A distribution free approach for comparing growth of knowledge

Tan, E.S.; Imbos, TJ.; Does, R.J.M.M.

Publication date
1994

Published in
Journal of Educational Measurement

Journal of Educational Measurement
Spring 1994, Vol. 31, No. 1, pp. 51–65

A Distribution-Free Approach for Comparing Growth of Knowledge

E. S. Tan
Tj. Imbos
University of Limburg
and
R. J. M. M. Does
University of Amsterdam/Centre for Quantitative Methods, Eindhoven

The longitudinal testing of student achievement requires the solution of several new problem areas. In this article, several small groups of medical students at the University of Limburg Medical School in Maastricht, The Netherlands, are compared with respect to their performances. The results indicate, that, despite the possession of more knowledge at entrance, students with a low rate of growth of knowledge in the first year demonstrate a lower level of knowledge after the second academic year and continue to do so throughout the academic program when compared to students who show a higher rate of growth of knowledge in the first year. The analysis has been carried out using a distribution-free version of a longitudinal IRT-model suggested by Albers, Does, Imbos, and Janssen (1989). Furthermore, growth of knowledge has been described by means of a general regression model. Statistical inferences are possible using a randomization design extended to the situation where the observations are time-dependent proportions of correct answers.

More than a decade ago, Goulet (1975) pleaded for the use of longitudinal and time-lag designs in educational research. Since then, there seems to be an increasing interest in these methods among educational researchers (Albers, Does, Imbos, & Janssen, 1989; Willoughby & Hutcheson, 1978). For example, at the University of Missouri Medical School, Kansas City, a within-school longitudinal method for achievement testing was developed (Willoughby & Hutcheson, 1978). Somewhat later, a similar method was independently developed at the University of Limburg Medical School, Maastricht, The Netherlands. At both schools, the growth of knowledge during a 6-year medical program was measured by within-school tests, called *progress tests* in Maastricht and *quarterly profile examinations* in Kansas City. We shall, from here on, adhere to the Maastricht label for these tests.

The progress test can be conceived as a repeated final examination. It contains many test items (about 250 to 300) randomly sampled from the entire cognitive domain of medicine. This sample is drawn from an item bank of about 15,000 test items. This "final" is given 4 times a year to all students. Participation is obligatory for all 6-year groups of students. A more detailed discussion

The authors thank the editor and the two anonymous reviewers for their valuable suggestions leading to this substantially improved version of the article.

of the motivation for longitudinal achievement testing is described elsewhere (Tan, Imbos, Does, & Theunissen, 1992). Progress tests provide schools with a strong evaluation tool (van der Linden, 1987) that allows for cohort-longitudinal as well as time-lag educational research designs (Goulet, 1975). The longitudinal testing enables us to evaluate the students' growth of knowledge. In general, we can compare growth of knowledge of different-year group cohorts of students or students with different social economic background, sex, and so forth.

In this article, we study the relationship between first-year results from some subgroup of students and achievement during the course of the study and/or at the end of medical school. This problem has been analyzed using an IRT-model, motivated by Rasch (1960), for the longitudinal measurement of change with stochastic parameters that has been developed by Albers et al. (1989). In the method section, we summarize the statistical model and estimation procedure. We have adapted this progress (test) model for the comparison of growth curves observed in an educational context. Moreover, we have relaxed the distributional assumption of normality by using a method that was originally developed by Zerbe (1979) in a biomedical context. The randomization design used by Zerbe has been extended to the situation where the observations are time-dependent, proportion-correct answers, called *observed ability curves.* The method can be applied to the analysis of differences in item difficulties as well as to the comparison of different observed ability curves. The latter procedure has been chosen for the evaluation of progress of different groups of students. The model has been used for the analysis of the 1978 and the 1982 cohort of medical students. In the results section, we present the results of our analysis, which are discussed in the last section.

Method

Specification of the Progress Model

We start this section with the description of the progress model as suggested by Albers et al. (1989). In the next subsection, this model has been generalized by relaxing the somewhat rigid normality assumption. Motivated by the methodology of Rasch (1960) to separate the individual's level of knowledge from item characteristics, Albers et al. (1989) have proposed an IRT-model that consists of two parts. The first part relates the probability of a correct answer to the level of knowledge and item difficulty according to the following IRT-model:

$$P(X_{ijk} = 1 \mid s_{ik}, t_{jk}) = \Phi\,(s_{ik} - t_{jk}), \tag{1}$$

where X_{ijk} is a dichotomous (true/false) item score of student $i = 1, \ldots, m$ measured at timepoint $k = 1, \ldots p$ for item $j = 1, \ldots, n$; s_{ik} is the ability parameter of person i at timepoint k, and t_{jk} is the difficulty of item j of a test given at timepoint k. Φ is the standard normal distribution function. The second part of the model describes rate of growth of knowledge according to an

ordinary linear regression model.

$$s_{ik} = a_i + b_i * \frac{k}{p} + z_{ik}, \tag{2}$$

where z_{ik} is the within-individual error term, a_i represents the level of knowledge of student i at the time of entrance to the medical school, and b_i represents the rate of growth of that student. Albers et al. (1989) have chosen this model, because the data they analyzed can very well be described by means of a simple linear regression model.

Note that the progress model specifies growth of knowledge for each student who participates in several progress tests. In general, the students need not participate in all tests administered, nor need the tests be of equal length (Tan, Imbos, et al., 1992). Moreover, the items may vary in difficulty over time. However, Albers et al. (1989) have imposed an additional restriction regarding the distribution of the item difficulties. Due to the fact that progress tests consist of items randomly sampled from the same item bank, all test items do have the same underlying item-difficulty distribution. Albers et al. (1989) assumed normally distributed item difficulties. Furthermore, they have proposed a closed form ability estimator which can be expressed unconditionally on the timepoints. This is a very useful property, because it provides estimates of variability in "true knowledge" within each person that would not be possible otherwise. By integrating with respect to the distribution of the item difficulties, we obtain the following marginal probability

$$P(X_{ijk} = 1 \mid s_{ik}) = \Phi\left(\frac{s_{ik}}{\sqrt{1 + \tau^2}}\right). \tag{3}$$

Hence, s_{ik} can be estimated by

$$\hat{s}_{ik} = \sqrt{(1 + \tau^2)} * \Phi^{-1}(\overline{X}_{ik}), \tag{4}$$

where $\overline{X}_{ik}$ is the proportion correct for student i at timepoint k and τ^2 is the variance of the item difficulties.

Although the progress model works perfectly well as a description of knowledge growth of one large sample cohort of students (Albers et al., 1989), this model cannot directly be applied to describe knowledge growth or to compare between groups of highly selected students. First of all, Albers et al. (1989) assumed the ability parameters to have an underlying normal distribution. However, the ability parameters of a group of highly selected students are unlikely to be normally distributed. Furthermore, the growth of knowledge does not need to be linear as specified by Equation 2. It may level off at the end of the medical school, or the rate of growth may accelerate at the beginning of the medical school, and so forth. Finally, an even larger problem is the small sample sizes in this study. We analyze a small group (3 to 10 individuals) of highly selected students, based on their performance in the first year. The following adaptation of the progress model is needed to deal, among others, with the above-mentioned problem.

A Distribution-Free Item-Response Model

The statistical model that we propose can be considered as a generalization of the progress model developed by Albers et al. (1989). First, we relax the assumption of normality of the first part of the progress model (Eq. 1). Second, Equation 2 is replaced by a general regression model, not necessarily linear. For the purpose of our study, consider the following general IRT-model (compare with Equation 1):

$$P(X_{ijkg} = 1 \mid s_{ikg}, t_{jkg}) = H(s_{ikg} - t_{jkg}) \tag{5}$$

where the subscript g denotes the different groups. H is, unlike the normal item characteristic function of Equation 1, an arbitrary item characteristic function, strictly increasing as a function of the person-related characteristics s_{ikg}. The monotone nature of the function H is a common assumption in IRT-models (Andersen, 1983). Note that the proportion correct $\overline{X}_{i.kg}(= 1/n\Sigma_{j=1}^{n} X_{ijkg})$ of subject i can still be any curve, not necessarily monotone. It might even be U-shaped. Note further that the students do not need to participate in all tests administered, nor do the test lengths need to be of equal length (Tan, Imbos, et al., 1992). Nevertheless, for ease of presentation, the specification of the model will be given for the case of equal test length, with all students participating on all test administrations. With the above specification, there is only one person-related parameter (ability) for each student and one item-related parameter (difficulty) for each item.

Furthermore, Albers et al. (1989) assumed that the underlying normal distributions of the person and item parameters, respectively, are independent (cf. Andersen, 1980). The independency seems a reasonable assumption in our educational setting, where each year a new class of students is drawn from those seeking entrance to medical school and the items of each test are drawn randomly from an item bank. However, we have relaxed the normality assumption (and retain the independence assumption) to make the model more flexible and appropriate to analyze highly selected students. With respect to the second part of the progress model (cf. Eq. 2), it appears that the expectation of the proportion correct can be expressed conditionally only in s_{ikg} and is

$$E(\overline{X}_{i.kg} \mid s_{ikg}) =: \Psi_F(s_{ikg}), \tag{6}$$

which in general still depends on the distributional form F of the item difficulties. For example, if F is the normal distribution, then the function Ψ still depends on the mean and variance (cf. Eq. 3). However, all groups of students to be compared participate in tests that consist of items randomly drawn from the same item bank. Hence, if the comparison is restricted to students who participate on tests coming from the same item bank, then the dependence on the distribution F for all groups is the same. This fact makes the subscript F irrelevant for comparison purposes. Consequently, this subscript will be dropped in the following text. Note, that by using Equation 6, the item difficulties can be considered as nuisance parameters. In particular, we need not estimate any of the item parameters. Within the one-dimensional IRT-

models, the proportion correct is apparently related to ability s_{ikg} according to a general regression of proportion correct on ability (Eq. 6).

Finally, we assume that the function $\Psi(s_{ikg})$—that is, the expected proportion correct—can be decomposed into a sum of three independent components according to the following additive model:

$$\Psi(s_{ikg}) = M(k) + D_g(k) + U_{ig}(k), \tag{7}$$

where $M(k)$ is a general mean ability curve; $D_g(k)$ is the "deviant" curve, specifying the amount of deviation around the general mean curve with mean zero due to differences between person-related characteristics from different groups; and $U_{ig}(k)$ is the deviant curve around the general mean with mean zero due to differences in person-related characteristics and irrespective of the differences between groups. Model 7 is one of the simplest models we can consider because no subject-versus-group interaction has been assumed. Despite the similarity of the common analysis of variance with repeated measures, note that no restrictions concerning the population covariance structure (compound symmetry) are imposed. Note further that, if the person-related characteristics between different groups are equally distributed, the amount of deviation around the general mean curve is zero for all groups. In this case, Model 7 reduces to

$$\Psi(s_{ikg}) = M(k) + U_{ig}(k). \tag{8}$$

Or, equivalently $D_1(k) = \ldots = D_G(k) = 0$ for all timepoints k.

In summary, we have specified the expected proportion correct by means of a distribution-free analysis of variance with repeated measures. Under conditions common in item response theory, this expected proportion correct is a monotonic transformation of the abilities. In fact, Equation 7 specifies the abilities as a function of an analysis of variance model. However, our current interest is not the expression of the expected proportion correct in terms of deviations from the mean $M(k)$. Instead, we want to make inferences about the expected proportion correct based on the observed proportion correct. Statistical inferences can be made feasible, if the distribution of the error terms (discrepancies between observed and expected proportion correct) can be assessed. In the following paragraph, we propose a statistical method for the estimation of the exact distribution of the error terms.

A Stochastic Model Based on a Randomization Design

In an experimental setting where subjects are randomly assigned into different groups, a randomization model could be considered rather than models based on some theoretical distribution. A randomization model basically takes into account the randomization mechanism regarding the assignment. No further assumptions are needed regarding the distribution of errors in the observations. Many textbooks in statistics (e.g., Edgington, 1987; Scheffé, 1959) give excellent outlines on this topic. Based on these models, inferences can be made regarding group differences. On the other hand, if no randomization mechanism exists, inferences can still be made based on randomization

models. Note that, with respect to the applications that we have in mind, the students are actually assigned to different groups according to their performances in the first year. However, this lack of a randomization design does not appear to be a major problem, because our main interest is the comparison of ability distributions between groups. Because statistical testing procedures are carried out in general under the null hypothesis of no group differences, we only need to estimate the ability distribution in the case that the null hypothesis is true. The exact ability distribution under the null hypothesis of no group differences equals the distribution of the error terms based on a randomized design. Therefore, evaluation of group differences in ability can be made by using the probability structure induced by the randomization process. Hence, the following additive statistical model can be considered (cf. Eq. 7):

$$\overline{X}_{i.kg} = \Psi(s_{i.kg}) + R_{ig}(k) = M(k) + D_g(k) + E_{ig}(k), \tag{9}$$

where $R_{ig}(k)$ is the error curve with mean zero and $E_{ig}(k)$ is a combination of the error curve and a deviant curve due to differences in ability. In other words, Model 9 describes the mean ability curve $M(k)$ and the deviations $D_g(k)$ of curves for abilities from different groups relative to the mean ability curve. The parameter $E_{ig}(k)$ in the equation is a random disturbance term, since only group differences are of interest.

The randomization process determines a probability structure on $E_{ig}(k)$ which is, under the null hypothesis of equally distributed group-specific abilities, exactly the same as that proposed by Zerbe (1979). In conformance with that article, we propose a test statistic for comparing groups based on the within- and between-groups sums of squares. The test statistic has an interpretation similar to the commonly used F statistic in an ANOVA with one-way classification (see Scheffé, 1959). If $E_{ig}(k)$ is normally distributed, the distribution of the F statistic is known. The p values can be obtained by comparing the observed F statistic with the F distribution function with the proper degrees of freedom. However, the error term $E_{ig}(k)$ is not assumed to be normally distributed. In fact, the distribution of this variable is unknown. Hence, the distribution of the statistic F is unknown. However, by permuting the observed data between the groups and calculating the new value of the test statistic F after each possible permutation, the distribution of F is then empirically determined, and the exact p values can then easily be obtained by calculating the percentage of test statistic values exceeding the observed test statistic value (see Zerbe, 1979, for a detailed description).

In summary, it is possible to test differences between groups following the methodology suggested by Zerbe (1979). In particular, the test statistic F described above is appropriate for testing equality in distribution of between-group abilities. In the following, we have given a brief description of an interactive FORTRAN program suitable as a computer tool for a distribution-free analysis of variance.

Description of MUCRA

MUCRA is an abbreviation of Multiple Comparison Randomization Analysis. In general, the program can perform a distribution-free analysis of variance with repeated measures of a completely randomized, or randomized-blocks, design extended to growth and response curves (cf. Tan, Roos, Volovics, van Baak, & Does, 1992). It tests equality between several groups at selected time intervals. In the present situation, growth of several cohorts of students can be studied even if the measurements between cohorts have been carried out at different preselected timepoints. To tackle the problem of multiple comparison regarding the Type I error rate, MUCRA provides a single-step Scheffé type procedure (cf. Zerbe & Murphy, 1986) as well as a step-down procedure (so-called Peritz's closed step-down) that can be considered as the most powerful method among step-down procedures (cf. Hochberg & Tamhane, 1987). In the result section, we demonstrate the use of both procedures.

Data Description

The model has been applied to groups of students who entered the medical school in 1978 and 1982. The detailed description of the 1978 cohort has been given elsewhere (Albers et al., 1989). In 1978, 71 students and, in 1982, 148 students entered the medical school—only 44 students of the 1978 cohort and 71 students of the 1982 cohort participated in all tests administered. Furthermore, no results of the first and the last test in the fifth year of the 1978 cohort (Albers et al., 1989) exist due to administration errors. Due to the same administration errors, there were also missing data in the corresponding first year of the 1982 cohort (see Tables 1 and 2). Finally, missing data also occur in the last test of the third year of the 1982 cohort. For each cohort, two groups were formed based on the students' performance in the first year. For each student, the slope of the linear trend of the proportion correct based on the first four measurements was determined. Students in the lowest and highest 10% of the distribution of the slopes were labeled as slow and fast groups, respectively.

Note that there was no methodological rationale to classify the lowest 10% as the slow group and not, for instance, the lowest 15% or 25%. Our purpose was to compare highly selected groups who substantially differ from one another with respect to their performance in the first year. Furthermore, given the one-dimensional one-parameter IRT-model, the small samples (3 to 10 individuals in each group) do not invalidate inferences based on the proposed model. In fact, Zerbe has developed the randomization model in order to be able to compare small sample groups.

The Tables 1 and 2 show the mean proportion correct and standard error of the mean. All seven students of the 1978 cohort finished the academic program within 6 years, whereas additional information regarding study delay in the 1982 cohort is needed. This cohort has been classified further into a subgroup of students who finished the academic program within 6 years ($j6$) and a subgroup who needed more than 6 years ($j > 6$). With respect to this last

TABLE 1

Descriptive Statistics of the Average Proportion Correct Answers
Cohort 1978

Sample Size	Growth	Acad. Test Number	Mean	Stand. Error
		1	.21	.05
		2	.24	.07
		3	.18	.05
		4	.17	.03
		5	.17	.03
		6	.31	.04
		7	.30	.03
		8	.36	.04
		9	.36	.03
		10	.39	.03
		11	.38	.03
3	Slow	12	.44	.02
		13	.39	.03
		14	.43	.02
		15	.45	.02
		16	.40	.03
		17	---	---
		18	.46	.04
		19	.42	.02
		20	---	---
		21	.43	.04
		22	.49	.05
		23	.53	.05
		24	.55	.06

-continued on the next page-

group, it appears that all three students in the fast group show a study delay in the sixth year only. All three have participated in one extra test. However, all six students in the slow group experienced study delay varying from in the first year to in the last year. Those who fail the first-year examinations are obligated to redo the first year. The same consequence applies to the fourth year (MD) and the sixth year (basic physician). During other years, the student has the option to redo the same program at his own request. In all these cases, we have only used the scores of the last (series of) tests taken successfully. At the time

Table 1, continued

Sample Size	Growth	Acad. Test Number	Mean	Stand. Error
		1	.03	.01
		2	.13	.03
		3	.17	.02
		4	.21	.03
		5	.21	.03
		6	.23	.03
		7	.25	.03
		8	.36	.05
		9	.33	.03
		10	.40	.02
		11	.46	.04
4	Fast	12	.56	.02
		13	.48	.03
		14	.50	.03
		15	.52	.02
		16	.54	.02
		17	---	---
		18	.58	.04
		19	.59	.01
		20	---	---
		21	.60	.02
		22	.63	.03
		23	.61	.01
		24	.67	.04

that these data were assembled for research purposes, these six students had not yet graduated. Hence, the scores of the last two tests are missing.

Results

The data has been analyzed with the aid of the computer program MUCRA. As mentioned before, the probability structure of the error term $E_{tg}(k)$ is exactly the same as proposed by Zerbe (1979). The Figures 1 and 2 show the mean observed-ability curve (proportion correct) of the slow (S) and the fast (F) group within each cohort, respectively. An exploration of these figures suggests that students with slow knowledge development in the first year ultimately show deficient performance as compared to those students with fast knowledge development in the first year. Furthermore, in both cohorts, we

TABLE 2

Descriptive Statistics of the Average Proportion Correct Answers
Cohort 1982

Sample Size j6*	Growth j6	Acad. Test Number j6	Mean j6	Stand. Error j6	Sample Size j>6*	Growth j>6	Acad. Test Number j>6	Mean j>6	Stand. Error j>6
		1	—	—			1	—	—
		2	.25	.04			2	.15	.02
		3	.24	.07			3	.15	.02
		4	—	—			4	—	—
		5	.16	.01			5	.18	.02
		6	.17	.02			6	.16	.03
		7	.24	.05			7	.23	.03
		8	.30	.02			8	.27	.04
		9	.23	.02			9	.26	.04
		10	.29	.03			10	.31	.04
		11	.29	.04			11	.35	.03
4	Slow	12	—	—	6	Slow	12	—	—
		13	.30	.03			13	.33	.03
		14	.39	.03			14	.44	.04
		15	.41	.04			15	.43	.03
		16	.43	.03			16	.43	.02
		17	.44	.04			17	.39	.04
		18	.47	.03			18	.45	.05
		19	.47	.06			19	.48	.05
		20	.50	.03			20	.48	.05
		21	.47	.03			21	.50	.05
		22	.55	.06			22	.55	.05
		23	.53	.06			23	—	—
		24	.53	.05			24	—	—

-continued on the next page-

found that students of the S group have a higher level of knowledge at entrance than the students of the F group (see also Tables 1 and 2). The above statement can be tested properly using the randomization procedure described in the second section. All test items were drawn randomly from a large item bank. Hence the item difficulties can be considered independent of the abilities. As noted before, no group versus subject interaction is assumed. As can be seen from the (more or less) constant standard errors (Tables 1 and 2), the students in the different subgroups show a rather homogeneous growth of proportion correct, which is indicative of the absence of interaction. Within the framework of a one-dimensional IRT-model (Equation 5), we have argued that inferences are possible using the methodology suggested by Zerbe (1979). No other assumptions are needed.

An overall randomization test for testing equality of the two curves of the 1978 cohort yields a p value of .006, based on the observed F value and empirically determined distribution of F (Zerbe, 1979). This p value has been calculated with the aid of the program MUCRA, using the procedure as

Table 2, continued

Sample Size j6*	Growth j6	Acad. Test Number j6	Mean j6	Stand. Error j6	Sample Size j>6*	Growth j>6	Acad. Test Number j>6	Mean j>6	Stand. Error j>6
		1	--	--			1	---	---
		2	.08	.01			2	.09	.03
		3	.13	.02			3	.11	.01
		4	--	--			4	---	---
		5	.21	.02			5	.22	.01
		6	.28	.03			6	.39	.01
		7	.31	.03			7	.36	.05
		8	.36	.03			8	.38	.1
		9	.30	.02			9	.31	.07
		10	.37	.03			10	.39	.07
		11	.40	.02			11	.40	.04
10	Fast	12	---	---	3	Fast	12	---	---
		13	.38	.03			13	.40	.05
		14	.47	.03			14	.45	.07
		15	.47	.02			15	.44	.05
		16	.50	.03			16	.48	.05
		17	.49	.03			17	.48	.05
		18	.54	.02			18	.49	.04
		19	.55	.03			19	.51	.06
		20	.56	.02			20	.53	.05
		21	.55	.02			21	.49	.04
		22	.57	.02			22	.52	.04
		23	.60	.02			23	---	---
		24	.63	.02			24	---	---

* j6: students who finished the academic program within 6 years
j>6: students who needed more than 6 years

described in the second section (part 3). Furthermore, MUCRA can test equality between two or more groups at selected time intervals. Hence, on closer inspection, we found a p value of .203 for testing equality of the curves based on the first four measurements of the first year, a p value of .073 for testing equality of the curves based on the last four measurements in the fourth year, and a p value of .037 for testing equality of the curves based on the last four measurements in the sixth year. The 1982 cohort yields slightly different results. The comparison is made irrespective of how long the students needed to finish the academic program (p values of .001 as an overall result, p value of .737 for the first four measurements, a p value of .004 for the comparison until the fourth year, and .360 for the last four measurements of the sixth year). Note that the above-mentioned p values are the adjusted p values following the Scheffé type procedure. Using this method, MUCRA controls the overall Type I error rate (Tan, Roos, et al., 1992). Although the Scheffé type procedure is rather conservative (i.e., overestimates the true probability of a Type I error), and despite the small samples, the results indicate significant differences at, say 5% significance level within those periods as already suggested through the exploration of the figures. However, because the performance of students with

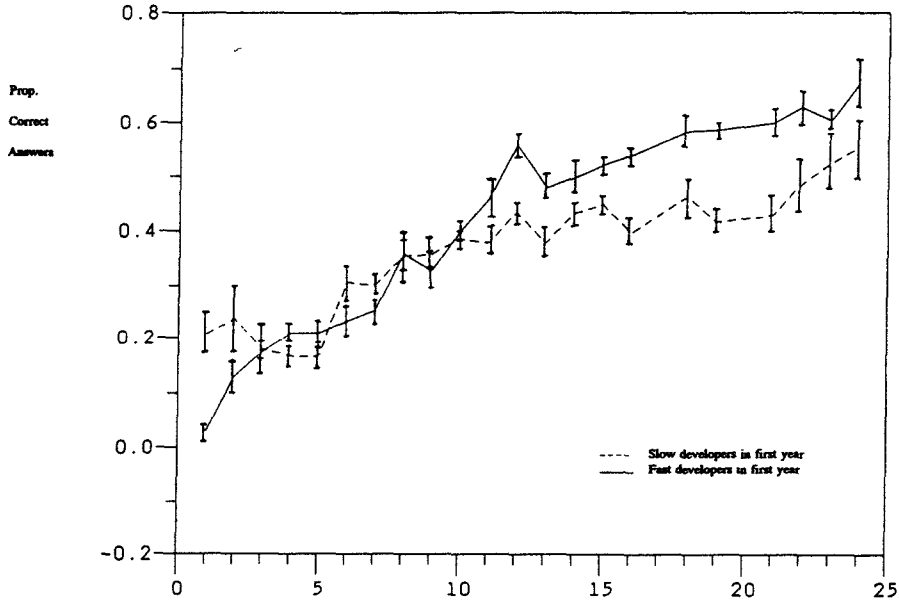


FIGURE 1. *Cohort 1978: Mean and standard error proportion correct*

no study delay can differ from those with study delay, we have analyzed all four groups of the 1982 cohort. Figure 3 shows the four groups of mean observed-ability curves. The results of a detailed analysis of the 1982 cohort are presented in Scheme 1, which is the result of a Peritz's closed step-down procedure. It starts with the most general hypothesis not implied by any other hypothesis and continues to less general hypotheses. If a particular hypothesis is not rejected, the other hypotheses implied by this particular hypothesis are then automatically not rejected without further testing. As can be seen from Scheme 1, no differences have been found between slow (fast) "growers" who finished their academic program within 6 years and those who needed more than 6 years (p value = .36 for testing H_0: $\mu_{S6} = \mu_{S>6}$, and p value = .33 for testing H_0: $\mu_{F6} = \mu_{F>6}$). Furthermore, the fast growers, irrespective of how long they needed to finish the academic program, differ significantly from slow growers, who do finish the academic program within 6 years. Finally, slow growers, who need more than 6 years to finish their academic program, differ significantly from fast growers, who do finish the academic program within 6 years (p value = .06 for testing H_0: $\mu_{S>6} = \mu_{F6}$; significant at 10% level, using the Peritz step-down procedure).

Summary and Discussion

The suggested method of analysis can be used to compare groups of students participating generally on different tests. This has been made feasible due to the constant distribution of the item difficulties. Moreover, growth of knowledge has been modeled using a general regression model. Because all results stated above hold whatever the underlying class of the one-dimensional

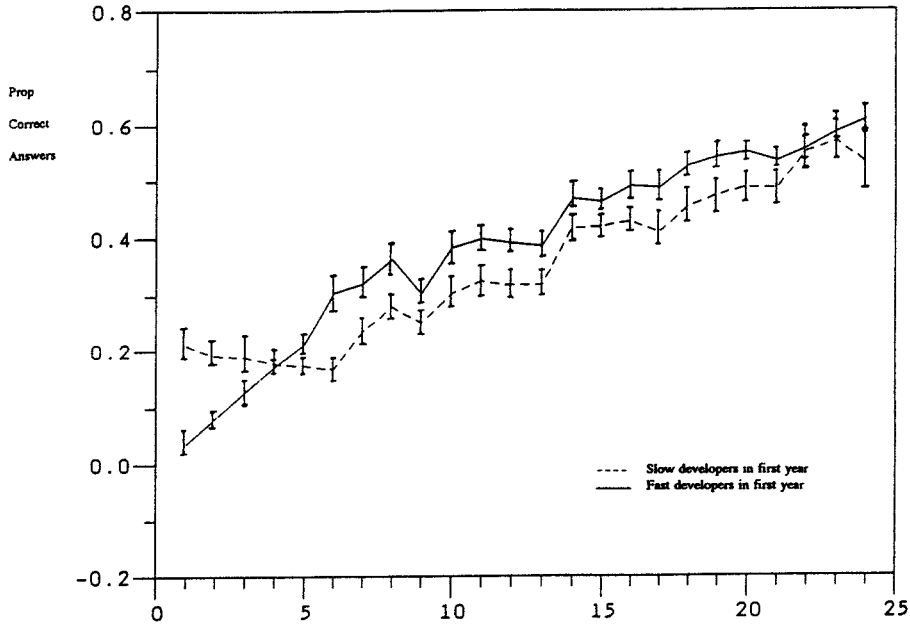


FIGURE 2. *Cohort 1982: Mean and standard error proportion correct*

IRT-models and whatever the structural form of knowledge may be, the proposed model can be considered as a generalization of the progress test model mentioned in the second section. In evaluating growth of knowledge between groups, the item difficulties are considered as nuisance parameters. By taking the expectation with respect to the distribution of these item difficulties, we manage to express the growth model unconditionally on the item difficulties. In particular, we need not estimate these item parameters. The different groups were constructed to investigate the role that growth in knowledge during the first year would play in summative as well as in formative evaluation of achievement. However, it goes beyond the objective of this article to give an explanation of the differences in growth of knowledge during the first year. Further research needs to focus on finding the explanations for these differences.

From the evidence of Figures 1–3 and the performed randomization tests, we argue that there is evidence for the following statement. The (S)low growers start with a higher level of knowledge at medical school entrance but lose their advantage after the first year and show a lower ability than the fast growers throughout the rest of the academic program. The difference for the 1978 cohort is more pronounced in the sixth year than in the first 4 years. With respect to the 1982 cohort, the difference is most pronounced during the first 4 years and tends to disappear at the end of the academic program. In this article, we have defined as slow and fast growers those students whose linear growth of knowledge during the first year of their academic program falls in the lowest and highest 10%, respectively, of the distribution of all possible first year's

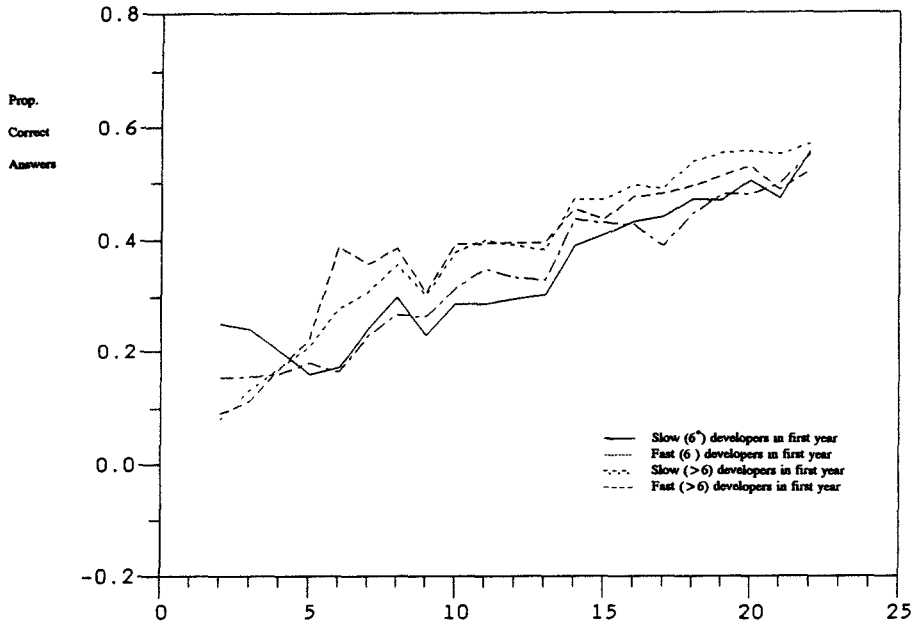
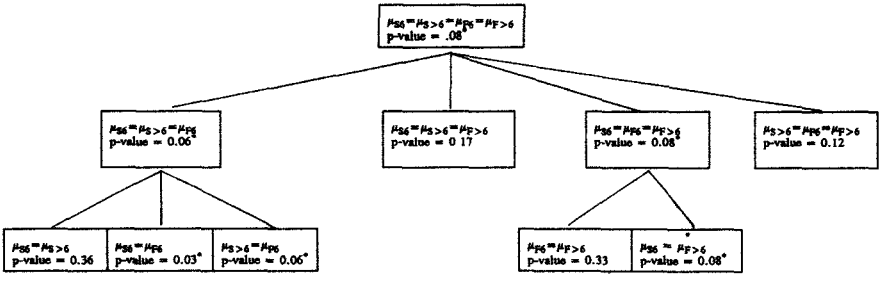


FIGURE 3. *Cohort 1982: Mean and standard error proportion correct*
*(6) finished within six years
(>6) needed more than six years

linear growth of knowledge. Considering the above analysis, we conclude that students who demonstrate little or no growth of knowledge in their first year (lowest 10 percent) will demonstrate a lower level of knowledge after the second academic year in comparison to the fast growers (upper 10%) and continue to do so throughout the rest of the academic program. Obviously, growth of knowledge in the first academic year has a predictive value towards the final level of knowledge at the end of the academic program. The speed of knowledge growth becomes an interesting evaluation device of students' performance.



SCHEME 1. *Group comparison Cohort 1982*
*Significant at 10% level according to the Peritz closure step-down procedure.

References

Albers, W., Does, R. J. M. M., Imbos, Tj., & Janssen, M. P. E. (1989). A stochastic growth model applied to repeated tests of academic knowledge. *Psychometrika, 54*(3), 451–466.

Andersen, E. B. (1980). Comparing latent distribution. *Psychometrika, 42*(1), 69–81.

Andersen, E. B. (1983). Latent trait models. *Journal of Econometrics, 22,* 215–227.

Edgington, E. S. (1987). *Randomization tests* (2nd ed.). New York: Marcel Dekker.

Goulet, L. R. (1975). Longitudinal and time-lag designs in educational research: An alternate sampling model. *Review of Educational Research, 45,* 505–523.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures.* New York: Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Tan, E. S., Imbos, Tj., Does, R. J. M. M., & Theunissen, M. (1992, April). *Predicting acquired knowledge using a longitudinal system of educational achievement.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Tan, E. S., Roos, J. M. A., Volovics, A., van Baak, M. A., & Does, R. J. M. M. (1992). An interactive computer program for randomization analysis of response curves with facilities for multiple comparisons. *Computers and Biomedical Research, 25,* 101–116.

van der Linden, W. J. (1987). *Het zwalkende niveau van het onderwijs* [The fleeting nature of educational level]. 26th anniversary address of the University of Twente, Netherlands.

Willoughby, T. L., & Hutcheson, S. J. (1978). Edumetric validity of the quarterly profile examination. *Educational and Psychological Measurement, 38,* 1057–1061.

Zerbe, G. O. (1979). Randomization analysis of the completely randomized design extended to growth response curves. *Journal of the American Statistical Association, 74,* 215–221.

Zerbe, G. O., & Murphy, J. R. (1986): On multiple comparisons in the randomization analysis of growth and response curves. *Biometrics, 42,* 795–804.

Authors

E. S. TAN is Senior Lecturer, Department of Methodology and Statistics, University of Limburg, Fac. 1, Room 2072, Peter Debeyeplein 1, Maastricht, The Netherlands. *Degree:* MD, University of Nÿmegen. *Specializations:* applied statistics and psychometric theory.

TJ. IMBOS is Senior Lecturer, Department of Methodology and Statistics, University of Limburg, Fac. 1, Room 2072, Peter Debeyeplein 1, Maastricht, The Netherlands. *Degrees:* MA, University of Groningen; PhD, University of Maastricht. *Specialization:* psychometric theory.

R. J. M. M. DOES is Professor, Industrial Statistics, Department of Mathematics, University of Amsterdam/Management Consultant, Centre for Quantitative Methods, Vonderweg III Building HC2-3, 5600 AK Eindhoven. *Degrees:* MD, PhD, University of Leiden. *Specializations:* mathematical and applied statistics.