GCSE variables passed at level C or greater while AVG_PASS is the average pass computed over all GCSE's taken). The correlations between these key variables are not so high as to indicate excessive multicollinearity, so, it was decided to examine whether we could optimise the NVQ maximum attainment prediction by including both previously identified GeneSys and GCSE variables in the same prediction equation. Several regressions were tried, leading to the optimal solution below, using just two variables ...

Table 15: The Relationships between GeneSys variables and 5 GCSE indicator variables

Regression Summary for Dependent Variable: INDIV						
MULTIPLE REGRESS.	R= .58980151 R ² = .34786583 Adjusted R ² = .33233882 F(2,84)=22.404 p<.00000 Std.Error of estimate: .30610					
N=87	BETA	St. Err. of BETA	B	St. Err. of B	t(84)	p-level
Intercept			.839265	.170112	4.933613	.000004
_GRT2_AR	.426831	.094481	.033344	.007381	4.517612	.000020
AVG_PASS	.281147	.094481	.094135	.031635	2.975679	.003817

Here we see that using just the GeneSys Abstract Reasoning ability score and the Average GCSE pass rate, we can attain an *adjusted* multiple R of **0.58**. This is in contrast to the **0.51** achieved using four GeneSys variables, or **0.44** when just using the AVG_PASS variable alone.

Perhaps the best way to compare the results to one another is in terms of the classification accuracy. Table 7 above showed a classification accuracy of 72% using four GeneSys variables to classify NVQ attainment predictions .. it is repeated below...

Table 7: NVQ Predicted Maximum Attainment: Classification Table using optimal GeneSys variable weights (R=0.51)

Summary Frequency Table (tempregr.sta)			
BASIC STATS	Marked cells have counts > 10 (Marginal summaries are not marked)		
INDIV	Predicted Level 1	Predicted Level 2	Row Totals
Level 1	103	54	157
Level 2	41	156	197
Level 3	0	6	6
Column Totals	144	216	360

The corresponding table using our new 2–variable prediction equation is:

Table 16: NVQ Predicted Maximum Attainment: Classification Table GeneSys Abstract Reasoning and GCSE Average Pass rate scores (R=0.58)

Summary Frequency Table (tempreg2.sta)			
Continue...	Marked cells have counts > 10 (Marginal summaries are not marked)		
INDIV	Predicted Level 1	Predicted Level 2	Row Totals
Level 1	1	10	11
Level 2	1	73	74
Level 3	0	2	2
Column Totals	2	85	87

The data in Table 16 give us a classification accuracy of 85% compared with that of 72% using just the GeneSys variables alone. However, because of the substantive drop in sample size (from 360 in the GeneSys analysis to just 87 in this current analysis), we can see that the frequency distribution of the values of our criterion INDIV variable is somewhat distorted. That is, the vast majority (85%) of our cases achieve Level 2 NVQs. So, some caution must be applied to these results – they may not be replicable on a new sample of data.

Result: GeneSys ability variables are related substantively to the GCSE indicator variables GCSE_C (the number of GCSE passes of C and above attained by an individual) and AVG_PASS (the average pass rate computed across all GCSEs taken). The maximum correlation observed was 0.45. Using a joint prediction function composed of GRT2 Abstract Reasoning and AVG_PASS, a multiple R of 0.58 was attained, with classification accuracy of 85%. However, some caution was indicated in the interpretation of this result as the sample size was low (N=87) and the distribution of NVQs in this sample was distorted (the majority of cases with an attainment level of 2).

Conclusion: There is a substantive relationship between GCSE average pass rate and the number of GCSEs attained at grade C and above, and general reasoning ability.

6. Are starting SOC code areas associated with GeneSys Occupational Interest variable profiles?

Analysis File: Uses ELTECFIN.sta, ELTECSOC.sta, and ELTECSTD.sta

Here, a simple program was used to categorise SOC codes into 8 distinct categories (SOC codes category membership was supplied by ELTEC).

{SOC CODE Categorisation program}

```
{Construction}
if (STARTSOC = 521) or (STARTSOC = 570) then SOCATS := 1;

{Engineering and Motor Trade}
if (STARTSOC = 516) or (STARTSOC = 540) then SOCATS := 2;

{Engineering}
if (STARTSOC = 519) or (STARTSOC = 599) or (STARTSOC = 859) then SOCATS := 3;

{Business Admin}
if (STARTSOC = 410) or (STARTSOC = 430) or (STARTSOC = 459) or (STARTSOC = 490)
then SOCATS := 4;

{Sales and Warehouse}
if (STARTSOC = 441) or (STARTSOC = 720) then SOCATS := 5;

{Hospitality, catering, hotel etc.}
if (STARTSOC = 620) then SOCATS := 6;

{Hairdressing and Beauty}
if (STARTSOC = 660) then SOCATS := 7;

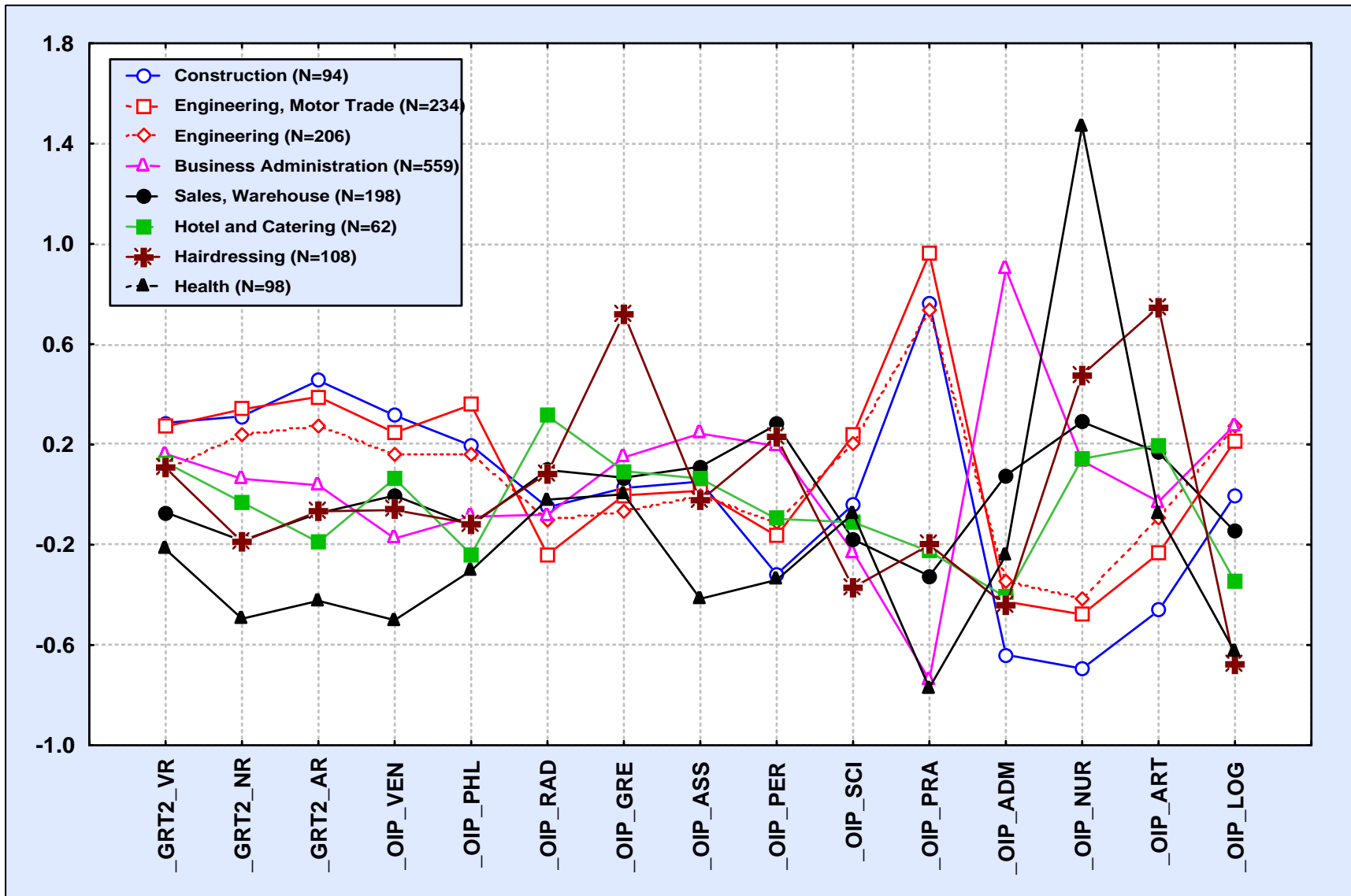
{Health Care, Nursing etc.}
if (STARTSOC = 644) then SOCATS := 8;
```

The resultant coding was set up in the variable SOCATS in the file ELTECSOC.sta. In order to compare each of the GeneSys variables with the SOC code groups, it was decided that a graphical profile comparison of mean scores for each GeneSys variable for each SOC code group would be an optimal approach. However, because some scales (ability at least) may differ in their maximum scores attainable, and hence display mean differences due solely to this fact, all GeneSys variables were standardized individually using their respective means and standard deviations. The file ELTECSTD.sta contains the standardized values. Means were then computed for each SOC group, and profiles created accordingly. Figure 1 below displays the SOC x GeneSys variable profile.

From Figure 1, it is clear that GeneSys variables are discriminating between certain Occupational Start SOC code categories. For example, there are clear distinctions between the Health category and say Engineering categories. This is a good example of the construct validity of the GeneSys scales.

Result and Conclusion: GeneSys variables are substantively associated with SOC categories.

Figure 1: The GeneSys Ability, Personality, and Interest Profile for each of 8 Occupational SOC code categories.



7. Is Success or Failure with NVQ training related to the overall level of Interest in Any Occupational categories?

Analysis File: Uses ELTECSTD.sta and ELTECINT.sta

Here, a hypothesis was tested concerning whether the overall LEVEL of interest is associated with success or failure to complete an NVQ qualification (using the definition of success and failure as in Hypothesis #1). Specifically, rather than look at each interest separately, it was decided to create a new variable (INTLEVEL) that would hold the count of the number of interests that exceeded 1 standard deviation discrepancy from the mean for each student. Using the standardised GeneSys interest variables ...

45	_OIP_PER	8.0	-9999	OIP - Persuasive
46	_OIP_SCI	8.0	-9999	OIP - Scientific
47	_OIP_PRA	8.0	-9999	OIP - Practical
48	_OIP_ADM	8.0	-9999	OIP - Administrative
49	_OIP_NUR	8.0	-9999	OIP - Nurturing
50	_OIP_ART	8.0	-9999	OIP - Artistic
51	_OIP_LOG	8.0	-9999	OIP - Logical

a small program calculated the appropriate score for the new variable. Then, this variable was cross-tabulated with the SUCCESS variable, and an appropriate statistical test made for the significance of any observed differences between the Successful and Fail groups. The results are given below ...

Table 17: Cross-Tabulation of Overall Occupational Interest level and Success or Failure to complete an NVQ qualification course.

BASIC STATS var: INTLEVEL	Marked cells have counts > 10 (Marginal summaries are not marked)		
	var: Success Fail	var: Success Success	Row Totals
0 interests >= 1sd	201	168	369
1 interest >= 1sd	181	145	326
2 interests >= 1sd	74	97	171
3 interests >= 1sd	42	43	85
4 interests >= 1sd	21	21	42
5 interests >= 1sd	7	7	14
6 interests >= 1sd	3	2	5
7 interests >= 1sd	2	1	3
Column Totals	531	484	1015

Given a Null hypothesis of "no-difference" between the observed cell frequencies for the columns "Fail" and "Success", the Chi-Square value for this table is 8.407, with 7 df, and $P = 0.298$. There is no statistically significant difference between the level of interests for an individual and whether they complete (with grade above level 0) or leave an NVQ course prematurely. The nominal measure of agreement between Interest level and the criterion variable SUCCESS is 0.09 (Cramer's V). An ordinal gamma coefficient (which used more information contained in the data) = 0.08. Both of course are conceptually as well as statistically not significant.

Result: Interest level, indexed by the count of the number of interests for a student that are greater than 1 standard deviation from each group mean interest, is not associated with success or failure to complete an NVQ qualification course.

Conclusion: There is no relationship between Interest level and success or failure to complete an NVQ qualification course.

Appendices

A.1 Statistica Fieldname Specification (ELTECFIN.sta)

No	Name	Format	MD Code	Long Label
1	SURNAME	8.0	-9999	Surname
2	FIRSTNAM	8.0	-9999	First Name
3	TITLE	8.0	-9999	Title
4	SEX	3.0	-9999	Gender
5	ENTRY_DA	DATE6	-9999	Genesys Entry Date
6	AGE	3.0	-9999	Age
7	REFERENC	DATE6	-9999	Date of Birth
8	USER_PIC	8.0	-9999	User_PIC
9	APPLICAN	8.0	-9999	Applicant
10	NVQ_LEVE	8.0	-9999	NVQ Level (all set to 0)
11	ORIGIN	8.0	-9999	Ethnic origin of individual
12	EDUCATIO	8.0	-9999	blank field
13	CAREER	8.0	-9999	blank field
14	ETHNICGR	8.0	-9999	Ethnic Group (coded 1-9)
15	STARTDAT	DATE6	-9999	Start Date of Training
16	MO	8.0	-9999	Month of Start Date
17	YR	8.0	-9999	Year of Start Date
18	SKEY	8.0	-9999	Sample Key - set if STARTDAT between Jan. 98 and
19	STRAND	6.0	-9999	Type of Training (Other, Modern Apprenticeship, o
20	EMPLOYED	8.0	-9999	Employed status
21	LITERACY	8.0	-9999	Requirement for literacy training
22	NUMERACY	8.0	-9999	Requirement for numeracy training
23	STN?	4.0	-9999	Special Training Need
24	STARTSOC	8.0	-9999	Start SOC code
25	ANTICIPA	8.0	-9999	Anticipated NVQ
26	FINISHDA	DATE6	-9999	Finish Date ... date when training finished or st
27	CURRENTS	8.0	-9999	Current SOC code
28	FINISHCO	8.0	-9999	Finish Code
29	COMPLETE	8.0	-9999	Completed individual training plan at the point o
30	STARTERC	8.0	-9999	Starter comments
31	LEAVERCO	8.0	-9999	Leaver Comments
32	_GRT2_TE	8.0	-9999	GRT2 Test Date
33	_GRT2_VR	8.0	-9999	Test Score - Verbal Reasoning
34	_GRT2_NR	8.0	-9999	Test Score - Numerical Reasoning
35	_GRT2_AR	8.0	-9999	Test Score - Abstract Reasoning
36	_GRT2_VR	8.0	-9999	Number Attempted - Verbal Reasoning
37	_GRT2_NR	8.0	-9999	Number Attempted - Numerical Reasoning
38	_GRT2_AR	8.0	-9999	Number Attempted - Abstract Reasoning
39	_OIP_TES	8.0	-9999	OIP Test Date
40	_OIP_VEN	8.0	-9999	OIP - Venturesome
41	_OIP_PHL	8.0	-9999	OIP - Phlegmatic
42	_OIP_RAD	8.0	-9999	OIP - Flexible
43	_OIP_GRE	8.0	-9999	OIP - Gregarious
44	_OIP_ASS	8.0	-9999	OIP - Assertive
45	_OIP_PER	8.0	-9999	OIP - Persuasive
46	_OIP_SCI	8.0	-9999	OIP - Scientific
47	_OIP_PRA	8.0	-9999	OIP - Practical
48	_OIP_ADM	8.0	-9999	OIP - Administrative
49	_OIP_NUR	8.0	-9999	OIP - Nurturing
50	_OIP_ART	8.0	-9999	OIP - Artistic
51	_OIP_LOG	8.0	-9999	OIP - Logical
52	_MRT2_TE	8.0	-9999	Mechanical Reasoning Test Date
53	_MRT2_MR	8.0	-9999	Test Score - Mechanical Reasoning
54	_MRT2_MR	8.0	-9999	Number Attempted - Mechanical Reasoning
55	_GCSE_TE	8.0	-9999	GCSE - Test Date

A.1 Statistica Fieldname Specification (ELTECFIN.sta) (cont.)

No	Name	Format	MD Code	Long Label
56	_GCSE_ET	8.0	-9999	GCSE - English Literature
57	_GCSE_EL	8.0	-9999	GCSE - English Language
58	_GCSE_MA	8.0	-9999	GCSE - Maths
59	_GCSE_SC	8.0	-9999	GCSE - Science
60	_GCSE_GO	8.0	-9999	GCSE - Geography
61	_GCSE_HI	8.0	-9999	GCSE - History
62	_GCSE_TK	8.0	-9999	GCSE - Technology
63	_GCSE_IT	8.0	-9999	GCSE - Information Technology
64	_GCSE_AR	8.0	-9999	GCSE - Art Design
65	_GCSE_TX	8.0	-9999	GCSE - Textiles
66	_GCSE_CR	8.0	-9999	GCSE - Craft Design
67	_GCSE_FO	8.0	-9999	GCSE - Food
68	_GCSE_RE	8.0	-9999	GCSE - Religious Education
69	_GCSE_FR	8.0	-9999	GCSE - French
70	_GCSE_GE	8.0	-9999	GCSE - German
71	_GCSE_SS	8.0	-9999	GCSE - Spanish
72	_GCSE_UR	8.0	-9999	GCSE - Urdu
73	_GCSE_CH	8.0	-9999	GCSE - Child Development
74	_GCSE_GS	8.0	-9999	GCSE - General Studies
75	_GCSE_BU	8.0	-9999	GCSE - Business Studies
76	_GCSE_GR	8.0	-9999	GCSE - Graphics
77	_GCSE_DR	8.0	-9999	GCSE - Drama
78	_GCSE_PE	8.0	-9999	GCSE - Physical Education
79	_GCSE_SO	8.0	-9999	GCSE - Sociology
80	_GCSE_ME	8.0	-9999	GCSE - Media Studies
81	_GCSE_MU	8.0	-9999	GCSE - Music
82	_GCSE_EK	8.0	-9999	GCSE - Electronics
83	_GCSE_EC	8.0	-9999	GCSE - Economics
84	_GCSE_HO	8.0	-9999	GCSE - Home Economy
85	_GCSE_SP	8.0	-9999	GCSE - Sports Studies
86	_GCSE_WO	8.0	-9999	GCSE - Word processing
87	_GCSE_CO	8.0	-9999	GCSE - Computing
88	GCSE_ENT	8.0	-9999	Number of GCSE's Entered
89	GCSEFAIL	8.0	-9999	Number of GCSE's Failed (Grade 1)
90	GCSE_F	8.0	-9999	Number of GCSEs with grade F and above
91	GCSE_C	8.0	-9999	Number of GCSEs with grade C and above
92	AVG_PASS	8.2	-9999	Average Pass Mark over all GCSEs taken
93	MAX_PASS	8.0	-9999	Maximum Passmark across all GCSEs taken
94	ID	8.0	-9999	Just a control validity field for the GCSEutil ST
95	QUAL1REF	8.0	-9999	Qualification Levels -1 - discrete codes
96	QUAL1LEV	8.0	-9999	Qualification Levels -1- actual NVQs achieved
97	QUAL1DAT	8.0	-9999	Qualification Levels -1- Dates
98	QUAL2REF	8.0	-9999	Qualification Levels -2 - discrete codes
99	QUAL2LEV	8.0	-9999	Qualification Levels -2- actual NVQs achieved
100	QUAL2DAT	8.0	-9999	Qualification Levels -2- Dates
101	QUAL3REF	8.0	-9999	Qualification Levels -3 - discrete codes
102	QUAL3LEV	8.0	-9999	Qualification Levels -3- actual NVQs achieved
103	QUAL3DAT	8.0	-9999	Qualification Levels -3- Dates
104	QUAL4REF	8.0	-9999	Qualification Levels -4 - discrete codes
105	QUAL4LEV	8.0	-9999	Qualification Levels -4- actual NVQs achieved
106	QUAL4DAT	8.0	-9999	Qualification Levels -4- Dates
107	JOINT123	8.0	-9999	Any 2 NVQ fields that contain a 1, 2, or 3
108	INDIV	8.0	-9999	The maximum level NVQ achieved by an individual
109	NODAT	8.0	-9999	No start date but an NVQ qualification level -
110	LEFTTRN	8.0	-9999	those who LEFT with no NVQ (not still in training)
111	NODAT1	8.0	-9999	No Start Date - but still flagged as STILL IN TRA
112	SUCCESS	8.0	-9999	0 = Left, Non-Achievers, 1 = NVQ Achievers Group

*ID(v.94), NODAT(v.109) and NODAT1(v.111) = validity check variables only