DOCUMENT RESUME

ED 409 345 TM 026 828

AUTHOR	Motika, Robert T.
TITLE	Generalizability of Performance Assessment Measures on the Florida Teacher Certification Examinations.
PUB DATE	Mar 97
NOTE	28p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; *Error of Measurement; French; *Generalizability Theory; *Language Teachers; *Licensing Examinations (Professions); Second Language Instruction; Spanish; *Teacher Certification; *Test Reliability
IDENTIFIERS	Florida; Variance (Statistical)

ABSTRACT

Data from performance measures that were part of two foreign language teacher certification examinations were used in a generalizability study of the quality of their performance ratings. A total of 775 examinees from the Spanish K-12 and 192 examinees from the French K-12 subject area tests of the Florida Teacher Certification Examinations were selected. Data groups for both examinations were subdivided by unique rater pair combinations to form a series of fully crossed designed (person x rater x scale) with two random facets (person and rater) and one fixed facet (scale). Variance component estimates were then determined for each of the 31 examinee subgroups for Spanish and 8 subgroups for French. The means of the resulting variance component estimate distributions were used to assess the overall quality of the ratings and the relative measurement error associated with rater and scale facets. Separate partially nested designs (person:form x rater) were used to estimate variance associated with the forms facet. Results indicate that for both the Spanish and French examination data, universe score or person variance represented the largest single component of the total observed variance while the magnitude of the variance component estimated for facets associated with measurement error are small. The overall quality of these data for use in decisions as assessed by estimates of the index of dependability or phi coefficient was high. An appendix contains equations for the theoretical sampling variance of variance components. (Contains 14 tables and 5 references.) (Author/SLD)

Generalizability of Performance Assessment Measures on the Florida Teacher Certification Examinations

Robert T. Motika
Institute for Instructional Research and Practice
University of South Florida

Paper presented at the annual meeting of the
American Educational Research Association
Chicago, March 27, 1997

Generalizability of Performance Assessment Measures on the Florida Teacher Certification Examinations

Robert T. Motika
University of South Florida
Institute for Instructional Research and Practice

Abstract: Examinee data from performance measures comprising a portion of 2 foreign language teacher certification exams were used in a generalizability study of the quality of these performance ratings. A total of 775 examinees from the Spanish K-12 and 192 examinees from the French K-12 subject area tests of the Florida Teacher Certification Examinations were selected for inclusion. The total data groups for both exams were sub-divided by unique rater pair combinations to form a series of (*Person* × *Rater* × *Scale*) fully crossed designs with 2 random facets (*Person* and *Rater*) and 1 fixed facet (*Scale*). Variance component estimates were then determined using this model for each of the 31 examinee subgroups for Spanish and 8 examinee subgroups for French. The means of the resulting variance component estimate distributions were used to assess the overall quality of the ratings and the relative measurement error associated with rater and scale facets. Separate (*Person:Form* × *Rater*) partially nested designs were utilized to estimate variance associated with the forms facet. Results of the study indicate that for both the Spanish and French certification exam data, universe score or person variance represented the largest single component of the total observed variance while the magnitude of the variance component estimates for facets associated with measurement error were small. The overall quality of these data for use in absolute decisions as assessed by estimates of the index of dependability or phi coefficient was high ($\phi \geq .90$).

Purpose

This study applies Generalizability Theory to examine the technical quality of performance assessment ratings for two subject matter certification exams administered as part of the Florida Teacher Certification Examination (FTCE) program. A passing score on these examinations is a prerequisite to teacher certification in the state of Florida. The use of performance assessment in testing programs of all kinds has increased substantially over the past decade even as the technical quality of these performance assessment applications remained at issue. The need for quality performance measurement is particularly acute in high-stakes areas such as certification and licensure testing.

In its current form, the Florida Teacher Certification Examination (FTCE) comprises several sub-tests including: 1) a multiple-choice exam covering general pedagogical skill, 2) a test of basic skills in math and English, and 3) subject area specific exams (54 subject area certifications currently exist). This study examined performance assessments used as part of the French K-12 and Spanish K-12 subject area examinations. These foreign language certification exams are part of a small group of 8 subject area exams which incorporate performance assessments or non-traditional item prompts as part of the exam. The large majority of the subject area tests comprising the FTCE program (47 out of 54 subject area exams) consist solely of traditional multiple choice item formats.

The use of generalizability theory to examine the dependability of performance measures provides advantages over classical measurement theory estimates of score reliability. Generalizability theory permits the multiple sources of measurement error and their interactions to be assessed in a single analysis whereas classical test theory based procedures such as test-retest reliability correlation permit only the examination of a single error source. In performance assessment measurement situations, which commonly involve several facets or sources of measurement error (such as raters, rating scales, and testing prompts) the use of generalizability theory is clearly warranted. (Thompson, 1994).

This study attempts to assess the variability associated with the following major sources:

- variance associated with differences among raters (inter-rater reliability)
- variance associated with differences among the scale elements or categories which define the ratings (scale reliability)
- variance associated with different forms of the exams (coefficients of equivalence)
- variance associated with persons (true-score or universe variance)

Spanish K-12 Certification Examination

For the Spanish K-12 subject area test, successful examinees must demonstrate proficiency in speaking skills, listening skills, reading skills, knowledge of Hispanic culture, knowledge of grammar, and teaching techniques. Of these test elements, speaking skill is assessed by use of a performance assessment. The Spanish examination is divided into three sections. In the speaking section, examinees tape-record responses to items that are printed in the test book. In the listening section, examinees listen and respond to multiple-choice items presented on an audio tape. The multiple-choice section requires examinees to respond to multiple-choice items that require no supplemental materials.

The audio taped responses to 11 prompts are the examinee performances assessed in this test. After listening to all eleven examinee responses, raters are asked to assign a holistic rating to the entire group of responses for the following scale categories: Grammar, Vocabulary, Pronunciation, and Fluency. As seen in Figure 1, each of the 4 scales range from a low of 1 to a high score of 6 and are anchored by bipolar scale descriptions

GRAMMAR	Inaccurate	1	2	3	4	5	6	Accurate
VOCABULARY	Inadequate	1	2	3	4	5	6	Adequate
PRONUNCIATION	Non-Native	1	2	3	4	5	6	Native
FLUENCY	Broken	1	2	3	4	5	6	Smooth

Figure 1. Spanish Certification Exam Performance Scales.

French K-12 Certification Examination

For the French K-12 subject area test, speaking skills, listening skills, writing skills, knowledge of grammar, and teaching techniques are major test elements. Examinee speaking skill assessment is the test component examined in this study of performance ratings. In the speaking assessment section, examinees tape-record oral recitations of a printed passage, listen to a short passage describing a scenario and then record a plausible ending, and make up a short story (4 sentences or more) based on a series of simple pictures The audio taped responses to these prompts are the examinee performances assessed by a team of 2 raters. After listening to all examinee responses, raters are asked to assign holistic ratings to the entire group of responses for the following scale categories: Content, Grammar, Vocabulary, Pronunciation, and Fluency. As seen in Figure 2, the sub-scales range from a low of 1 to a high of 6 and are anchored by bipolar descriptions.

CONTENT	Inappropriate	1	2	3	4	5	6	Appropriate
GRAMMAR	Inaccurate	1	2	3	4	5	6	Accurate
VOCABULARY	Inadequate	1	2	3	4	5	6	Adequate
PRONUNCIATION	Non-Native	1	2	3	4	5	6	Native
FLUENCY	Broken	1	2	3	4	5	6	Smooth

Figure 2. French Certification Exam Performance Scales

Method

Since their inception in 1991, the Spanish and French subject area exams have been administered to a total of 1,672 and 397 teacher candidate examinees, respectively. From this data set, examinees were grouped by the unique ID codes of the 2 raters who supplied the holistic ratings for that individual. Of these rater combination groupings, 31 unique rater pair combinations were selected for Spanish each of which contained data for 25 or more examinees. For French, 8 rater pair groupings were selected each of which contained data for 24 or more examinees. From the Spanish subgroups, 25 examinees were randomly selected from within each of the 31 subgroups yielding a total sample size of 775 (25 × 31). From the French subgroups, 24 examinees were randomly selected from each of the 8 groups yielding a total sample size of 192 (24 × 8).

Performance data for both foreign language exams were analyzed first by use of a fully random (*Person × Rater × Scale*) fully crossed design. Variance components were estimated using this design for each of the 31 subgroups of Spanish examinees and for each of the 8 subgroups for French examinees. The resulting variance component estimates from the fully random model were then used to estimate the variance components for a mixed design consisting of two random facets (*Persons and Raters*) and one fixed facet (*Scale*) by averaging over conditions for the fixed facet. Variance components for the persons ($p*$), raters ($r*$) and the ($pr,e*$) interaction in the fixed model were estimated using equations 1-3.

$$\sigma^2_{p*} = \sigma^2_{p} + \frac{1}{n_s}\left(\sigma^2_{ps}\right) \qquad (1)$$

$$\sigma^2_{r*} = \sigma^2_{r} + \frac{1}{n_s}\left(\sigma^2_{rs}\right) \tag{2}$$

$$\sigma^2_{pr,e*} = \sigma^2_{pr} + \frac{1}{n_s}\left(\sigma^2_{prs,e}\right) \tag{3}$$

Relative error for each subgroup was calculated using equation 4. Absolute error for each subgroup was calculated using equation 5.

$$\sigma^2_{\mathrm{Re}l} = \frac{\sigma^2_{pr,e*}}{n'_r} \tag{4}$$

$$\sigma^2_{Abs} = \frac{\sigma^2_{r*}}{n'_r} + \frac{\sigma^2_{pr,e*}}{n'_r} \tag{5}$$

A phi coefficient or index of dependability for absolute decisions was estimated for each data subgroup using equation 6.

$$\phi = \frac{\sigma^2_p}{\left(\sigma^2_p + \sigma^2_{Abs}\right)} \tag{6}$$

To examine variance associated with different forms of the exams, a separate (*Person:Form × Rater)* partially nested design was used for both the Spanish and French data. In this design, all raters were treated as the same pair (rater #1 and rater #2, crossed with *Persons* and *Forms*) regardless of the actual rater ID. Reducing the rater facet in this way was considered justified in light of the small variance associated with the rater component found in the first design of this study. For Spanish , 200 examinee cases for each of 3 forms (A, B, and C) were randomly selected from the existing data. For French, 80 examinees for each of three forms were randomly selected. The resulting 600 Spanish and 240 French examinee cases were then used to estimate the magnitude of the 5 variance components for the nested design: *Form, Rater, Person:Form*, *Form × Rater*, and *Person:Form × Rater.*

Results

Performance assessment data were analyzed using the SAS VARCOMP procedure, using the default MIVQUE0 option for estimation of the design variance components. The variance components associated with the fully random ($p \times r \times s$) design were estimated for each of the 31 subgroups (n per subgroup = 25) for Spanish and each of the 8 subgroups (n per subgroup = 24) for French. The resulting variance component estimates for the fully random design are listed in Table 1 for the Spanish data and Table 8 for the French data. The variance components for the same samples using a mixed design (fixed facet = *Scale*), averaging over levels of the fixed facet, are shown in Table 2 (Spanish) and Table 9 (French). The variance component data from these 4 tables form distributions of variance component estimates from which mean estimates and standard errors may be calculated. Tables 3 and 4 show the means of the variance components and their standard errors for both the fully random and mixed models for Spanish. The means of the variance component estimates and their standard errors for the French data can be seen in Tables 10 and 11. Standard errors of the distributions are shown both for the set of subgroup means (labeled as empirical standard error) and for the theoretical standard error for each subgroup estimated by use of the formulas for theoretical sampling variance described by Smith (Smith, 1978). Theoretical sampling variance approximations are listed in the appendix.

As can be seen from the data for standard errors, the estimates for both the empirical and theoretical methods of estimation substantially agree.

The mean variance component estimates from all subgroups were used to assess the quality of the performance ratings. The summary of the variance component magnitudes for both random and fixed models can be seen in Table 5 (Spanish) and Table 12 (French). The data indicate that for both examination data sets, universe or person variance accounts for the largest single component of the total variance. Variance component estimates for facets associated with measurement error (such as the variance component for *Raters* and *Scale*) were small.

For the Spanish data, the mean absolute error variance was 0.058 and the mean phi coefficient was 0.923. The small magnitude of the absolute error variance and the correspondingly high value for the phi coefficient show that the performance ratings for Spanish reflect primarily universe score variance. The mean absolute error variance of 0.101 and the mean phi coefficient of 0.901 for the French data show that these ratings, likewise, reflect primarily universe score variance. The phi coefficients in these contexts can be viewed as generalizability coefficients for absolute decisions.

Variance Component magnitudes resulting from the partially nested (*Person:Form × Rater*) design seen in Table 7 (Spanish) and Table 14 (French) show that the magnitude of the *Forms* facet is also small--about 1% of the total variance. As with the fully crossed design, the partially nested design also indicates that person variance accounts for over 80% of the total variance of the model.

It should be noted, finally, that the raw performance data used in this study were taken from the ratings of judges before use of any arbitration procedures. In the scoring procedures for Spanish, for example, a third referee or arbitration rater is used to reduce disparity between rater scores whenever the absolute magnitude of the difference between the sums of the 2 raters' assigned scores equals or exceeds 5 points on the scale of 4 to 24 points. In French, an arbitration rater is used if the magnitude of the difference between the mean ratings per rater equals or exceeds ±2.0. Arbitration was used in only about 2% of the examinee cases in this data set. Since the raw performance data used in this study did not incorporate arbitration ratings, the estimates of the overall generalizability of the scores and the magnitude of the various error facets can be considered as lower bound estimates, as the use of arbitration should only decrease error variance.

Conclusions

The evidence produced by this generalizability theory based analysis indicates that these performance ratings for certification examinations reflect primarily universe score variance, that is, variance actually associated with differences among examinees in knowledge and proficiency in the domains of French and Spanish. Variance associated with potential major sources of measurement error such as differences among raters, scales, or test forms does not seem to be of a large enough magnitude to constitute a serious threat to the generalizability of these ratings. Task related variance, which has been found to be a significant problem in other studies of the generalizability of performance ratings (e.g. Shavelson, et. al 1993) does not seem to be a threat to the quality of ratings in this case (assuming that task variance would become manifest as *Forms* related variance in this design). Overall generalizability of these performance ratings, as assessed by the phi coefficient, was high enough (> .90) to warrant confidence in their use for certification decisions.

Table 1

Spanish Certification Examination Performance Data

Estimated Variance Components for Fully Random *(p × r × s)* Design for 31 Unique Rater Pairs

Unique Rater Pair	Persons *(p)* σ_p^2	Raters *(r)* σ_r^2	Scale *(s)* σ_s^2	*pr* σ_{pr}^2	*ps* σ_{ps}^2	*rs* σ_{rs}^2	*prs,e* $\sigma_{prs,e}^2$
1	0.67500	0.00278	0.033333	0.00722	0.00667	0.02556	0.17944
2	0.70694	0.00	0.049861	0.04042	0.02014	0.06292	0.21875
3	0.68347	0.00	0.062639	0.00542	0.08069	0.00167	0.15000
4	0.72750	0.00	0.009167	0.04569	0.01083	0.00444	0.26222
5	0.72528	0.01139	0.027361	0.07194	0.06597	0.02986	0.24681
6	0.79958	0.00	0.009583	0.12403	0.05375	0.02403	0.15097
7	0.75347	0.00472	0.055556	0.09194	0.11111	0.00653	0.12181
8	0.53958	0.11153	0.000000	0.04181	0.16000	0.03806	0.22861
9	0.78097	0.02222	0.012639	0.04111	0.06403	0.00361	0.16806
10	0.60417	0.00014	0.004167	0.05319	0.00	0.00111	0.24556
11	0.56181	0.02639	0.00	0.09694	0.16778	0.02778	0.19889
12	0.95069	0.00597	0.056111	0.06736	0.06722	0.00	0.18972
13	0.45611	0.00	0.019444	0.06583	0.03389	0.00	0.21250
14	0.73472	0.03667	0.005139	0.01333	0.00	0.02375	0.18625
15	0.35306	0.00	0.056806	0.01903	0.12653	0.01319	0.16847
16	0.52778	0.00	0.051528	0.09667	0.05181	0.01000	0.15500
17	0.84389	0.00542	0.030556	0.03125	0.04611	0.00	0.15750
18	0.60903	0.00	0.004861	0.01875	0.09181	0.07583	0.25750
19	0.71708	0.01514	0.017083	0.03153	0.11292	0.03153	0.21681
20	1.13444	0.00	0.057778	0.05694	0.14222	0.02361	0.19639
21	0.59417	0.03556	0.00	0.07111	0.10625	0.01944	0.17389
22	1.25361	0.00	0.025278	0.04611	0.01806	0.00	0.20806
23	0.47444	0.00069	0.017361	0.02931	0.06597	0.00181	0.17319
24	1.19236	0.00	0.009861	0.11194	0.08681	0.01069	0.15097
25	0.87417	0.00	0.083333	0.05319	0.16000	0.01153	0.17181
26	0.87972	0.00	0.00	0.06486	0.06111	0.11528	0.25639
27	0.54333	0.00111	0.090417	0.01556	0.02958	0.00	0.20569
28	0.50903	0.00	0.001111	0.07236	0.11889	0.00694	0.15139
29	0.61319	0.00444	0.059444	0.16556	0.00	0.00431	0.23069
30	0.59583	0.00	0.082083	0.05736	0.18792	0.00	0.17681
31	0.58319	0.00389	0.00	0.01944	0.02431	0.11611	0.27556

Note. Negative variance estimates set to 0.00.

Table 2

Spanish Certification Examination Performance Data

Estimated Variance Components for Mixed ($p \times r \times s$) Design for 31 Unique Rater Pairs

	σ_p^2		σ_r^2		σ_{pr}^2	
	Estimated Variance Component	Percent of Total Variance	Estimated Variance Component	Percent of Total Variance	Estimated Variance Component	Percent of Total Variance
1	0.67667	92	0.00917	1	0.05208	7
2	0.71198	88	0.00	0	0.09510	12
3	0.70365	94	0.00	0	0.04292	6
4	0.73021	87	0.00	0	0.11125	13
5	0.74177	83	0.01885	2	0.13365	15
6	0.81302	83	0.00198	0	0.16177	17
7	0.78125	86	0.00635	1	0.12240	13
8	0.57958	72	0.12104	15	0.09896	12
9	0.79698	88	0.02312	3	0.08313	9
10	0.60146	84	0.00042	0	0.11458	16
11	0.60375	77	0.03333	4	0.14667	19
12	0.96750	89	0.00521	0	0.11479	11
13	0.46458	80	0.00	0	0.11896	21
14	0.72927	88	0.04260	5	0.05990	7
15	0.38469	87	0.00	0	0.06115	14
16	0.54073	80	0.00	0	0.13542	20
17	0.85542	92	0.00438	0	0.07063	8
18	0.63198	89	0.00	0	0.08313	12
19	0.74531	87	0.02302	3	0.08573	10
20	1.17000	92	0.00	0	0.10604	8
21	0.62073	80	0.04042	5	0.11458	15
22	1.25813	93	0.00	0	0.09813	7
23	0.49094	87	0.00115	0	0.07260	13
24	1.21406	89	0.00	0	0.14969	11
25	0.91417	91	0.00	0	0.09615	10
26	0.89500	88	0.00	0	0.12896	13
27	0.55073	89	0.00052	0	0.06698	11
28	0.53875	84	0.00	0	0.11021	17
29	0.60833	73	0.00552	1	0.22323	27
30	0.64281	87	0.00	0	0.10156	14
31	0.58927	83	0.03292	5	0.08833	12

Note. Negative variance estimates set to 0.00.

Table 3

Spanish Certification Examination Performance Data

Mean Variance Components and Standard Errors for Random $(p \times r \times s)$ Design for 31 Unique Rater Pairs

Source of Variation	*Mean*	*Empirical Standard Error*	*Theoretical Standard Error*
σ_{p}^{2}	0.7096	0.2106	0.2259
σ_{r}^{2}	0.0051	0.0245	0.0238
σ_{s}^{2}	0.0282	0.0313	0.0413
σ_{pr}^{2}	0.0557	0.0368	0.0315
σ_{ps}^{2}	0.0716	0.0572	0.0331
σ_{rs}^{2}	0.0217	0.0313	0.0242
$\sigma_{prs,e}^{2}$	0.1963	0.0399	0.0327

Note. Negative variance component estimates included in calculations of means.

Table 4

Spanish Certification Examination Performance Data

Mean Variance Component Estimates and Standard Errors for Mixed ($p \times r \times s$) Design for 31 Unique Rater Pairs

Source of Variation	*Mean*	*Standard Error*
σ^2_p	0.7275	0.2104
σ^2_r	0.0106	0.0249
$\sigma^2_{pr,e}$	0.1048	0.0366

Note. Negative variance component estimates included in calculations of means.

Table 5

Spanish Certification Examination Performance Data

Summary Analysis of *(p × r × s)* Design for Both Random and Mixed Models

Random Design			*Mixed Design (Fixed Facet = Scale)*			
Source	*Variance Component*	*Estimated Value*	*Source*	*Variance Component*	*Estimated Value*	*Percent of Total Variance*
Examinees *(p)*	σ_p^2	0.7096	Examinees *(p)*	σ_p^2	0.7275	86%
Raters *(r)*	σ_r^2	0.0051	Raters *(r)*	σ_r^2	0.0106	1.3%
Scale *(s)*	σ_s^2	0.0282				
pr	σ_{pr}^2	0.0557		$\sigma_{pr,e}^2$	0.1048	12.9%
ps	σ_{ps}^2	0.0716				
rs	σ_{rs}^2	0.0217				
prs,e	$\sigma_{prs,e}^2$	0.1963				

Table 6

Spanish Certification Examination Performance Data

Relative Error, Absolute Error and Phi Coefficients for Mixed $(p \times r \times s)$ Design for 31 Unique Rater Pairs

Relative Error $\left(\sigma^2_{\mathrm{Re}l}\right)$	Absolute Error $\left(\sigma^2_{Abs}\right)$	Phi Coefficient (ϕ)
0.02604	0.03063	0.95670
0.04755	0.04688	0.93823
0.02146	0.02063	0.97152
0.05562	0.05500	0.92995
0.06682	0.07625	0.90679
0.08089	0.08188	0.90851
0.06120	0.06437	0.92387
0.04948	0.11000	0.84048
0.04156	0.05312	0.93751
0.05729	0.05750	0.91274
0.07333	0.09000	0.87027
0.05740	0.06000	0.94161
0.05948	0.05750	0.88986
0.02995	0.05125	0.93434
0.03057	0.02938	0.92906
0.06771	0.06563	0.89177
0.03531	0.03750	0.95800
0.04156	0.04000	0.94047
0.04286	0.05438	0.93200
0.05302	0.05250	0.95706
0.05729	0.07750	0.88900
0.04906	0.04750	0.96362
0.03630	0.03688	0.93014
0.07484	0.07187	0.94411
0.04807	0.04625	0.95184
0.06448	0.06312	0.93412
0.03349	0.03375	0.94226
0.05510	0.05313	0.91024
0.11161	0.11438	0.84174
0.05078	0.04875	0.92951
0.04417	0.06062	0.90672

Table 7

Spanish Certification Examination Performance Data

Summary Analysis of *((p: f) × r)* Partially Nested Design

Source	*Variance Component*	*Estimated Value*	*Percent of Total Variance*
Forms *(f)*	σ^2_f	0.0715	1.0%
Raters *(r)*	σ^2_r	0.0037	0.0%
Persons:Forms *(p:f)*	$\sigma^2_{p,pf}$	11.3425	84.0%
fr	σ^2_{fr}	0.00*	0.0%
p:fr,e	$\sigma^2_{pr,fpr,e}$	2.0747	15.0%

Table 8

French Certification Examination Performance Data

Estimated Variance Components for Fully Random *(p × r × s)* Design for 8 Unique Rater Pairs

Unique Rater Pair	Persons *(p)*	Raters *(r)*	Scale *(s)*	*pr*	*ps*	*rs*	*prs,e*
	σ^2_p	σ^2_r	σ^2_s	σ^2_{pr}	σ^2_{ps}	σ^2_{rs}	$\sigma^2_{prs,e}$
1	0.85679	0.00	0.017935	0.25534	0.15082	0.037772	0.35806
2	1.04447	0.00	0.040217	0.05217	0.23270	0.037319	0.25435
3	0.94538	0.00	0.030163	0.14339	0.11984	0.00	0.18089
4	1.21522	0.002808	0.013134	0.14719	0.00145	0.00	0.20734
5	0.65942	0.00	0.029710	0.06658	0.12029	0.003804	0.31911
6	0.76902	0.006884	0.082971	0.07853	0.12120	0.005435	0.31748
7	0.89058	0.00	0.005797	0.28351	0.07754	0.00	0.25996
8	0.89737	0.00	0.037772	0.18062	0.02473	0.014764	0.20399

Note. Negative variance estimates replaced by 0.00.

Table 9

French Certification Examination Performance Data

Estimated Variance Components for Mixed *(p* × *r* × *s)* Design for 8 Unique Rater Pairs

	σ^2_p		σ^2_r		σ^2_{pr}	
	Estimated Variance Component	Percent of Total Variance	Estimated Variance Component	Percent of Total Variance	Estimated Variance Component	Percent of Total Variance
1	0.88696	73	0.00638	1	0.32696	27
2	1.09101	92	0.00	0	0.10304	9
3	0.96935	85	0.00	0	0.17957	16
4	1.21551	86	0.0022	0	0.18866	13
5	0.68348	84	0.00	0	0.13040	16
6	0.79326	84	0.00797	1	0.14203	15
7	0.90609	73	0.00	0	0.33551	27
8	0.90232	81	0.00	0	0.22141	20

Note. Negative variance estimates replaced by 0.00.

Table 10

French Certification Examination Performance Data

Mean Variance Components and Standard Errors for Random ($p \times r \times s$) Design for 8 Unique Rater Pairs

Source of Variation	*Mean*	*Empirical Standard Error*	*Theoretical Standard Error*
σ^2_p	0.9098	0.1684	0.3063
σ^2_r	-0.0035	0.0064	0.0119
σ^2_s	0.0322	0.0237	0.0350
σ^2_{pr}	0.1509	0.0858	0.0606
σ^2_{ps}	0.1060	0.0727	0.0403
σ^2_{rs}	0.0113	0.0173	0.0159
$\sigma^2_{prs,e}$	0.2626	0.0639	0.0387

Note. Negative variance component estimates included in calculations of means.

Table 11

French Certification Exam Performance Data

Mean Variance Component Estimates and Standard Errors for Mixed ($p \times r \times s$) Design for 8 Unique rater Pairs

Source of Variation	*Mean*	*Standard Error*
σ_{p}^{2}	0.9310	0.1655
σ_{r}^{2}	-0.0013	0.0063
$\sigma_{pr,e}^{2}$	0.2034	0.0870

Note. Negative variance component estimates included in calculations of means.

Table 12

French Certification Exam Performance Data

Summary Analysis of *(p × r × s)* Design for Both Random and Mixed Models

Random Design			*Mixed Design (Fixed Facet = Scale)*			
Source	*Variance Component*	*Estimated Value*	*Source*	*Variance Component*	*Estimated Value*	*Percent of Total Variance*
Examinees *(p)*	σ_p^2	0.9098	Examinees *(p)*	σ_p^2	0.9310	82%
Raters *(r)*	σ_r^2	0.00*	Raters *(r)*	σ_r^2	0.00*	0%
Scale *(s)*	σ_s^2	0.0322				
pr	σ_{pr}^2	0.1509	*pr,e*	$\sigma_{pr,e}^2$	0.2034	17%
ps	σ_{ps}^2	0.1061				
rs	σ_{rs}^2	0.0113				
prs,e	$\sigma_{prs,e}^2$	0.2626				

* *Negative Estimate set to 0.00*

Table 13

French Certification Examination Performance Data

Relative Error, Absolute Error and Phi Coefficients for Mixed $(p \times r \times s)$ Design for 8 Unique Rater Pairs

Relative Error $\left(\sigma^2_{\mathrm{Re}l}\right)$	Absolute Error $\left(\sigma^2_{Abs}\right)$	Phi Coefficient (ϕ)
0.1635	0.1667	0.8418
0.0500	0.0500	0.9562
0.0898	0.0867	0.9179
0.0943	0.0954	0.9272
0.0652	0.0646	0.9137
0.0710	0.0750	0.9136
0.1677	0.1642	0.8466
0.1107	0.1063	0.8947

Table 14

French Certification Examination Performance Data

Summary Analysis of *((p: f) × r)* Partially Nested Design

Source	*Variance Component*	*Estimated Value*	*Percent of Total Variance*
Forms *(f)*	σ^2_f	0.00	0.0%
Raters *(r)*	σ^2_r	0.0781	0.0%
Persons:Forms *(p:f)*	$\sigma^2_{p,pf}$	24.9672	82.0%
fr	σ^2_{fr}	0.0601	0.0%
p:fr,e	$\sigma^2_{pr,fpr;e}$	5.4409	18.0%

References

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

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30,* 215-232.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A primer.* Newbury Park, CA: SAGE.

Smith, P. L. (1978). Sampling errors of variance components in small sample multifacet generalizability studies. *Journal of Educational Statistics, 3*(4) , 319-346.

Thompson, B. , & Crowley, S. L. (1994, April). *When classical measurement theory is insufficient and generalizability theory is essential.* Paper presented at the annual meeting of the Western Psychological Association, Kailua-Kona, Hawaii.

Appendix

Theoretical Sampling Variance of Variance Components for *(p × r × s)* Fully Crossed Random Design (Smith, 1978)

$$\text{Variance } \sigma_p^2 = \frac{2}{n_p - 1} \times \left[\left(\sigma_p^2 + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{ps}^2}{n_s} + \frac{\sigma_{prs}^2}{n_s n_r} \right)^2 + \frac{1}{(n_r - 1)} \left(\frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{prs,e}^2}{n_s n_r} \right)^2 + \frac{1}{(n_s - 1)} \left(\frac{\sigma_{ps}^2}{n_s} + \frac{\sigma_{prs,e}^2}{n_s n_r} \right)^2 + \frac{1}{(n_s - 1)(n_r - 1)} \left(\frac{\sigma_{prs,e}^2}{n_s n_r} \right)^2 \right]$$

$$\text{Variance } \sigma_s^2 = \frac{2}{n_s - 1} \times \left[\left(\sigma_s^2 + \frac{\sigma_{ps}^2}{n_p} + \frac{\sigma_{rs}^2}{n_p} + \frac{\sigma_{prs}^2}{n_p n_r} \right)^2 + \frac{1}{(n_p - 1)} \left(\frac{\sigma_{ps}^2}{n_p} + \frac{\sigma_{prs,e}^2}{n_p n_r} \right)^2 + \frac{1}{(n_r - 1)} \left(\frac{\sigma_{rs}^2}{n_r} + \frac{\sigma_{prs,e}^2}{n_p n_r} \right)^2 + \frac{1}{(n_p - 1)(n_r - 1)} \left(\frac{\sigma_{prs,e}^2}{n_p n_r} \right)^2 \right]$$

$$\text{Variance } \sigma_r^2 = \frac{2}{n_r - 1} \times \left[\left(\sigma_r^2 + \frac{\sigma_{pr}^2}{n_p} + \frac{\sigma_{rs}^2}{n_s} + \frac{\sigma_{prs}^2}{n_p n_s} \right)^2 + \frac{1}{(n_p - 1)} \left(\frac{\sigma_{pr}^2}{n_p} + \frac{\sigma_{prs,e}^2}{n_p n_s} \right)^2 + \frac{1}{(n_s - 1)} \left(\frac{\sigma_{rs}^2}{n_s} + \frac{\sigma_{prs,e}^2}{n_p n_s} \right)^2 + \frac{1}{(n_p - 1)(n_s - 1)} \left(\frac{\sigma_{prs,e}^2}{n_p n_s} \right)^2 \right]$$

$$\text{Variance } \sigma^2_{pr} = \frac{2}{(n_p-1)(n_r-1)} \times \left[\left(\sigma^2_{pr} + \frac{\sigma^2_{prs,e}}{n_s}\right)^2 + \frac{1}{(n_s-1)}\left(\frac{\sigma^2_{prs,e}}{n_s}\right)^2\right]$$

$$\text{Variance } \sigma^2_{ps} = \frac{2}{(n_p-1)(n_s-1)} \times \left[\left(\sigma^2_{ps} + \frac{\sigma^2_{prs,e}}{n_r}\right)^2 + \frac{1}{(n_r-1)}\left(\frac{\sigma^2_{prs,e}}{n_r}\right)^2\right]$$

$$\text{Variance } \sigma^2_{rs} = \frac{2}{(n_s-1)(n_r-1)} \times \left[\left(\sigma^2_{rs} + \frac{\sigma^2_{prs,e}}{n_p}\right)^2 + \frac{1}{(n_p-1)}\left(\frac{\sigma^2_{prs,e}}{n_p}\right)^2\right]$$

$$\text{Variance } \sigma^2_{prs,e} = \left(\frac{2}{(n_p-1)(n_r-1)(n_s-1)}\right)\sigma^4_{prs,e}$$

AERA CONFERENCE 1997

TM026828

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: GENERALIZABILITY OF PERFORMANCE ASSESSMENT MEASURES ON THE FLORIDA TEACHER CERTIFICATION EXAMINATIONS

Author(s): ROBERT T. MOTIKA

Corporate Source: UNIVERSITY OF SOUTH FLORIDA

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☑

Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all **Level 1** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all **Level 2** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

□

Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature:

Printed Name/Position/Title: ROBERT T. MOTIKA STATISTICIAN

Organization/Address:
UNIVERSITY OF SOUTH FLORIDA
HMS 401
4202 EAST FOWLER AVE.
TAMPA FL 33620

Telephone: (813) 974-3700

FAX: (813) 974-5132

E-Mail Address: BOB@iirp.coedu.usf.edu

Date: 3-27-97

(over)