DOCUMENT RESUME

ED 323 217 TM 015 400

AUTHOR	Kim, Yang Boon; Lee, Jong Sung
TITLE	A Study Applying Generalizability Theory to the Scientific Thinking and Research Skill Test.
PUB DATE	Apr 90
NOTE	27p.; Revision of a paper presented at the Annual Meetings of the American Educational Research Association (Boston, MA, April 16-20, 1990) and the National Council on Measurement in Education (Boston, MA, April 17-19, 1990). Top of document pages contain light type.
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Analysis of Variance; Decision Making; *Elementary School Students; *Generalizability Theory; Grade 5; Grade 6; Intermediate Grades; *Logical Thinking; Research Design; Scientific Concepts; Test Bias; Test Interpretation; Test Items; *Test Validity; Theory Practice Relationship
IDENTIFIERS	*Scientific Thinking and Research Skill Test

ABSTRACT

The empirical validity of generalizability theory was investigated by applying two three-facet designs to data obtained in 1988 from administration of the Scientific Thinking and Research Skill Test (STRST). The decision validity of the STRST was also examined. Subjects were 125 fifth-grade and 125 sixth-grade students who were administered the STRST in a test and retest. The STRST contains 13.items on scientific skills domain and 13 items within the logical thinking domain. Applying generalizability theory to the data resulted in the observed score variance being partitioned in two ways, identifying different sources of error and their relative magnitudes. Test domain was one of the large variance components of the total score variance, a finding suggesting that test scores should be interpreted from the separate domain scores rather than from the total score. The interaction effect for persons and items within test domains suggests the possibility of biased item sampling. Findings from the generalizability study further imply that the number of test items should be increased to more than 20 in each domain to attain satisfactory generalization of the STRST into its universe. Recommendations for improvement of the STRST are presented. Nine tables present study data, and two figures illustrate the research designs. (SLD)

A Study Applying Generalizability Theory to the Scientific Thinking and Research Skill Test

Yang Boon Kim

Korean Educational Development Institute

Jong Sung Lee

Yonsei University

Running head: A STUDY APPLYING GENERALIZABILITY THEORY

This is a revision of a paper presented at the annual meeting of AERA and NCME, Boston, MA., April, 1990.

Yang Boon Kim is Senior Researcher, Korean Educational Development Institute, Seoul, Korea. She specializes in educational measurement and evaluation.

Jong Sung Lee is Professor, Department of Education, Yonsei University, Seoul, Korea. He specializes in educational statistics and measurement

Abstract

The purpose of the study was to investigate the empirical validity of generalizability theory by applying two, three-facet designs to data from the Scientific Thinking and Research Skill Test (STRST) and to investigate the decision validity of the STRST. Implications for test developers and program evaluators were suggested, and also recommendations were provided for the further improvement of the STRST.

A Study Applying Generalizability Theory to the Scientific Thinking and Research Skill Test

Generalizability theory has been applied in various fields such as program evaluation (Rothman, 1982), behavior observation (Smith & Teeter, 1982), rating of instruction (Gillmore, 1980), evaluation of students' attitudes to school subjects (Carloni & Molen, 1980), and so forth. However, most of these practical studies did not deal with more than three facets of the theory.

The purpose of this study was to investigate the empirical validity of generalizability theory by applying two, three-facet designs of the theory to the Scientific Thinking and Research Skill Test (STRST) data (Cho & Kim, 1988) and to investigate the validity of the STRST. There are three parts in the study: (a) estimation of STRST variances, (b) decision study of STRST for follow-up studies, and (c) comparison of the STRST classical reliability coefficients with generalizability coefficients.

Method

Research Design

The research design of the study used both a G (Generalizability) study px(i:h*)xo design, D (Decision) study px(I:H*)xO design and also a G study (p:j)x(i:h*) design, D study (p:J)x(I:H*) design. Figures 1 and 2 are Venn diagrams for the px(i:h*)xo design and the (p:j).(i:h*) design, respectively.[1]

Insert Figures 1 and 2.about here

In the G stud,' $px(i:h^*)xo$ design, persons (p) were crossed with test items (i), test content domains (h), and test occasions (o), and test items were nested within each content domain. It was assumed that persons were randomly sampled from an infinite population ($n_p<N_p\rightarrow\infty$), that test items and test occasions were randomly sampled from infinite universes, respectively ($n_i<N_i\rightarrow\infty$, $n_o<N_o\rightarrow\infty$), and that test contents were fixed as two domains ($n_h=N_h=2$). Also, it was assumed that the D study $px(I:H^*)xO$ design had the same structure as the G study design ($n_i'<N_i'\rightarrow\infty$, $n_o'<N_o'\rightarrow\infty$, $n_h'=N_h'=2$).[2]

In both the G study $(p:j)x(i:h^*)$ design and the D study $(p:J)x(I:H^*)$ design, persons (p) were nested within schools (j) and crossed with test items (i), which were nested within test content domains (h). It was assumed that persons, schools, and test items were random samples from an infinite population and universe, respectively ($n_p<N_p\rightarrow\infty$, $n_j<N_j\rightarrow\infty$, $n_i<N_i\rightarrow\infty$), and that test contents were fixed as two domains ($n_h=N_h=2$).

The two designs above were balanced designs in which there were the same number of items within the two test content domains and the same number of students within each school.

Research Subjects and Data

Research subjects were students in grades 5 and 6 who were administered both the test and retest during the development of the STRST. Two hundred fifty students, 125 from grade 5 and 125 from grade 6 were randomly sampled from those who were administered both the test and retest of the STRST in order to obtain an equal number of subjects from each of five schools which had been randomly selected from a random sample of 28 primary schools from 12 cities in Korea.

The STRST consists of 13 items within the scientific skills and 13 items within the logical thinking domain. STRST data were coded 1 for right answers and 0 for wrong answers. The data to be analyzed were the scores on the STRST that were obtained in 1988 during the project to develop identification instruments for the scientifically gifted in grades 5 and 6.

Results

Variance Estimates of the STRST

Tables 1 and 2 show the results of estimating variance components of the G study px(i:h*)xo design in grade 5 and 6, respectively. The total variance of STRST was partitioned into 11 components. Among them, the largest variance component in the proportion of the total variance was the residual effect and/or interaction effect for persons and test occasions and test items within test domains (39.20% in grade 5 and 37.63% in grade 6), the second largest variance component was the test domain effect in grade 5 (23.67%), and the interaction effect for persons and items within test domains in grade 6 (25.69%), and the third iargest variance component was the interaction effect for persons and items within test domains in grade 5 (18.54%), and the test domain effect in grade 6 (17.16%). The item effect within test domains was the fourth largest variance component contributing to the differences of the test scores (8.64% in grade 5 and 6.53% in grade 6). The main effect and interaction effects of the test occasion which was not a facet in the (p:j)x(i:h*) design had a little contribution to the differences of the test scores.

Insert Tables 1 and 2 about here

Tables 3 and 4 show the results of estimating the variances of the G study (p:j)x(i:h*) design in grade 5 and 6, respectively. The total variance or STRST scores was partitioned into 8 components. Among them, the largest variance component in the proportion of the total variance was the residual and/or interaction effects for persons and items within schools and test domains (55.98% in grade 5 and 59.92% in grade 6), the second largest variance component was the test domain effect (26.20% in grade 5 and 20.82% in grade 6), and the third largest variance component was the item effect within test domains (8.80% in grade 5 and 7.51% in grade 6). The main effect and interaction effects of school which was not a facet in the px(i:h*)xo design made little contribution to the differences of the test scores.

Insert Tables 3 and 4 about here

In both the px(i:h*)xo design and the (p:j)x(i:h*) design, test domain, test item, and interaction for persons and items within test domains were all significant effects on the differences among the test scores. In both two designs, the estimates of universe variances, $\hat{\sigma}^2(p\backslash H)$ and $\hat{\sigma}^2(p:j\backslash H)$, did not show significant differences, suggesting that school effect made little contribution to the variability of the test scores.

Decision Study of the STRST

The results of the D study px(I:H*)xO design are reported in Tables 5 and 6. When the number of test items was increased from $n_i'=13$ to $n_i'=25$, the generalizability coefficient, $E\hat{\rho}^2$, was increased by .09 for $n_o'=2$ and by .14 for $n_o'=1$ in grade 5, and by .08 for $n_o'=2$ and by .10 for $n_o'=1$ in grade 6. Similar figures could be found in dependability, $\hat{\Phi}(\lambda)$, for domain-referenced interpretations (Brennan, 1983, p.108). These results indicate that the coefficients of the D study px(I:H*)xo design were more influenced by the number of test items than by test occasions.

Insert Tables 5 and 6 about here

The results of the D study (p:J)x(I:H*) design are reported in Tables 7 and 8. The coefficients of the D study (p:J)x(I:H*) design were dependent upon the number of test items, but not upon the number of schools. Considering the same number of test items, the generalizability coefficients and dependability indices of the D study (p:J)x(I:H*) design ($n_h'=2$ and $n_j'=5$) were lower than those of the D study px(I:H*)x0 design ($n_h'=2$ and $n_o'=2$) in both grades 5 and 6. However, the generalizability coefficients and dependability indices of the D study (p:J)x(I:H*) design ($n_j'=5$) were higher than those of the D study px(I:H*)x0 deisgn ($n_o'=1$) in grade 6, while the coefficients and indices of the (p:J)x(I:H*) design were lower than those of the px(I:H*)x0 design in grade 5.

Insert Tables 7 and 8 about here

Comparisons of Coefficients

Table 9 compares generalizability coefficients, dependability indices, and classical reliability coefficients. KR-20 and Cronbach's alpha coefficient are the same as the generalizability coefficient of the D study pxI design (Brennan, 1983, p.13). The generalizability coefficient of the D study px(I:H*)x0 design was only exceeded by the test-retest reliability coefficient. The generalizability coefficient of the D study (p:J)x(I:H*) design was lower than all others except KR-21. Also, the dependability index of the D study px(I:H*)x0 design was higher than that of the D study (p:J)x(I:H*) design.

Insert Table 9 about here

Discussion

Variance Estimates of the STRST

The observed score variance in G studies is partitioned into various components according to the research design, whereas in classical test theory analysis, the observed score variance is partitioned into a true score component and an error score component. In this study, applying generalizability theory to the STRST data resulted in the observed score variance being partitioned in two ways, the px(i:h*)xo design and the (p:j)x(i:h*) design. As a consequence, different sources of error and their relative magnitudes were identified.

In both the G study px(i:h*)xo design and the G study (p:j)x(i:h*) design, test domain, test item, and interaction for persons and items within test domains were all significant effects on the differences among the test scores. Among these effects, the test domain, which was assumed to be a fixed effect, was one of the large variance components of the total score variance. This finding suggests that test scores should be interpreted from the separate domain scores rather than from the total score, and that the assumption of two fixed domains should be reconsidered.

Another of the large variance components in both the G study px(i:h*)xo design and the G study (p:j)x(i:h*) design was the interaction effect for persons and items within test domains. This fact indicates the possibility of biased item sampling. In addition, the reason why the interaction variance for persons and items in the (p:j)x(i:h*) design is much larger than in the px(i:h*)xo design could be explained with the following two equations: $\sigma^2(pi:h\backslash H) = \sigma^2(pih\backslash H) + \sigma^2(pi\backslash H)$ in the px(i:h*)xo design, and $\sigma^2(pi:jh\backslash H) = \sigma^2(pijh\backslash H) + \sigma^2(pij\backslash H) + \sigma^2(pih\backslash H) + \sigma^2(pi\backslash H)$ in the (p:j)x(i:h*) design. That is, the interaction for persons and items in the (p:j)x(i:h*) design has more confounding effects than in the px(i:h*)xo design. More presicely, $\sigma^2(pi:jh\backslash H) = \sigma^2(pi:h\backslash H) + \sigma^2(pijh\backslash H) + \sigma^2(pij\backslash H)$.

Decision Study of the STRST

As a result of the D study with the application of generalizability theory to the STRST data, generalizability coefficients and dependability indices were estimated. When based upon the same number of test items, the generalizability coefficients and dependability indices in the D study px(I:H*)x0 design were higher than those in the D study (p:J)x(I:H*) design. The fact that the school variance in the (p:J)x(I:H*) design was separated from the universe variance and merged into error variance resulted in a smaller universe variance and larger error variance proportion within the total variance, while the test occasion effect and its interaction effect variances in the px(I:H*)xO design contributed less to the proportion of the total variance. This separation explains why lower generalizability coefficients and dependability indices occured in the D study (p:J)x(I:H*) design, and also means that the universe of generalizability in the D study (p:J)x(I:H*) design was larger than the one in the D study px(I:H*)xO design.

The study found that the coefficients of the D study px(I:H*)xO design depended more on the number of test items than on the number of test occasions. This finding was shown by the interaction variance for persons and items being larger than the one for persons and occasions.

When based upon the same number of test items, the generalizability coefficients of the D study px(I:H*)xO design (n_o'=2) were higher than those of the D study (p:J)x(I:H*) design (n_j'=5) except in the grade 6 D study (p:J)x(I:H*) design (n_j'=5) and D study px(I:H*)xO design (n_o'=1). This exception occured because the universe variance of the D study (p:J)x(I:H*) design in grade 6 was larger than in grade 5, while the universe variance of the D study px(I:H*)xO design in grade 6 was smaller than in grade 5.

In the D study (p:J)x(I:H*) design, the generalizability coefficients were highly dependent upon the number of test items, but not upon the number of schools. The reason for this dependency is that the generalizability coefficient of the D study (p:J)x(I:H*) design is determined by the universe variance and the relative error variance, and they are influenced by the number of test items rather than the number of schools.

The findings from the D study imply that the number of test items should be increased to more than 20 in each domain in order to attain satisfactory generalization of the STRST into its universe.

Comparisons of Coefficients

Assuming that the number of schools, test items, and test occasions in the D study are the same as in the G study, the study found that the generalizability coefficient of the D study px(I:H*)x0 design was only exceeded by the test-retest reliability coefficient. This finding follows from the fact that the generalizability coefficient for the D study px(I:H*)x0 design, $E\hat{\rho}^2 = \hat{\sigma}^2(p\backslash H)/[\hat{\sigma}^2(p\backslash H) + \hat{\sigma}^2(pO\backslash H) + \hat{\sigma}^2(pI{:}H\backslash H) + \hat{\sigma}^2(pOI{:}H\backslash H)]$, and the test-retest reliability coefficient are closely related to the generalizability coefficient for the D study pxI*x0 design $(n_O'=2)$, $E\hat{\rho}^2 = [\hat{\sigma}^2(p) + \hat{\sigma}^2(pI)]/[\hat{\sigma}^2(p) + \hat{\sigma}^2(pI) + \hat{\sigma}^2(pO) + \hat{\sigma}^2(pIO)]$, and the test-retest reliability coefficient for the D study pxI*x0 design $(n_O'=2)$, $\hat{\rho}_{XX'} = [\hat{\sigma}^2(p) + \hat{\sigma}^2(pI)]/\hat{\sigma}(X_{pI*O})\hat{\sigma}(X_{pI*K})$. It is known from the above three equations that the test-retest reliability coefficient has the same numerator as the generalizability coefficient for the D study pxI*x0 design $(n_O'=2)$ which has a larger universe variance than the D study px(I:H*)x0 design $(n_O'=2)$ (Brennan, 1983, p. 74; Lee, 1988, p. 166). Therefore, it is clear that the test-retest reliability coefficient for the D study px(I:H*)x0 design is higher than the generalizability coefficient for the same design.

References

Brennan, R. L. (1983). Elements of generalizability theory. IA: The American College Testing Program.

Carloni. J. A., & . Molen, K. J. (1980). The generalizability of elementary school student ratings of attitudes toward school subjects. Paper presented at the annual meeting of the Eastern Educational Research Association, Norfork, VA.

Cho, S. K. & Kim, Y. B. (1988). Development of identification instruments of the scientifically gifted of grades 5 and 6 in primary school. Korean Educational Development Institute. Research Report RR 88-3.

Gillmore, G. M. (1980). Student instruction ratings: to what universe can we dependably generalize results? Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Lee, J. S. (1988). Generalizability theory. Seoul, Korea: Yonsei University Press.

Rothman, M. L. (1982). Generalizability in program evaluation. Paper presented at the annual meeting of the American Psychological Association, Washington, D. C.

Smith, P. L., & Teeter, P. A. (1982). The use of generalizability theory with behavioral observation. Paper presented at the annual meeting of the American Psychological Association, Washington, D. C.

Footnotes

[1] This study followed Brennan's notational conventions (Brennan, 1983). * stands for fixed facets.

[2] n and N denote G study sample sizes and population and/or universe sizes, respectively, while n' and N' denote D study sample sizes and universe sizes, respectively.

Table 1

Variance Estimates of STRST for the G Study px(i:h*)xo Design: Grade 5 (n_p=125, n_o=2, n_h=2, n_i=13)

Effect (α\H)	df(α)	SS(α)	MS(α)	$\hat{\sigma}^2$(α\H)	Proportion of estimate(%)
p\H	124	137.03938	1.10516	.01706	5.89
o\H	1	4.18846	4.18846	.00120	.41
h\H	1	230.11215	230.11215	.06855	23.67
i:h\H	24	159.41600	6.64233	.02503	8.64
po\H	124	13.75385	.11092	(-.00011)0 [a]	0.00
ph\H	124	52.90708	.42667	.00712	2.46
oh\H	1	.77554	.77554	.00029	.10
pi:h\H	2,976	658.04554	.22112	.05368	18.54
oi:h\H	24	6.63200	.27633	.00130	.45
poh\H	124	16.62831	.13410	.00157	.54
poi:h\H	2,976	338.52185	.11375	.11375	39.20
Total	6,499	1618.02016		.28955	100.00

[a] Negative estimate was replaced by 0.

Table 2

Variance Estimates of STRST for the G Study px(i:h*)xo Design:

Grade 6 (n_p=125, n_o=2, n_h=2, n_i=13)

Effect (α\H)		SS(α)	MS(α)	$\hat{\sigma}^2$ (α\H)	Proportion of estimate(%)
p\H	124	137.8606	1.111779	.01620	6.47
o\H	1	7.1115	7.111578	.00206	.82
h\H	1	146.2500	146.250000	.04294	17.16
i:h\H	24	110.2578	4.594077	.01633	6.53
po\H	124	17.4846	.141005	.00180	.72
ph\H	124	52.1154	.420285	.00622	2.49
oh\H	1	2.3275	2.327538	.00117	.47
pi:h\H	2,976	662.8960	.222747	.06429	25.69
oi:h\H	24	9.1969	.383205	.00231	.92
poh\H	124	16.1148	.129958	.00275	1.10
poi:h\H	2,976	280.2646	.094175	.094175	37.63
Total	6,499	1441.8797		.250245	100.00

Table 3

Variance Estimates of STRST for the G Study (p:j)x(i:h*) Design: Grade 5 (n_p=25, n_j=5, n_h=2, n_i=13)

Effect (α\H)	df(α)	SS(α)	MS(α)	$\hat{\sigma}^2$ (α\H)	Proportion of estimate(%)
p:j\H	120	57.92000	.48267	.01222	4.15
j\H	4	19.21046	4.80262	.00661	2.24
h\H	1	128.80277	128.80277	.07717	26.20
i:h\H	24	82.33600	3.43067	.02593	8.80
jh\H	4	.62954	.15738	(-.00037)0[a]	0.00
ph:j\H	120	30.37538	.25313	.00679	2.30
ji:h\H	96	18.14400	.18900	.00096	.33
pi:jh\H	2,880	474.90462	.16490	.16490	55.98
Total	3,249	812.32277		.29458	100.00

[a] Negative estimate was replaced by 0.

Table 4

Variance Estimates of STRST for the G Study (p:j)x(i:h*) Design: Grade 6 (n_p=25, n_j=5, n_h=2, n_i=13)

Effect (α\H)	df(α)	SS(α)	MS(α)	$\hat{\sigma}^2(\alpha\backslash H)$	Proportion of estimate(%)
p:j\H	120	70.8123	.59010	.01659	6.26
j\H	4	5.7742	1.44354	.00113	.43
h\H	1	92.7388	92.73877	.05517	20.82
i:h\H	24	66.3963	2.76651	.01990	7.51
jh\H	4	2.3797	.59492	.00066	<.01
ph:j\H	120	31.4585	.26215	.00795	3.00
ji:h\H	96	26.7422	.27856	.00479	1.81
pi:jh\H	2,880	457.1692	.15874	.15874	59.92
Total	3,249	753.4112		.26493	100.00

Table 5

Results of the D Study px(I:H*)xO Design of STRST: Grade 5

	$\sigma^2(\bar{\alpha}\backslash H)$							
Effect $(\bar{\alpha}\backslash H)$	$n_h'= 2,\ n_o'= 2$				$n_h'= 2,\ n_o'= 1$			
	$n_i'= 13$	$n_i'= 15$	$n_i'= 20$	$n_i'= 25$	$n_i'= 13$	$n_i'= 15$	$n_i'= 20$	$n_i'= 25$
p\H	.01706	.01706	.01706	.01706	.01706	.01706	.01706	.01706
O\H	.00060	.00060	.00060	.00060	.00120	.00120	.00120	.00120
H\H	-	-	-	-	-	-	-	-
I:H\H	.00096	.00083	.00063	.00050	.00096	.00083	.00063	.00050
pO\H	<.00001	<.00001	<.00001	<.00001	<.00001	<.00001	<.00001	<.00001
pH\H	-	-	-	-	-	-	-	-
OH\H	-	-	-	-	-	-	-	-
pI:H\H	.00206	.00179	.00134	.00107	.00206	.00179	.00134	.00107
OI:H\H	.00003	.00002	.00002	.00001	.00006	.00004	.00004	.00002
pO H\H	-	-	-	-	-	-	-	-
pOI:H\H	.00219	.00190	.00142	.00114	.00538	.00380	.00284	.00228
$\hat{\sigma}^2(\tau)$	.01706	.01706	.01706	.01706	.01706	.01706	.01706	.01706
$\hat{\sigma}^2(\delta)$	.00425	.00369	.00276	.00221	.00744	.00559	.00418	.00335
$\hat{\sigma}^2(\Delta)$	.00584	.00514	.00401	.00332	.00966	.00766	.00605	.00507
$E\hat{\sigma}^2(\bar{X})$	.02131	.02075	.01982	.01927	.02450	.02265	.02124	.02041
$E\hat{\rho}^2$	.80	.82	.86	.89	.70	.75	.80	.84
$\hat{\sigma}^2(\bar{X})$ a	.00176	.00162	.00141	.00126	.00242	.00225	.00204	.00188
$\hat{\Phi}(\lambda)$ b	.72	.75	.80	.83	.60	.66	.71	.75
$\hat{\Phi}$	.74	.77	.81	.83	.64	.69	.74	.77

[a] Estimates assumed $n_p'=125$.

[b] Estimates assumed $\bar{X}=\lambda$.

Table 6

Results of the D Study px(I:H*)xO Design of STRST: Grade 6

Effect	$\sigma^2(\bar{\alpha}\backslash H)$							
$(\bar{\alpha}\backslash H)$	$n_h'= 2$, $n_o'= 2$				$n_h'= 2$, $n_o'= 1$			
	$n_i'= 13$	$n_i'= 15$	$n_i'= 20$	$n_i'=25$	$n_i'= 13$	$n_i'= 15$	$n_i'= 20$	$n_i'= 25$
p\H	.01620	.01620	.01620	.01620	.01620	.01620	.01620	.01620
O\H	.00103	.00103	.00103	.00103	.00206	.00206	.00206	.00206
H\H	-	-	-	-	-	-	-	-
I:H\H	.00063	.00054	.00041	.00033	.00063	.00054	.00041	.00033
pO\H	.00090	.00090	.00090	.00090	.00180	.00180	.00180	.00180
pH\H	-	-	-	-	-	-	-	-
OH\H	-	-	-	-	-	-	-	-
OI:H\H	.00004	.00004	.00003	.00002	.00004	.00004	.00003	.00002
pI:H\H	.00247	.00214	.00161	.00129	.00247	.00214	.00161	.00129
pOH\H	-	-	-	-	-	-	-	-
pOI:H\H	.00181	.00157	.00118	.00094	.00362	.00314	.00236	.00188
$\hat{\sigma}^2(\tau)$	.01620	.01620	.01620	.01620	.01620	.01620	.01620	.01620
$\hat{\sigma}^2(\delta)$	.00518	.00461	.00363	.00313	.00789	.00708	.00577	.00497
$\hat{\sigma}^2(\Delta)$	.00683	.00622	.00516	.00451	.01062	.00972	.00827	.00738
$E\hat{\sigma}^2(\bar{X})$	.02138	.02081	.01983	.01933	.02409	.02328	.02197	.02117
$E\hat{\rho}^2$	.76	.78	.82	.84	.67	.70	.74	.77
$\hat{\sigma}^2(\bar{X})$ [a]	.00182	.00178	.00169	.00153	.00292	.00283	.00268	.00258
$\hat{\Phi}(\lambda)$ [b]	.68	.70	.74	.76	.55	.58	.62	.65
$\hat{\Phi}$	.70	.72	.76	.78	.60	.62	.66	.69

[a] Estimates assumed $n_p'=125$.

[b] Estimates assumed $\bar{X}=\lambda$.

Table 7

Results of the D Study (p:J)x(I:H*) Design of STRST: Grade 5

Effect	$\sigma^2(\bar{\alpha}\backslash H)$							
$(\bar{\alpha}\backslash H)$	$n_h'= 2,\ n_j'= 5$				$n_h'= 2$			
	$n_i'= 13$	$n_i'= 15$	$n_i'= 20$	$n_i'= 25$	$n_j'= 30$ $n_i'= 20$	$n_j'= 40$ $n_i'= 20$	$n_j'= 30$ $n_i'= 25$	$n_j'= 40$ $n_i'= 25$
p:J\H	.01222	.01222	.01222	.01222	.01222	.01222	.01222	.01222
J\H	.00132	.00132	.00132	.00132	.00022	.00017	.00022	.00017
H\H	-	-	-	-	-	-	-	-
I:H\H	.00100	.00086	.00065	.00052	.00065	.00065	.00052	.00052
JH\H	-	-	-	-	-	-	-	-
pH:J\H	-	-	-	-	-	-	-	-
JI:H\H	.00001	.00001	<.00001	<.00001	<.00001	<.00001	<.00001	<.00001
pI:JH\H	.00634	.00550	.00412	.00330	.00412	.00412	.00330	.00330
$\hat{\sigma}^2(\tau)$	.01222	.01222	.01222	.01222	.01222	.01222	.01222	.01222
$\hat{\sigma}^2(\delta)$	.00634	.00550	.00412	.00330	.00412	.00412	.00330	.00330
$\hat{\sigma}^2(\Delta)$	.00867	.00769	.00609	.00514	.00499	.00494	.00404	.00399
$E\hat{\sigma}^2(X)$	.01856	.01772	.01634	.01552	.01634	.01634	.01552	.01552
$E\hat{\rho}^2$	.66	.69	.75	.79	.75	.75	.79	.79
$\hat{\sigma}^2(\bar{X})$[a]	.00307	.00290	.00262	.00246	.00152	.00147	.00136	.00131
$\hat{\Phi}(\lambda)$[b]	.51	.55	.61	.66	.68	.69	.73	.73
$\hat{\Phi}$	.58	.61	.67	.70	.71	.71	.75	.75

[a] Estimates assumed $n_p'=125$.

[b] Estimates assumed $\bar{X}=\lambda$.

Table 8

Results of the D Study (p:J)x(I:H*) Design of STRST: Grade 6

Effect	$\sigma^2(\bar{\alpha}\backslash H)$							
$(\bar{\alpha}\backslash H)$	$n_h'= 2$, $n_j'= 5$				$n_h'= 2$			
	$n_i'= 13$	$n_i'= 15$	$n_i'= 20$	$n_i'= 25$	$n_j'= 30$	$n_j'= 40$	$n_j'= 30$	$n_j'= 40$
					$n_i'= 20$	$n_i'= 20$	$n_i'= 25$	$n_i'= 25$
p:J\H	.01659	.01659	.01659	.01659	.01659	.01659	.01659	.01659
J\H	.00023	.00023	.00023	.00023	.00004	.00003	.00004	.00003
H\H	-	-	-	-	-	-	-	-
I:H\H	.00077	.00066	.00050	.00040	.00050	.00050	.00040	.00040
JH\H	-	-	-	-	-	-	-	-
pH:J\H	-	-	-	-	-	-	-	-
JI:H\H	.00004	.00003	.00002	.00002	<.00001	<.00001	<.00001	<.00001
pI:JH\H	.00611	.00529	.00440	.00317	.00440	.00440	.00317	.00317
$\hat{\sigma}^2(\tau)$	.01659	.01659	.01659	.01659	.01659	.01659	.01659	.01659
$\hat{\sigma}^2(\delta)$	.00611	.00529	.00440	.00317	.00440	.00440	.00317	.00317
$\hat{\sigma}^2(\Delta)$	.00715	.00621	.00515	.00382	.00494	.00493	.00361	.00360
$E\hat{\sigma}^2(\bar{X})$	.02270	.02188	.02099	.01976	.02099	.02099	.01976	.01975
$E\hat{\rho}^2$	.73	.75	.78	.83	.79	.79	.83	.83
$\hat{\sigma}^2(\bar{X})$ [a]	.00195	.00180	.00159	.00144	.00138	.00138	.00123	.00122
$\hat{\Phi}(\lambda)$ [b]	.67	.70	.74	.80	.75	.75	.81	.81
$\hat{\Phi}$	.70	.73	.76	.81	.77	.77	.82	.82

[a] Estimates assumed $n_p'=125$.

[b] Estimates assumed $\bar{X}=\lambda$.

Table 9

Comparison of Generalizability, Dependability, and Classical Reliability Coefficients

Coefficients	Grade	
Generalizability[a]		
px(I:H*)xO design	5	.80
	6	.76
(p:J)x(I:H*) design	5	.66
	6	.73
Classical reliability		
Cronbach's alpha	5	.73
	6	.73
KR-20	5	.73
	6	.73
KR-21	5	.62
	6	.65
Test-retest	5	.82
	6	.77
Dependability[b]		
px(I:H*)xO design	5	.72
	6	.68
(p:J)x(I:H*) design	5	.51
	6	.57

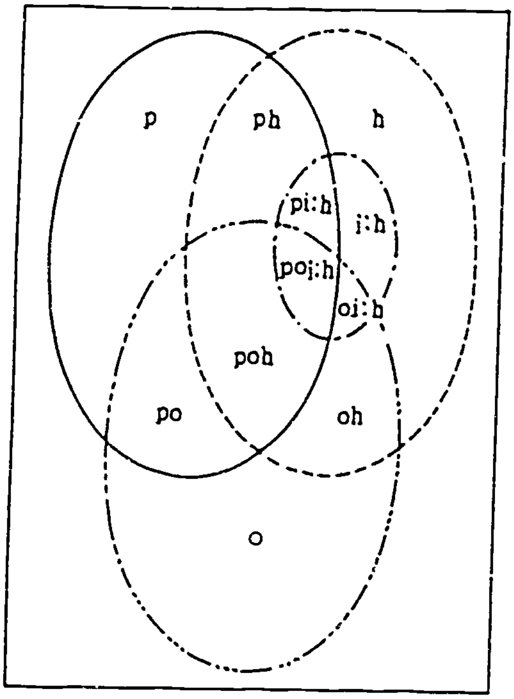
[a] Estimates of generalizability coefficients ($E\hat{\rho}^2$) were assumed that $n_i=n_i'$, $n_j=n_j'$, and $n_o=n_o'$.

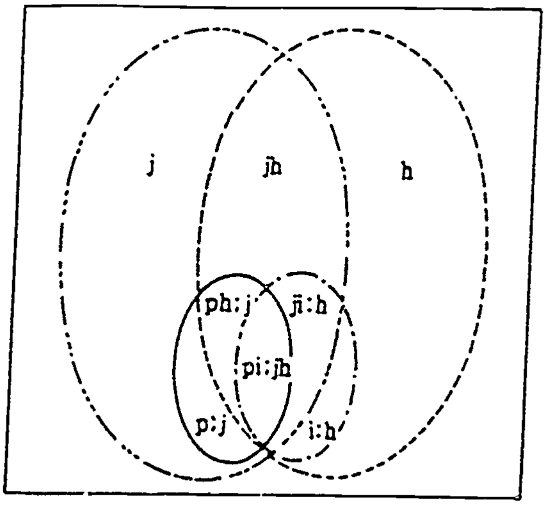
[b] Criterion score (λ) for dependability indices ($\Phi(\lambda)$) was the test mean score of the research subjects.

Figure Caption

Figure 1. Venn diagram for px(i:h*)xo design.

Figure 2. Venn diagram for (p:j)x(i:h*) design.


p
ph
h
pi:h
i:h
poi:h
oi:h
poh
po
oh
o


j
jh
h
ph:j
ji:h
pi:jh
p:j
i:h