The Economic and Social Review, Vol. 18, No. 3, April, 1987, pp. 137–147.

A Second-order Factor Analysis of the Reliability and Validity of the 11 plus Examination in Northern Ireland

BRENDAN BUNTING
University of Ulster at Jordanstown

WILLEM E. SARIS
University of Amsterdam

JOE McCORMACK
University of Ulster at Coleraine

Abstract: In Northern Ireland the transition from primary to secondary education is a critical process since the type of secondary school attended is a powerful determinant of the life-chances of individuals. In Northern Ireland children enter the secondary sector of education at about age eleven and the allocation process for non-fee payers is decided by performance in two parallel tests each lasting for 50 minutes. The examination takes place early in the last year at primary school with an interval of about six weeks between the two tests. These tests, commonly referred to as the 11 plus, are of profound importance to the children, parents and the allocating agency. This paper attempts to assess the reliability and validity of this allocation mechanism. The two test papers for three consecutive years namely 1982, 1983 and 1984 (six papers in all) were completed by the children (n = 47) over a period of weeks. The analysis of the participants' marks was based on a second order factor analysis using both the structural and measurement model in LISREL. This permits the variance to be partitioned into three components: (a) a common component, (b) a unique component, i.e., that which is specific to any particular test, and (c) the measurement error variance. The validity of a test is measured by (a), the invalidity by (b) and the unreliability by (c). The reliability and validity of this restricted sample, suggests that should such results be generalised, a large number of children would have been incorrectly classified by these test papers.

I INTRODUCTION

The 11 plus examination, in common with many educational and psychological tests, is based on the assumption that there is some fixed relatively stable attribute or trait which is real and can be measured. There is also the implicit assumption that this fixed quantity of intelligence or ability can be used to order individuals hierarchically in terms of the attribute. Hence it

follows that individuals can be separated using some arbitrary cut-off point, into those who would benefit from an élite and privileged education, and those (the great majority) who could be largely eliminated from the race for educational attainment.

These are commonplace and workable assumptions. They are commonplace in so far as (a) they direct our attention to the relative stability and predictability of much human behaviour and (b) because they present a model of human ability which implies and perpetuates inequality. However, they are also workable assumptions, to the extent that it is possible to identify the criteria used to justify educational segregation on the basis of "ability". The most overt criterion that is used to justify the policy of educational segregation is the 11 plus test. Of course, there may be many other criteria involved which are implicit within the structure of education and the wider society. An example would be the idea that females, who because they do better on average than males in this examination, should then be discriminated against by an external criterion of maturity, which may or may not be related to ability. For a more wide ranging discussion of the criteria involved in the selection procedure see McCormack and Bunting (1985).

The most crucial criteria in the construction and application of tests should be their validity and reliability. Few would deny the importance of knowing whether or not the items in a test are indeed measuring what they are intended to measure (i.e., whether they are valid). Likewise it is commonly required that those taking a test should maintain their rank order on a subsequent application of the same, or similar test, thus indicating reliability. This study examines the validity and reliability of the 11 plus examination which is used to determine the type of secondary school attended in Northern Ireland.

II METHOD

Respondents

In the run-up to the 11 plus examination 47 pupils (all girls) were asked, over a period of three months, to complete six 11 plus papers; those of 1982, 1983 and 1984. The 11 plus examination consists of two papers each year. These tests were administered by, and completed under, the supervision of the class teacher. This is not a random sample of children taking the test in Northern Ireland. However, this is not a major problem since the purpose of the study was to compare the results of the *same pupils on different test papers.*

Tests

The 11 plus examination is based on two 50 minute test papers, each with 100 questions. The two papers are completed by the children in the first term of their seventh year at primary school with an interval of about six weeks between tests.

A child is graded on the basis of his/her performance over the two papers. The top 20 per cent are given grade A which entitles them to a free place in a grammar school. A further 10 per cent receive grade M, which does not automatically assure them of a place in a grammar school. The acceptance of these candidates as non-fee-paying pupils is at the discretion of the grammar school. The great majority of those taking the 11 plus are given grade G, i.e., a fail. These children, with the exception of fee-payers, have no option but to attend a non-grammar secondary school.

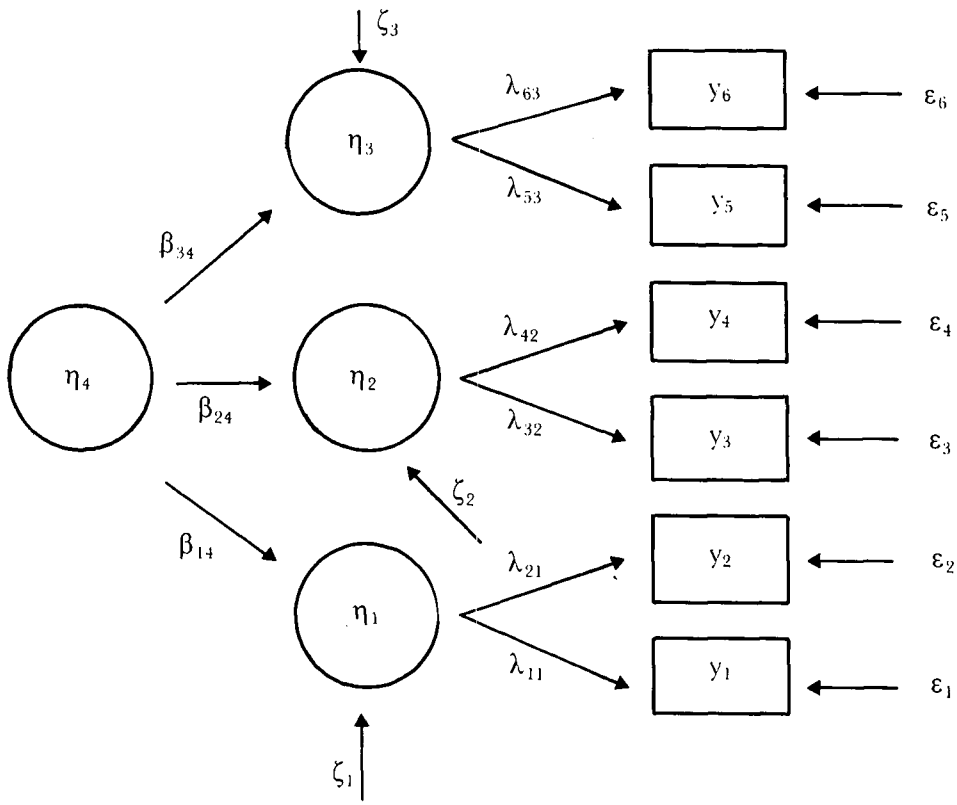
Statistical Procedure

The validity, invalidity and reliability of the 11 plus test was assessed by procedures outlined by Lord and Novick (1968) and Heise and Bohrnstedt (1970).

Scores are said to be reliable to the extent that they are repeatable and hence reproducible in terms of (a) an alternative form of the test or (b) by repeating the test with the same individuals at different points in time. On the other hand, the central issue regarding validity is whether or not a test measures what it is intended to measure. It is sometimes assumed that validity is the square root of reliability. However, as Heise and Bohrnstedt (1970) have demonstrated, this will only hold when invalidity is zero. When invalidity is not assessed, then unwittingly, reliability can be high due to items from a domain of content other than those being examined.

In the present study, validity, invalidity and reliability were assessed using a second order factor analysis as described by Joreskog (1971, 1974) and Saris (1982). (See Diagram 1 and Table 1 for a presentation of the model.) Within this framework it is postulated that the two papers in each year are combined and that they are both attempting to measure the same thing(s). It is further postulated that the combined tests for each year are themselves an attempt to measure the same ability(ies) over a number of years, since we would want to minimise the possibility of candidates passing on one year's papers and failing on another year's.

Diagram 1: *The Structural and Measurement Model for Test Papers*



Note: y_1 = Paper 1 '82
y_2 = Paper 2 '82
y_3 = Paper 1 '83
y_4 = Paper 2 '83
y_5 = Paper 1 '84
y_6 = Paper 2 '84

Table 1: *The Structural and Measurement Model can be Written in the Following Matrix Equations.*

Structural Model: $\eta = \beta\eta + \zeta$

$$\begin{bmatrix}\eta_1\\ \eta_2\\ \eta_3\\ \eta_4\end{bmatrix} = \begin{bmatrix}0 & 0 & 0 & \beta_{14}\\ 0 & 0 & 0 & \beta_{24}\\ 0 & 0 & 0 & \beta_{34}\\ 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\\ \eta_3\\ \eta_4\end{bmatrix} + \begin{bmatrix}\zeta_1\\ \zeta_2\\ \zeta_3\\ \zeta_4\end{bmatrix}$$

$$\psi = E(\zeta\zeta') = \begin{bmatrix}\psi_{11} & 0 & 0 & 0\\ 0 & \psi_{22} & 0 & 0\\ 0 & 0 & \psi_{33} & 0\\ 0 & 0 & 0 & \psi_{44}\end{bmatrix}$$

Measurement Model: $y = \Lambda_y\eta + \varepsilon$

$$\begin{bmatrix}y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\end{bmatrix} = \begin{bmatrix}\lambda_{11} & 0 & 0\\ \lambda_{21} & 0 & 0\\ 0 & \lambda_{32} & 0\\ 0 & \lambda_{42} & 0\\ 0 & 0 & \lambda_{53}\\ 0 & 0 & \lambda_{63}\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\\ \eta_3\end{bmatrix} + \begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \varepsilon_4\\ \varepsilon_5\\ \varepsilon_6\end{bmatrix}$$

$$\theta_\varepsilon = E(\varepsilon\varepsilon') = \begin{bmatrix}\theta_{\varepsilon_{11}} & 0 & 0 & 0 & 0 & 0\\ 0 & \theta_{\varepsilon_{22}} & 0 & 0 & 0 & 0\\ 0 & 0 & \theta_{\varepsilon_{33}} & 0 & 0 & 0\\ 0 & 0 & 0 & \theta_{\varepsilon_{44}} & 0 & 0\\ 0 & 0 & 0 & 0 & \theta_{\varepsilon_{55}} & 0\\ 0 & 0 & 0 & 0 & 0 & \theta_{\varepsilon_{66}}\end{bmatrix}$$

$E(\varepsilon) = 0$
$E(y) = 0$; $E(\eta) = 0$; $E(\zeta) = 0$

while it is assumed that $E(\eta\varepsilon') = 0$; $E(\zeta\varepsilon') = 0$; $E(\eta_4\zeta') = 0$
For identification purposes the following restrictions are introduced without loss of generality $\psi_{44} = 1$; $\lambda_{11} = 1$; $\lambda_{32} = 1$; $\lambda_{53} = 1$.

In order to obtain validity coefficients true scores were distinguished for the papers in 1982, 1983 and 1984 (η_1, η_2 and η_3). Since these tests are attempting to measure the same ability over a period of time, the question arises as to whether or not these true scores are indeed measuring a common ability (η_4). This can be formulated as a factor analytic model, where the common factor (η_4) indicates the shared component of the three true scores. These effects are indicated by the β coefficients. In this way validity coefficients are obtained which relate the dimension measured without error to the latent concept. The unique component for the true score for each set of tests in the model is indicated by ζ (zeta). The variance of these unique components indicates the extent of invalidity in the test. In other words, it is the variance which is due to some other factor(s) not held in common with the latent factor (η_4). The effects of the true scores (η_1, η_2 and η_3) on the observed scores are represented by the λ coefficients which when standardised and squared indicate the reliability of the individual tests.

Within this model the total variance equals the sum of the common factor variance, the unique variance and the error variance. The division of the obtained score variance can be described within the LISREL system as:

$$\sigma^2 y_i = \lambda_{ik}^2 \beta_{k_4}^2 + \lambda_{ik}^2 \psi_{kk} + \theta_{\varepsilon_{ii}}$$

where
$\sigma^2 y_i$ = total variance of y_i
$\lambda_{ik}^2 \beta_{k_4}^2$ = common variance in the i^{th} variable
$\lambda_{ik}^2 \psi_{kk}$ = unique variance in the i^{th} variable
$\theta_{\varepsilon_{ii}}$ = measurement error in the i^{th} variable

It is assumed that (1) the errors are normally distributed, (2) that the distribution of the errors for all values of the true score are the same, and (3) the errors in the two tests are independent of each other.

All the coefficients are estimated using the program LISREL (Joreskog and Sorbom, 1984). The LISREL procedures provide under the given conditions for efficient estimators of structural equation models containing unobserved variables. This approach permits the simultaneous estimation of both the parameters linking the observed variables to latent, unobserved variables (the measurement model) and the parameters linking the latent variables to each other (the structural model). This problem also provides a goodness of fit statistic, which while not without limitations (Saris *et al*, 1979) can nevertheless be used as a guide to the appropriateness of the model. Saris and Satorra (1984) and Satorra and Saris (1985) also provide an alternative approach.

Such a model can be shown to be identified in both its first and second order factors (e.g., Costner, 1969; Wiley, 1973); and since the parts are identified the

entire model must also be identified. Also, the LISREL program routinely searches for identification problems by checking if the information matrix is positive definite. As noted by Joreskog and Sorbom (1984): "If the model is identified then the information matrix is almost certainly positive definite."

III RESULTS

The correlation matrix (Table 2) indicates the relationship between scores on the six papers. These range from .44 between the second papers in 1982 and 1983 to .84 for the corresponding first papers in the same years. The means for each of the papers are also very variable, ranging from below 50 for the first paper in 1984 to nearly 67 in the first paper of 1983. However, the standard deviations are relatively consistent.

Table 2: *Correlation, Matrix, Means and Standard Deviations for the Six 11 Plus Papers*

Variables	P1 '82	P2 '82	P1 '83	P2 '83	P1 '84	P2 '84
Paper 1 '82	1.0000					
Paper 2 '82	.6970	1.0000				
Paper 1 '83	.8406	.5884	1.0000			
Paper 2 '83	.7344	.4398	.8105	1.0000		
Paper 1 '84	.7371	.6938	.7247	.5661	1.0000	
Paper 2 '84	.7566	.6100	.7727	.6828	.8102	1.0000
Means	55.6596	61.3404	66.7447	64.6596	49.4894	58.4043
Standard Deviation	16.1707	16.8462	14.8184	14.2330	15.8497	14.6994

On the basis of these data the statistical model, which has been described, was tested using the likelihood ratio test available in the LISREL program. The test statistic had a value of 10.731 with 6 degrees of freedom (probability level = .097). The proposed model cannot be rejected at the .05 level and hence the model is accepted for the time being as a good description of the data. The estimated values for the parameters in the model are presented in Table 3.

These parameter values can be used to estimate both the common variance (validity) and the error variance of each test. The error variance is the sum of (a) the systematic error variance in the true scores which is not explained by η_4 (invalidity) and (b) the random error variance present in the observed variable (unreliability).

Table 3: *Values for Coefficients in Diagram 1*

Coefficient	*Unstandardised*	*Residual*	*Variance*	*Reliability*
		Measurement Model		
λ_{11}	1.000	$\theta_{\varepsilon_{11}}$	18.445	.929
λ_{21}	.781	$\theta_{\varepsilon_{22}}$	135.462	.523
λ_{32}	1.000	$\theta_{\varepsilon_{33}}$	12.064	.945
λ_{42}	.842	$\theta_{\varepsilon_{44}}$	61.766	.695
λ_{53}	1.000	$\theta_{\varepsilon_{55}}$	56.436	.775
λ_{63}	.969	$\theta_{\varepsilon_{66}}$	33.141	.847
		Structural Model		Validity of true scores
β_{14}	14.843	ψ_{11}	22.746	.906
β_{24}	13.514	ψ_{22}	24.879	.880
β_{34}	12.730	ψ_{33}	32.728	.832

It can be seen from Table 4 that the common variance is very changeable. The first paper in 1982 has over 84 per cent of its variation coming from a common factor, but the second paper has less than 48 per cent of its variation in common with the latent construct. In three of the remaining tests there is a large amount of variance unaccounted for by the common factor variance as shown by the error variance.

These errors, signifying both test invalidity and unreliability, have consequences for the selection procedures, since the greater the amount of error, the more likely it is that those taking the test will be incorrectly placed. The degree of incorrect selection can be obtained since (a) the probability of all deviations from the true score (τ) can be calculated and (b) the probability of a value less than a given critical value ($<$cv) or cut-off point can be estimated, given that τ has a certain value. Accepting the previous assumptions, relating to the errors, then: $Z = cv - \tau/\sigma_{error}$ is distributed as a standard normal variable, which means that the probability for all deviations from τ can be estimated.

Since it is assumed that the errors in the two tests are independent of each other, the error variance for the combined tests is then the sum of the error variance in each test. The standard error can then be determined as follows:

$$\sigma_{error} = \sqrt{\sigma_e^2\,test_1 + \sigma_e^2\,test_2}$$

Where $\sigma_e^2\,test_1 = \lambda_{ik}^2\psi_{kk} + \theta_{\varepsilon_{ii}}$

$$\sigma_e^2\,test_2 = \lambda_i^2 + {}_{lk}\psi_{kk} + \theta_{\varepsilon_{i+1,i+1}}$$

And where $\lambda_{ik}^2\psi_{kk} + \theta_{\varepsilon_{ii}}$ = the error variance for the first paper in a given year.

$\lambda_i^2 + {}_{lk}\psi_{kk} + \theta_{\varepsilon_{i+1,i+1}}$ = the error variance for the second paper in a given year.

Table 4: *The Total Variance Partitioned Into (a) the Common Variance and (b) the Error Variance*

	Common Variance $\lambda_{ik}^2\beta_{k_4}^2$	*Error Variance* (σe^2) $\lambda_{ik}^2\psi_{kk} + \theta_{\varepsilon ii}$	*Total Variance*
P1 '82	220.315	41.191	261.506
P2 '82	134.383	149.336	283.720
P1 '83	182.628	36.943	219.571
P2 '83	124.000	78.658	202.658
P1 '84	162.053	89.164	251.217
P2 '84	152.161	63.871	216.032

The error variance for the first test paper in 1982 is 41.191, i.e. $(1.0^2)(22.746) + (18.445) = 41.191$. For the second test paper in 1982 the error variance is 149.336, i.e. $(.781^2)(22.746) + (135.462) = 149.336$. The combined error variance for the two test papers in 1982 is: 41.191 + 149.336 = 190.527. To obtain the standard error (σ_e) we take the $\sqrt{}$ of the error variance, i.e. $\sqrt{190.527} = 13.803$.

The misclassification of candidates can be obtained by dividing the difference between the cut-off point and the true score by the standard error and consulting Z tables. The percentage incorrectly classified, for selected differences between the cut-off point and the true score, is given in Table 5.

Table 5: *Chance of Incorrect Selection for Varying Differences Between the Cut-Off Point and the True-Score*

Difference Between Cut-Off Point and True-Score	*Chance of being Incorrectly Classified (%)*		
	1982	1983	1984
1	47	46	47
2	44	42	44
3	41	39	41
4	39	36	37
5	36	32	34
6	33	29	31
7	31	26	28
8	28	23	26
9	26	20	23
10	24	18	21
15	14	8	11
20	7	3	5

A difference of 1 to 3 points between a person's true score and the cut-off mark in the examination, could lead to more than a 40 per cent chance of being incorrectly classified. When the discrepancy between the true score and the cut-off point is 10 marks then there is a 20 per cent chance of the person being incorrectly classified. Even with a 20 mark discrepancy between the true score and the cut-off point on a scale of 0-200 we would still expect about 5 per cent of the candidates to be placed in the wrong category.

IV DISCUSSION

Any test which sets out to assess children in two 50 minute periods after some 6 (or 7) years of schooling must itself conform to the highest possible criteria. This is even more so when the test has real consequences for a child's future educational and occupational chances.

On the basis of the results (from this restricted sample) the 11 plus examination leaves a lot to be desired. At the correlational level there is a marked variability between supposedly parallel papers. Too much of the variation between the papers is left to extraneous and unexplained factors, e.g., item selection, nervousness, parental influences, etc. The papers also have fluctuating standards as shown by the means. This may not have an effect on the selection mechanism, but it does raise other issues.

However, the central issue in any test is its reliability and validity. Like the correlations and the means, there is a marked variation between the test papers. Some, like the first papers in 1982 and 1983 are very reliable and have also a reasonably high validity component. On the other hand, the second papers for the same years leave a lot to be desired in terms of both reliability and validity. Sizeable errors, as present in the unreliable and invalid components of a set of tests, can adversely affect the selection procedure.

Of course, in any selection procedure some children/candidates will be misplaced. During the time that the 11 plus was used as a means for selection to grammar schools in Britain, it was estimated that some 70,000 children a year were misplaced (Pidgeon, 1970). In essence, in order to use a selection procedure, we have to be willing to accept some degree of unreliability in our instrument. A crucial question then becomes (for those who wish to pursue this course) what level of inaccuracy are we willing to tolerate in our test? On the basis of the calculations made for differences between the test cut-off point and person's true score, the developers of this test are apparently willing to tolerate a very large degree of inappropriate selection. Using our procedures and given the nature of the sample it is not possible to state the numbers of children who are inappropriately placed each year, but to misclassify a child who is 10 points above an arbitrary cut-off point 1 time in 5 is a cause for concern. However, care is required in generalising from such a restrictive sample.

The present estimates are probably conservative for a number of reasons. In the present study only one school was used. Since we are to some degree controlling for a number of factors which might also have an effect on these scores, e.g., variability between teachers, school ethos, social class, gender, etc., there is good reason to suppose that should this study be extended to include other schools, that there would be even greater variability and hence more error present. It is also possible that some serial order effect is present in the data, though since the model fitted without the necessity to include correlated error this does not seem to be as strong a possibility as might be thought.

It is obvious that more data would be of great assistance in coming to a more complete understanding of this test. Much of the required data is already available in tests taken by students in previous years. On the basis of the data presented in this paper there would appear to be some cause for concern, over the validity, invalidity and reliability of this examination.

REFERENCES

COSTNER, H. L., 1969. "Theory, Deduction and Rules of Correspondence", *American Journal of Sociology*, 75, pp. 245-263.

HEISE, D. R., and G. W. BOHRNSTEDT, 1970. "Validity, Invalidity and Reliability", in E. F. Borgatta and G. W. Bohrnstedt (eds.), *Sociological Methodoloty*, San Francisco: Jossey-Bass, pp. 104-129.

JORESKOG, K. G., 1971. "Statistical Analysis of Sets of Congeneric Tests", *Psychometrika* Vol. 36, No. 2, pp. 109-133.

JORESKOG, K. G., 1974. "Analyzing Psychological Data by Structural Analysis of Covariance Matrices", in R. C. Atkinson, D. H. Krantz, R. D. Luce and P. Suppes (eds.), *Contemporary Developments in Mathematical Psychology - Vol. II*, San Francisco: W. H. Freeman, pp. 1-56.

JORESKOG, K. G., and D. SORBOM, 1984. LISREL VI: *A General Computer Program for Estimation of Linear Structural Equation Systems by Maximum Likelihood Methods*, Mooresville, Indiana: Scientific Software, Inc.

LORD, F. M., and M. R. NOVICK, 1968. *Statistical Theories of Mental Test Score*, Reading, Mass: Addison-Wesley Publishing.

McCORMACK, J., and B. BUNTING, 1985. "The Eleven Plus: Unjust, Unreliable and Invalid", *Fortnight*, No. 218, April-May, pp. 12, 16.

PIDGEON, D., 1970. "The Selection Shambles", *Where*, Vol. 48, pp. 46-48.

SARIS, W. E., 1982. "Different Questions, Different Variables", in W. E. Saris, *Linear Structural Relationship: Measurement Models*, Amsterdam: Sociometric Research Foundation, pp. 78-83.

SARIS, W. E., W. M. DE PIJPER and P. ZEGWAART, 1979. "Detection of Specification Errors", in K. Schuessler (ed.), *Sociological Methodology*, San Francisco: Jossey-Bass, pp. 151-171.

SARIS, W. E., and A. SATORRA, 1984. "The Likelihood Ratio Test of Structural Equation Models", in W. E. Saris (ed.), *Sociometric Research*, Amsterdam: Sociometric Research Foundation.

SATORRA, A., and W. E. SARIS, 1985. "Power of the Likelihood Ratio Test in Covariance Structure Analysis", *Psychometrika*, Vol. 50, No. 1, pp. 83-90.

WILEY, D., 1973. "The Identification Problem for Structural Equation Models with Unmeasured Variables", in A. S. Goldberger and O. D. Duncan, (eds.), *Structural Equation Models in the Social Sciences*, New York: Seminar Press, pp. 69-83.