KM 14(1984)
pag 97 -109

INDIVIDUAL GROWTH PATTERNS IN EARLY READING PERFORMANCE

Margo G.H. Jansen
Adriana G. Bus

University of Groningen

Abstract

The present paper concerns the characterization of observed individual learning curves obtained in a small sample of first grade pupils. These data were analysed with the well known logistic regression model and it is shown how hypotheses concerning these data can be tested.

1. INTRODUCTION

In the present paper we discuss an application of bio-assay methods to data obtained in a social science context. The study we are referring to was concerned with the acquisition of early reading skills. Because individual development patterns were considered as to be among the main points of interest, a longitudinal design was adopted.

Relatively few individuals were observed regularly over a relatively long period of time, in order to get an accurate picture of the individual growth patterns. At each occasion the individuals were tested with several short tests especially designed to measure different aspects of reading performance. One of the crucial features of the study was that the tests consisted of items on which the answer could be "condensed" to two forms: correct or incorrect. Thus, each response could be represented by a binary random variable taking values of one or zero only, where a one corresponded to a correct answer (a "success") and a zero to an incorrect one. The basic problem is to find good methods of analysis for the dependence of the probabilities of a correct answer on the time variable and also for dependencies on other explanatory variables, such as groupings of the individuals.

Address: Dep. of Education
Westerhaven 16
9718 AW Groningen

2. DESIGN OF THE STUDY

The study mentioned in the introduction was set up as follows. A group of seven first grade children was tested at regular intervals (once a week) during the school-year. A total of forty-seven testing sessions was planned, but due to holidays, illness, etc., data were missing on several occasions. We excluded here one of the seven children because of an excessive amount of missing data. The remaining six could be subdivided into two groups according to two different reading programs which were used in their schools. The major purpose of the study was to assess individual growth patterns. Secondly, we were interested to see if these individual patterns showed consistent differences which could be related to characteristics of the reading programs, though the small sample sizes and the essentially quasi-experimental nature of the study made this difficult.

During each testing session the children were tested with four different short tests (two of 10 items and two of 15 items respectively) designed to measure different aspects of reading performance. For each of the four different types of tests, several (4 for the 15-item tests and 8 for the 10-item tests) nominally parallel versions were available. Although no formal scheme was used for assigning test versions to testing sessions, it is unlikely that pupils were tested twice with the same version in close succession.

We will refer to the four types of tests as P, Q, R and S. The tests differ in such variables as length, orthografic regularity and presence of a context. We will not go into the theoretical background and the detailed research questions of the study which are fully treated elsewhere (Bus, 1982). Instead we restrict ourselves to the analysis of the data without a substantive interpretation of the results.

As may be inferred from the information given above, the data material consisted of item responses collected at a relatively large number of time points for six individuals who could be grouped a priori in two subgroups. Furthermore, the item responses which were recorded could be coded as either incorrect or correct. In the research project itself other codings were also used, but these do not concern us here. Our analysis tries to fit curves describing the relation of the attainment of reading skills, as measured by the tests, with the time variable, assuming the model we will describe in the next section.

3. THE MODEL

The task of finding a suitable model for characterizing growth in attainment within individuals is not an easy one. Possible solutions could be sought in the application of item response theory, where a probabilistic relation between observed item response and an underlying so-called latent trait is assumed (Lord & Novick, 1968). An extra assumption, that is not included in item response theory but is needed here, would be that the latent trait value of a given individual increases over time. Estimated trait values at several time points could be used to display the information concerning the individual's growth in attainment over time.

For a similar situation, Bock proposes the use of an adaption of the so-called normal ogive model (Bock, 1976; Lord & Novick, 1968).

Even more attractive are the logistic latent trait models with linear constraints on the parameters (LLTM) proposed by Fischer (Fischer, 1974, 1977, 1982, 1983). The linear logistic model with relaxed assumptions, which can be viewed as a special case of an LLTM is especially suited for measuring change. This model allows the characterization of individuals in a multi-dimensional latent space and the testing of hypotheses regarding the effects of different treatments (groups).

Given the very small number of individuals and the large number of time-points the LLTM approach was not feasable in our study.

The model we finally chose was a so-called bio-assay model. Though developed in the context of biological and psychological experiments, there is a close connection between bio-assay models and item response theory (Lord & Novick, 1968, p. 420). A fairly typical example of the kind of experiment for which bio-assay models are used is the following. There is an independent variable, the "stimulus level", which is controlled by the experimentor. Each experimental subject is assigned to a certain level of the variable and a binary response is observed. In bio-assay experiments the independent variable is usually the log-dose of some poison and methods to fit these models to the data have long been available. See, for example, Finney (1952,1978). Other references are the monograph by Ashton (1972) on (binary) logistic regression analysis applied to bio-assay data and a book by Cox (1970) which is entirely devoted to the analysis of binary data. Several (more advanced) text books on mathematical statistics devote some pages to the topic of binary regression (Cox and Hinkley, 1974; Andersen, 1980, chapter 8).

Although situations formally similar to bio-assay experiments may arise in many other fields, for example psychophysics (Bock and Jones, 1968), the bio-assay methods are seldom applied outside the bio-assay field. An example is the study by Micko (1969), also described by Andersen (1980), where the independent variable is the intensity of an acoustic stimulus and the response the action/non-action of one subject on a number of trials.

To keep the notation as simple as possible we will at present ignore the fact that we have to formulate the model for a series of attainment curves, e.g. for several individuals and tests.

For a certain non-specified individual and a certain non-specified test, the statistical model can be formulated as follows (Cox, 1970, chapter 2; Andersen, 1980, chapter 8). Assume we have a single regressor variable X, in our case the time, which takes m distinct values. At each level of X observations are obtained. And we denote by Y_{gj} the g-th observation at level j, where Y_{gj} denote the response on item g. Y_{gj} is a binary variable taking values of one and zero only.

$$E(Y_{gj}|X_j) = P(Y_{gj} = 1|X_j) \tag{1}$$

Now , let n_j responses be observed at level j of the independent variable X, and let the number of correct answers be

$$Y_j = \sum_{g}^{n_j} {}_{=1} Y_{gj}$$

Then, assuming that the Y_{gj}'s are independent given X_j, it is easily seen that Y_j has a compound binomial distribution. This compound binomial distribution can be considered binomial with parameters (n_j, P_j), if one of the following sets of assumptions concerning the individual items of the test holds. First, it is obvious that the total number correct is binomially distributed if all items are of equal difficulty for the individual being tested. This is a rather stringent assumption, which is fortunately not necessary.

Alternately, we may assume that we have well-defined domain of items from which the items given at time point X_j are sampled. At time X_j the true proportion correct is equal to P_j. Then the observed number correct has a distribution which is binomial with parameter P_j (Lord & Novick, 1968;

p. 250 and p. 508 ff) which is then treated as a function of X_j. In addition we assume that randomly parallel versions of the test are given at the different time points.

For the actual form of the function, we have a number of commonly used options. One of the possibilities is the cumulative normal distribution

$$P_j = \Phi(\alpha+\beta X_j) \tag{2}$$

Analysis based on (2) is known in the literature as probit-analysis and has been thoroughly treated by Finney (1952, 1978). Another model originally suggested by Berkson in a series of papers, is the logistic model

$$P_j = \exp(\alpha+\beta X_j) \; / \; \left(1+\exp(\alpha+\beta X_j)\right) \tag{3}$$

Analysis based on (3) is called logit-analysis, based on the fact that we can, by applying the logit transformation, write model (3) as follows

$$\lambda_j = \ell n \frac{P_j}{1-P_j} = \alpha+\beta X_j \tag{4}$$

If P_j given by (3) is plotted as a function of X we obtain a symmetrical monotonic increasing curve, provided that β is positive. The curve is symmetric around the value of X_j at which the probability of a success is ½, namely $-\alpha/\beta$. The slope of the curve at this point is $\beta/4$, so the numerical value of β measures the steepness.

Although numerically models (2) and (3) agree closely, the logistic model has theoretical advantages which the normal model lacks. This is primarily due to the fact that model (3) belongs to the exponential family and thus sufficient statistics are available for the model parameters (Andersen, 1980; Cox, 1970). In general, maximum likelihood estimates for the logistic model parameters α and β may be obtained by solutions of

$$\sum_j Y_j = \sum_j n_j P_j \tag{5}$$

and

$$\sum_j X_j Y_j = \sum_j n_j X_j P_j \tag{6}$$

As (5) and (6) do not have explicit solutions, numerical methods have to be employed.

To indicate that we have a series of response curves instead of only one we introduce an extra index i, referring to the individuals. For a given test the probability of a correct answer to a randomly chosen item g at time point X_j for individual i is given by

$$P_{ij} = P(Y_{igj} = 1 \mid X_j) = \exp(\alpha_i+\beta_i X_j) \; / \; (1+\exp(\alpha_i+\beta_i X_j)) \quad . \quad (7)$$

The total number of correct answers at time X_j is denoted by $Y_{ij} = \sum_g Y_{igj}$.

Using the estimation methods mentioned leads to estimates of the α's and β's and their asymptotic standard errors (Cox, 1970).

To check the fit of the model, the simplest procedure is to find the fitted probabilities, $\hat{P}_{ij}$, and from them the fitted number of correct items $n_j\hat{P}_{ij}$, and compare the fitted number of correct items with the observed number correct, possibly by calculating standardised residuals. Another informative procedure is to compare likelihood ratio statistics achieved under different models obtained by imposing various restrictions on the parameters. A goodness of fit test for model (7) against the general model in which all P_{ij}'s are left free and are estimated by the observed proportions Y_{ij}/n_j is given by the following test statistic

$$Z_i = 2\left(\sum_{j=1}^{m} Y_{ij}(\ln Y_{ij} - \ln n_j\hat{P}_{ij}) + \sum_{j=1}^{m} (n_j-Y_{ij})(\ln(n_j-Y_{ij}) - \ln n_j(1-\hat{P}_{ij}))\right), \quad (8)$$

which is, if model (7) holds, asymptotically distributed as a chi-square with m-2 degrees of freedom (Anderson, 1980). An overall measure of fit is given by

$$Z = \sum_i Z_i, \text{ where } i = 1, \ldots, N \quad , \quad (9)$$

which is also asymptotically chi-square with degrees of freedom N(m-2) if model (7) holds for each individual. Various other hypotheses concerning the α_i's and β_i's can be tested with appropriate LR-tests.

Table 1: Parameter estimates based on the model where all α's and β's are allowed to be different: test P, test Q, test R and test S.

	Test P			Test Q			Test R			Test S		
param.	est.	st.err.	95% conf.int.	est.	st.err.	95% conf.int.	est.	st.err.	95% conf.int.	est.	st.err.	95% conf.int.
α_1	-7.4	1.07	(-9.5,-5.3)	-6.6	0.77	(-8.1,-5.1)	-4.1	0.35	(-4.8,-3.4)	-3.1	0.33	(-3.8,-2.5)
α_2	-4.7	0.55	(-5.8,-3.6)	-4.7	0.92	(-6.5,-2.9)	-5.1	0.48	(-6.0,-4.2)	-1.6	0.23	(-2.1,-1.2)
α_3	-5.1	0.59	(-6.3,-4.3)	-5.6	0.97	(-7.5,-3.7)	-4.6	0.41	(-5.4,-3.8)	-2.5	0.28	(-3.1,-2.0)
α_4	-1.0	0.51	(-3.4,-1.4)	-0.3	0.84	(-2.0,+1.4)	-1.4	0.22	(-1.8,-1.0)	-1.8	0.70	(-3.2,-0.4)
α_5	-3.2	0.53	(-5.6,-0.8)	-2.7	0.85	(-4.4,-1.0)	-4.3	0.37	(-5.0,-3.6)	-2.1	0.35	(-2.8,-1.4)
α_6	-2.4	0.38	(-4.7,-0.1)	-3.0	0.85	(-4.3,-1.3)	-3.6	0.32	(-4.2,-3.0)	-2.2	0.30	(-2.8,-1.6)
β_1	0.50	0.071	(.36,.64)	0.35	0.040	(.27,.43)	0.15	0.013	(.12,.18)	0.23	0.02	(.14,.27)
β_2	0.29	0.032	(.23,.35)	0.20	0.021	(.16,.24)	0.15	0.016	(.12,.18)	0.12	0.01	(.10,.14)
β_3	0.30	0.033	(.24,.36)	0.23	0.024	(.18,.28)	0.14	0.014	(.11,.17)	0.18	0.02	(.14,.22)
β_4	0.26	0.052	(.16,.36)	0.13	0.023	(.08,.18)	0.10	0.010	(.08,.12)	0.48	0.10	(.18,.68)
β_5	0.31	0.043	(.22,.40)	0.19	0.022	(.15,.23)	0.18	0.014	(.15,.21)	0.24	0.03	(.18,.30)
β_6	0.20	0.025	(.15,.25)	0.16	0.017	(.13,.20)	0.13	0.012	(.11,.15)	0.19	0.02	(.15,.23)
LR-test	z = 184.7	df = 188		z = 306.7	df = 188		z = 257.6	df = 188		z = 182.2	df = 188	

4. RESULTS

Based on (7), we can define several "submodels" by systematically placing restrictions on the α_i's and β_i's. The first model we tried to fit is the case where all individual α_i's and β_i's are allowed to be different. Maximum likelihood estimates of the parameters, assymptotic standard errors and the value of the overall test statistic Z per test are given in Table 1. (The computer program GLIM (Baker and Nelder, 1978) was used for all computations.) As can be seen from Table 1 the model fits reasonable well for test P en test S. The fit is markedly less good in the case of test Q (especially for the individuals 5 and 6). It is also clear that, although per test the parameter estimates differ across the individuals, the confidence intervals overlap considerably.

To study the extent of the difference between the person parameters, several more restricted models were fitted. A summary of the various models is given in Table 2, and the results of the estimation procedures in terms of fit measures can be found in Table 3. The second model assumed different individual intercepts and equal slopes within groups. Individuals 1, 2 and 3 are in the first, individuals 4, 5 and 6 in the second group. The grouping is based on teaching method.

Table 2: A summary of the tested models.

model	comment
model I : $Y_{ijk} = \alpha_i + \beta_i X_j$	all parameters different.
model II : $Y_{ijk} = \alpha_i + \beta_k X_j$	β different for the groups, α's different across individuals.
model III : $Y_{ijk} = \alpha_i + \beta X_j$	β the same for all individuals, α different.
model IV : $Y_{ijk} = \alpha_k + \beta_k X_j$	α, β different across groups.
model V : $Y_{ijk} = \alpha + \beta X_j$	all α's and β's the same.

$i = 1, .., n$ $\quad j = 1, .., m$ $\quad k = 1, 2$

Table 3: Measures of fit for the models of table 2.

		test P		test Q		test R		test S	
model	df	Z		Z		Z		Z	
I	188	184.7	16.7*	306.7	17.4*	257.6	23.2*	182.2	37.4*
	4								
II	192	201.4	8.9*	324.1	18.7*	280.8	1.6	219.6	3.8*
	1								
III	193	210.3		342.8		282.4		223.4	
IV	196	264.4		460.6		400.4		343.7	
V	198	431.1		654.1		574.5		475.4	

*significant difference (α=0.05)

Model II as well as model III (which assumed all slopes to be equal) show a statistically significant decrease in fit compared to model I, but the decrease is relatively small. The models with equal intercepts within groups (model IV) or even over all individuals (model V) however, show a pronounced decrease of fit. This phenomenon holds for all four tests. We are inclined to accept model III as the "best" descriptive model.

Note that we are reluctant to attach "exact" P-values to the observed likelihood-ratio statistics. This statistic is known to be distributed as chi-square only asymptotically and little is known about how good the approximation is for small sets of data. The interpretation of differences however, expressing the effect of adding terms to the model, seems less risky (Baker and Nelder, 1978). Another problem is that the null-hypotheses may be false, as in our study seems to be the case with test Q and test R. Also in that case, comparisons between the "best fitting logistic curves" for different models may still have some value.

The results in Table 1 also indicate that the tests differ from each other in a consistent way.

Another useful way to summarize the differences between the attainment curves is to estimate the so-called 50-percent point, the time-point at which 50 percent of the items are answered correctly, for each individual and test. This point is a function of both the intercept and the slope of the curve ($-\alpha/\beta$). The results are given in Table 4.

Table 4: The estimated points on the time-axis at which a performance level of 50 percent correct responses is reached.

		tests			
individuals		P	Q	R	S
	1	15	19	28	14
group 1	2	17	23	34	13
	3	17	24	33	14
	4	4	2	15	4
group 2	5	11	12	24	9
	6	12	19	27	12

Treating the estimated 50-percent points as observations on a derived dependent variable, an analysis of variance may be carried out, with 'group' as a between-subjects and 'test' as a within-subjects factor. Table 5 summarizes the results. Only the test effect was significant (alpha = 0.05). The group effect and the interaction between grouping and tests were not significant.

Table 5: A summary of an Analysis of Variance performed on the data in table 4.

Source	SS	df	MS	F	P
Between-subjects					
group	416.7	1	416.7	5.8	.07
subjects w groups	287.3	4	71.8		
Within-subjects					
test	908.8	3	302.9	80.2	.00
group by test	28.3	3	9.4	2.5	.11
test by subjects w group	45.3	12	3.8		

5. CONCLUSIONS

As can be inferred from the tables in the foregoing sections, the logistic model fits the observations reasonably well. In other words, the regression of reading performance, measured in logits, on time is nearly linear. Measured on the original scale, the regression of test performance on time can be approximated by a symmetric S-shaped curve. This holds in varying degrees for all four different tests and all individuals. We may also conclude that the individual curves differ, but that is hardly surprising. Furthermore, the differences between individual curves seem to be more a question of intercept than of slope. Comparing the tests, we note that slopes and intercepts differ. Test R, for example, seems consistently harder than the others. This is manifested in the 50-percent points, which are reached relatively late in the observation period. The relatively small slope parameters imply a slower growth rate.

A major advantage of the analysis chosen is that the individual curves can be characterized in a very simple manner by a limited number of parameters which are also easy to understand for persons without an extensive training in statistics.

The interpretation of the foregoing is hampered, however by a number of theoretical and statistical/methodological problems. Worth mentioning at this stage is the difficulty concerning the meaning of the time variable. In fact this independent variable, which is of course not controlled by the investigator, incorporates all kinds of variables influencing reading performance in addition to teaching, such as parental influence and maturation. It seems likely that the reading program has influenced the shape of the attainment curves, but with extremely small groups such as those employed in the study, and the absence of random assingment of individuals to groups, interpretations in this direction cannot be based on the statistical analysis alone.

Another problem is that it is difficult to infer anything about the extent of individual differences in a larger group. Having a reasonably large number of observations per individual one can assume that the individual curves are estimated rather precisely, but, because of the fact that the subjects were not chosen with a random selection procedure, it is virtually impossible to judge wether these pupils are typical for the group(s) they are chosen from or not.

Background information provided by their teachers indicated that the individuals were "ordinary" pupils with the exception of one who seemed brighter than this fellows.

Last but not least, in the derivation of the estimates and of the statistical tests, it was assumed that the item responses of an individual were conditionally independent. Although violations of this assumption do not affect the estimates of the parameters in a serious way, the estimated standard errors may be suspect.

Although it is possible to deal with correlated errors, these methods are more complicated and demand a relatively large number of observations for reasonable results (Visser, 1982). Moreover the results are more difficult to understand.

REFERENCES

Andersen, E.B., Discrete Statistical Models with Social Science Applications. North-Holland, Amsterdam, 1980.

Ashton, W.D., The Logit Transformation Griffin & Co., London, 1972.

Baker, R.J. and Nelder, The GLIM system: release 3, NAG, Oxford, 1978.

Berkson, J., Application of the Logistic Function to Bio-Assay, 1944, Journ. Amer. Stat. Assoc. 39, 357-365

Bock, R.D. Basic Issues in the Measurement of Change.
In: D.N.M. de Gruijer, L.J.Th. van der Kamp (Eds.), Advances in Psychological and Educational Measurement.
John Wiley and Sons, London, 1976.

Bock, R.D. & Jones, L.V., The Measurement and Prediction of Judgement and Choice.

Bus, A.G., A. longitudinal Study in Learning to Read. Paper presented at the 9th Conference of the International Reading Association. Dublin, 1982.

Cox, D.R., The Analysis of Binary Data. Chapman & Hall, London, 1970.

Cox, D.R. and Hinkley, Theoretical Statistics. Chapman & Hall, London, 1974.

Finney, D.J., Probit Analysis. Cambridge Un. Press, Cambridge, 1952.

Finney, D.J., Statistical Methods in Biological Assay. Griffin & Co., London, 1978, 3rd ed.

Fischer, G.H., Linear Logistic Test Models: Theory and Application.
In: H. Spada & B. Kampff (Eds.). Structural Models of Thinking and Learning. Huber, Bern, 1977.

Fischer, G.H., Logistic Latent Trait Models with Linear Constraints. Psychometrika, 1983, 48, 3-26.

Fischer, G.H. and Formann, A.H., Some Applications of Logistic Latent Trait Models with Linear Constraints on the Parameters. Applied Psychological Measurement, 1982, 6, 397-415.

Goldstein, H., The Design and Analysis of Longitudinal Studies. Academic Press, London, 1979.

Lord, F.M. and Novick, M.R., Statistical Theories of Mental Test Scores. Addison Wesley, Reading, 1968.

Lord, F.M. Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Associates, Hillsdale, 1980.

Linden, W.J. van der, Binomial Test Models and Item Difficulty. Applied Psychological Measurement, 1979, 3, 1-411.

Micko, H.C., A Psychological Scale for Reaction Time Measurement. Acta Psychologica, 1969, 30, 324-335.

Visser, R.A., On Quantitative Longitudinal Data in Psychological Research. Doct. Diss., Leyden, 1982.

Ontvangen: 10-10-1983