G Theory Project: Math Placement Exam Data Analyzed Via Multivariate Generalizability Theory

Brad Thiessen

07P:455 – Generalizability Theory

Introduction

In order to determine the mathematics knowledge and ability of new students, most incoming freshmen [redacted] are administered a short math placement test. The results from this test are used to select an appropriate math course for these new students. These courses cost money, so it is important that students be accurately placed in the appropriate course. In this study, data from an unusual 1999 administration of this math placement exam will be analyzed via Generalizability Theory to determine the optimal number of items needed in order to make efficient, accurate placement decisions.

Data Description

[redacted]. The placement exam was specified to have 30 dichotomously-scored items split evenly into three content categories: Algebra, Geometry, and Functions. The examinees, who identified their intended plans-of-study on the test, fell into three groups: 50 students intended to major in a scientific field (mathematics, biology, chemistry, etc.); 51 students intended to major in the social sciences (psychology, sociology, education, etc.); and 57 students indicated that they were undecided about their intended major.

[redacted], the students were administered the same test later [redacted].

G-Study Design

The data for this analysis consist of the 0/1 responses of 150 students to two administrations of a 30-item test. Students, the objects of measurement, were nested within the three types of majors. To balance the design, one social science major and seven undeclared majors were randomly eliminated from the study.

The 50 students nested within each major all took the same exam twice. Thus, the students were crossed with respect to the two administration occasions. Furthermore, each test administration consisted of the same three content categories with 10 items nested within each.

If the conditions within each facet are treated as being random samples from all possible conditions within each facet, and we let:

s = students,

m = major categories (science, social science, undecided),

o = occasions,

i = items, and

c = content categories (Algebra, Geometry, Functions),

then this data follows an **(s:m)** x **o** x **(i:c)** design, with the following observed sample sizes:

$n_s = 50$,

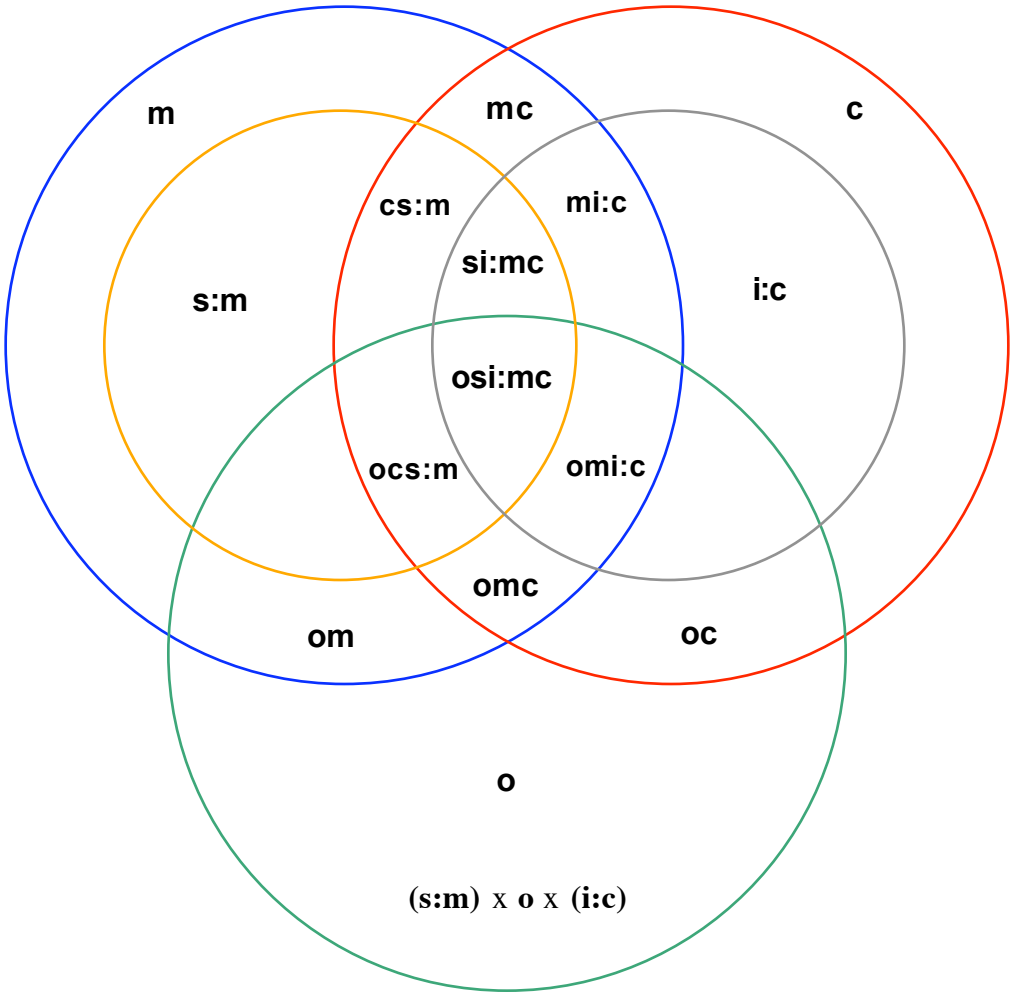
$n_m = 3$,

$n_o = 2$,

$n_i = 10$, and

$n_c = 3$.

The following Venn diagram displays this design along with its linear model and variance components.



Linear Model: $X_{simco} = \mu + \nu_m + \nu_c + \nu_o + \nu_{s:m} + \nu_{i:c} +$ (5 main effects)

$\nu_{om} + \nu_{oc} + \nu_{mc} + \nu_{omc} + \nu_{os:m} + \nu_{mi:c} + \nu_{cs:m} + \nu_{oi:c} +$

$\nu_{oc:sm} + \nu_{si:mc} + \nu_{moi:c} + \nu_{ois:mc}$ (12 interaction effects)

Variance Components: $\sigma^2(m) + \sigma^2(c) + \sigma^2(o) + \sigma^2(s;m) + \sigma^2(i:c) + \sigma^2(om) + \sigma^2(oc) +$

$\sigma^2(mc) + \sigma^2(omc) + \sigma^2(os:m) + \sigma^2(mi:c) + \sigma^2(cs:m) + \sigma^2(oi:c) +$

$\sigma^2(oc:sm) + \sigma^2(si:mc) + \sigma^2(moi:c) + \sigma^2(ois:mc)$

Note: Fixed versus Random Facets

While defining a facet as fixed or random is required for the D-Study, some discussion is warranted here. Whereas the student and test item facets are obviously random facets (sampled from a much larger population of students and test items), the other facets are not so clear. As will be seen later, the occasion facet will be treated as though it is random – it is sampled from a large population of possible occasions and we will want to generalize over occasions. The student major facet is also random, due to the way in which all possible majors were split into three artificial categories (all possible majors could have been split into other categories). The test content categories, on the other hand, can be considered to be a fixed facet. [redacted].

G-Study Attempt #1: GENOVA

In the first attempted G-Study, all facets were treated as being random and the data were formatted for GENOVA. The research could not get GENOVA to analyze the data without crashing – apparently due to the fact that students were nested within majors. When the student major facet was ignored, GENOVA was able to estimate the variance components. The researcher did not want to ignore the student major facet without reason, so the GENOVA results were ignored.

G-Study Attempt #2: urGENOVA

While urGENOVA has the requirement that all facets must be random (Brennan, 2001c), it was able to analyze the data in this **(s:m) x o x (i:c)** design. The following control cards yielded the results in Tables 1 and 2.

```
urGENOVA Control Cards

GSTUDY    EXAMPLE: (P:C) X O X (I:R) DESIGN
OPTIONS   NREC 4  "*.out" EMS TIME SECI .8
EFFECT      M     3
EFFECT    * S:M   50 50 50
EFFECT      O     2
EFFECT      C     3
EFFECT      I:C   10 10 10
FORMAT    0 2
PROCESS   "data"
```

Table 1: Group Means

Social Science Major	0.476
Science Major	0.523
Undecided Major	0.496
Student (lowest)	0.100
Student (highest)	0.933
Occasion 1	0.483
Occasion 2	0.513

	Algebra	**Geometry**	**Functions**
Item 1	0.763	0.780	0.733
Item 2	0.827	0.787	0.710
Item 3	0.707	0.567	0.700
Item 4	0.530	0.553	0.590
Item 5	0.560	0.643	0.387
Item 6	0.483	0.447	0.377
Item 7	0.467	0.253	0.247
Item 8	0.380	0.417	0.280
Item 9	0.267	0.350	0.357
Item 10	0.263	0.193	0.330
Content Mean	0.525	0.499	0.471

Perhaps unsurprisingly, the students intending to major in a scientific field had the highest average score among the three types of majors. Another unsurprising result is that students overall did slightly better on the second administration of the placement exam (when they were given the full 40-minutes to answer the 30-items). We are not interested in the mean scores, however; we are interested in the estimated variance components for this **(s:m) x o x (i:c)** design.

Table 2 displays the estimated values of these variance components. The variance component estimates were calculated using Henderson's Method #1 (Brennan, 2001c).

Table 2: G-Study Results from urGENOVA

Effect	df	SS	MS	Variance	SE	80% Confidence
m	2	3.43756	1.71878	**-0.00024**	0.00042	(0.0000, 0.0046)
s:m	147	352.86733	2.40046	**0.03369**	0.00466	(0.0284, 0.0406)
o	1	2.05511	2.05511	**0.00045**	0.00037	(0.0002, 0.0289)
c	2	4.32289	2.16145	**-0.00314**	0.00114	(0.0000, 0.0029)
i:c	27	314.19533	11.63686	**0.03737**	0.01019	(0.0271, 0.0564)
mo	2	0.10156	0.05078	**0.00000**	0.00003	(0.0000, 0.0003)
mc	4	1.55711	0.38928	**0.00004**	0.00024	(0.0000, 0.0011)
mi:c	54	17.25867	0.31961	**-0.00026**	0.00061	(0.0000, 0.0007)
so:m	147	5.37667	0.03658	**0.00046**	0.00015	(0.0003, 0.0007)
sc:m	294	107.38667	0.36526	**0.00083**	0.00155	(0.0000, 0.0030)
si:mc	3969	1361.94600	0.34315	**0.16299**	0.00386	(0.1582, 0.1681)
oc	2	0.05356	0.02678	**-0.00008**	0.00003	(0.0000, 0.0001)
oi:c	27	3.39800	0.12585	**0.00071**	0.00022	(0.0005, 0.0011)
moc	4	0.14778	0.03695	**0.00002**	0.00004	(0.0000, 0.0002)
moi:c	54	1.04400	0.01933	**0.00004**	0.00007	(0.0000, 0.0002)
soc:m	294	6.66533	0.02267	**0.00055**	0.00019	(0.0003, 0.0008)
soi:mc	3969	68.15800	0.01717	**0.01717**	0.00039	(0.0167, 0.0177)
Total	8999	2249.97156				

The column with bold type displays the estimated variance components for this G-Study design. Note that some of the estimated variances are negative due to the fact that these are variance *estimates* and, therefore, are subject to sampling variability. It appears as though most of the variance is due to the combination of students within majors, items within content classifications, and the interactions among students and items. The standard errors of the variance components (SE) and 80% Confidence Interval boundaries displayed in Table 2 are calculated using the Ting et. al (1990) procedure (Brennan, 2001c). This procedure requires the assumption that the score effects follow a normal distribution. Since the data in this study are dichotomously scored items, this assumption is not met. So while the confidence interval boundaries are not appropriate for this data, they are displayed to provide at least a rough idea about the real values of the variance components.

Recall that the purpose of this study was to determine the optimal number of items needed in a mathematics placement test in order to make accurate placement decisions. This purpose necessitates a D-Study. Unfortunately, urGENOVA does not provide D-Study results (Brennan, 2001c). Even if urGENOVA provided D-Study results, they would not be of interest in this study. In this study, the

content categories of Algebra, Geometry, and Functions are fixed, giving us a mixed model. urGENOVA only deals with random designs.

G-Study Attempt #3: mGENOVA

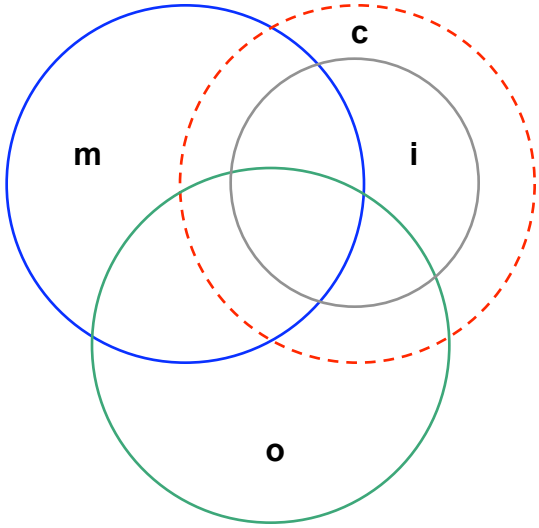
In order to complete a D-Study with a fixed content facet, the data will need to be analyzed via multivariate Generalizability Theory using mGENOVA. Unfortunately, mGENOVA only works with five canonical study structures (Brennan, 2001b). This study's **(s:m) x o x (i:c)** design will not work with mGENOVA.

If the m facet (student majors) can be ignored, then this data does match up with one of mGENOVA's canonical structures. From a univariate perspective, the data can then be characterized as an **s x o x (i:c)** design with c being a fixed content facet. Based on the negative variance estimate of the m facet from urGENOVA (and the fact that scores are never reported for student major groups), the researcher believes the majors facet can be ignored. In other words, students will no longer be considered to be nested within majors.

Ignoring the major facet, this study can now be characterized as a multivariate $s^{\bullet}$ x $o^{\bullet}$ x i° design, where items differ across the content categories of Algebra, Geometry, and Functions. The following control cards were used to analyze the data via mGENOVA. The G-Study design is displayed in a Venn diagram on the following page.

```
mGENOVA Control Cards

GSTUDY    s X o X (i:c) Design
OPTIONS   NREC 2  "*.out"
MULT      3 Algebra Geometry Functions
EFFECT     * p    150 150 150
EFFECT    # o    2 2 2
EFFECT      i    10 10 10
FORMAT    2 2
PROCESS   "projdata"
```



In addition to calculating the observed means and mean squares/products, mGENOVA provided the following G-Study variance component estimates:

Table 3: mGENOVA G-Study Results			
Variance, Covariance, & Correlation estimates			
	Algebra	Geometry	Functions
student	**0.03449**	0.91954	1.01051
	0.03311	**0.03758**	1.00645
	0.03309	0.03440	**0.03108**
occasion	**0.00047**		
	0.00039	**0.00017**	
	0.00056	0.00039	**0.00050**
item	**0.03668**		
		0.04014	
			0.03503
so	**0.00161**		
	0.00038	**0.00020**	
	0.00064	0.00037	**0.00128**
si	**0.16184**		
		0.16379	
			0.16282
oi	**0.00068**		
		0.00092	
			0.00058
soi	**0.01792**		
		0.01171	
			0.02198

Looking at the variance component estimates (the bold numbers in the above table), we once again see that the item facet, the student facet, and the student-item interaction facet have the largest variance estimates. This tells us that the items differed in difficulty and students differ in ability. The occasion facet, once again, had a very small estimated variance.

Estimates of the disattenuated correlations are reported above the diagonal of the table. One can see that two of the reported correlations are above 1.0. According to Brennan (2001a), correlations above 1.0 can indicate (a) sampling error, (b) small sample sizes, or (c) hidden facets in the G-Study design.

In order to complete this analysis, a series of D-Studies must be conducted. Before running a D-Study, the researcher must specify: (a) the design of the D-Study, (b) the nature of the facets (fixed versus random), and (c) the D-Study sample sizes. For this analysis, the researcher is interested in the true score and error variances for a design in which students, on one occasion, take a test consisting of Algebra, Geometry, and Functions questions. As a univariate design, it would be characterized as an **s x o x (i:c)** design. As a multivariate design, it is characterized as an $s^{\bullet}$ x $o^{\bullet}$ x i° design.

Since the researcher obviously wants to generalize over students and test items, those facets will be defined as being random. For those two facets, the researcher will keep the same sample sizes as were used in the G-Study. Because students in the future will only take the exam on one occasion, the D-Study sample size for occasions will be 1. This single occasion will be treated as a random facet, so the researcher will be able to generalize over occasions. The content facet will be treated as being fixed, as is required under multivariate Generalizability theory.

The following control cards were entered into mGENOVA to determine true score and error variances for this D-Study with 150 students, 1 occasion, and 10 items classified into three fixed content categories.

```
mGENOVA D-Study Control Cards

DSTUDY   One occasion 10 items
DEFFECT  $ p 150 150 150
DEFFECT  # O 1 1 1
DEFFECT    I 10 10 10
ENDDSTUDY
```

Table 4 displays the results from this D-Study. Since decisions are made based on the student's total score on the test, the results for the *composite score* are of most interest. Furthermore, the composite score is the simple sum of the 30-item scores on the exam, the composite score can be calculated using the following weights:

$$\text{Composite Score} = \frac{1}{3}(\text{Algebra score}) + \frac{1}{3}(\text{Geometry score}) + \frac{1}{3}(\text{Functions score}).$$

Table 4: D-Study Results from mGENOVA (10 items within each content category)

	Algebra	Geometry	Functions	Composite
Universe Score Variance	0.03449	0.03758	0.03108	0.03382
Relative Error Variance	0.01958	0.01775	0.01976	0.00665
Absolute Error Variance	0.02379	0.02203	0.02382	0.00834
Mean Error Variance	0.00456	0.00465	0.00440	0.00196
Universe SD	0.18572	0.19386	0.17630	0.18389
Relative SD	0.13994	0.13322	0.14057	0.08156
Absolute SD	0.15422	0.14841	0.15433	0.09135
Error SD for Mean	0.06754	0.06816	0.06631	0.04428
Generalizability	0.63783	0.67923	0.61134	0.83561
Phi	0.59185	0.63049	0.56615	0.80208
S/N Relative	1.76113	2.11748	1.57292	5.08294
S/N Absolute	1.45007	1.70630	1.30494	4.05262
Contributions to…				
Universe Score Variance	33.08%	34.53%	32.39%	
Relative Error Variance	34.41%	30.90%	34.69%	
Absolute Error Variance	34.29%	31.37%	34.34%	
Nominal Weights for Composite	0.33333	0.33333	0.33333	1.00000

The results show a generalizability coefficient of $E\rho^2 = 0.836$ and a phi coefficient of $\phi = .802$ for the composite scores. The relative and absolute signal-to-noise ratios are 5.08 and 4.05, respectively. Of the three content categories, Geometry had the highest generalizability coefficient of $E\rho^2 = 0.679$. The Geometry content category also contributed more to the composite universe score variance and less to the composite error variances than the other two content categories.

The following table displays the error variances, covariances (below the diagonal), and correlations (above the diagonal) for the three content categories. From this table, one can see that the error variances are only slightly correlated with each other.

	Algebra	Geometry	Functions
Relative Error	**0.01958**	0.02040	0.03241
	0.00038	**0.01775**	0.01989
	0.00064	0.00037	**0.01976**
Absolute Error	**0.02379**	0.03350	0.05042
	0.00077	**0.02203**	0.03347
	0.00120	0.00077	**0.02382**

Side Note: Other Reliability Estimates

To compare the estimated generalizability coefficient with other reliability estimates, the data were entered into STATA and the following procedures were followed:

(1) First, all the data from the first test administration were entered into the computer. Cronbach's Alpha was calculated to be 0.843 for this test. Cronbach's Alpha for the second test administration was calculated at 0.858. Averaging these two reliability estimates, it is estimated that the "reliability of this test" is 0.851.

(2) Total scores from the first test administration were correlated with total scores from the second administration. This correlation was calculated to be 0.971. Therefore, the "reliability of this test" is 0.971.

Why is the coefficient calculated via Generalizability Theory lower than both of these estimates which are based on Classical Test Theory? The answer is, in large part, due to the way in which the occasion and item facets are treated. In the G Theory analysis, both the occasion and the items were treated as being random facets (we would like to generalize over both items and a single random occasion). In averaging the Cronbach's Alpha coefficients, the occasion was treated as being fixed. In calculating the test-retest correlation coefficient, the items were treated as being fixed. Treating a facet as being fixed will reduce error variance and increase estimates of reliability.

Determining the Optimal Number of Items

A generalizability coefficient of $E\rho^2 = 0.836$ seems pretty good for a placement test put together piecemeal by faculty over one summer. But finding the time and space to administer exams to incoming freshmen is difficult. What would happen to the generalizability of the test scores if the test length were shortened?

To answer this question, further D-Studies were conducted. It was always assumed that this math placement exam must have an equal number of items in each content category, although this is not a required assumption to conduct the D-Studies. The graphs in Figure 1 show what happens to the generalizability and phi coefficients as test length increases from 1 item per content category (3 item test) to 20 items per content category (60-item test). Based on these graphs, it does not appear as though adding additional items will significantly increase the generalizability of the exam scores. If the lowest acceptable generalizability coefficient is artificially set at 0.80, then the test length could be reduced to 24 items (8 items per content category).

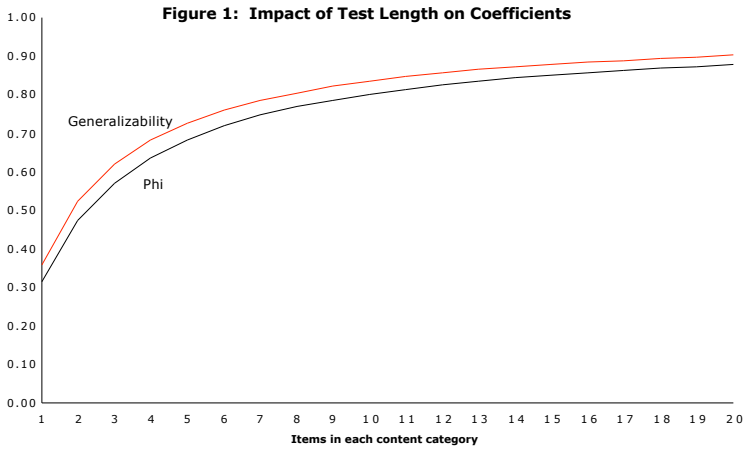
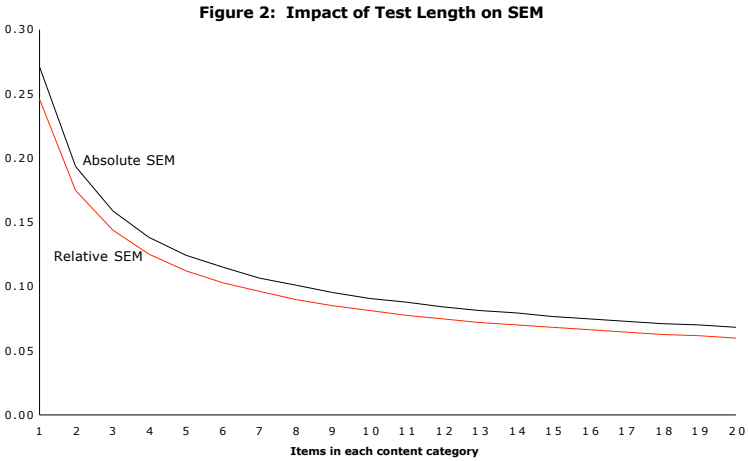


Figure 2 shows a similar graph showing the impact of test length on the absolute and relative standard errors (square root of error variances). The same conclusions are reached by examining this graph: if each test content category has fewer than 8 items, then the generalizability of the test scores starts to decline significantly.

Perhaps the generalizability of an even shorter exam could be increased by allowing a different number of items within each content category or by changing the weights of each content category. For example, the Geometry content category had the lowest amount of error variance. Perhaps if one Geometry item were added in place of 2 items from the other content categories, the generalizability would not decline so rapidly. The cost of doing this would be sacrificing the content specifications of the placement exam.



Impact of Time Limits on Occasions

Throughout this study, the difference in time limits between the two testing occasions has been ignored. Remember that on the first occasion, students were only given 20 minutes to complete the 40-minute exam. To see what impact, if any, the different time limits had on this study, separately analyses were conducted for each occasion. The following two tables display the G-Study variance component estimates and the D-Study results for the first and second occasions separately. Both the G- and D-Studies followed an $s^{\bullet} \times i^{\circ}$ design with three fixed content categories. The D-Study was conducted for 10 items within each content category.

G-Study Results for Occasions Separately
Variance, Covariance, and Correlation Estimates

	Occasion #1			Occasion #2		
	Algebra	**Geometry**	**Functions**	**Algebra**	**Geometry**	**Functions**
student	**0.03264**	0.90288	0.91555	**0.03955**	0.91015	1.05732
	0.03057	**0.03512**	0.97029	0.03640	**0.04044**	1.02017
	0.03024	0.03324	**0.03341**	0.03721	0.03630	**0.03131**
item	**0.04296**			**0.03175**		
		0.04563			**0.03649**	
			0.03777			**0.03345**
si	**0.17896**			**0.18054**		
		0.17400			**0.17699**	
			0.18082			**0.18878**

D-Study Results for Occasions Separately

		Algebra	**Geometry**	**Functions**	**Composite**
Universe Score Variance	Occasion 1	0.03264	0.03512	0.03341	0.03214
	Occasion 2	0.03955	0.04044	0.03131	0.03679
Relative Error Variance	Occasion 1	0.01790	0.01740	0.01808	0.00593
	Occasion 2	0.01805	0.01770	0.01888	0.00607
Absolute Error Variance	Occasion 1	0.02219	0.02196	0.02186	0.00733
	Occasion 2	0.02123	0.02135	0.02222	0.00720
Mean Error Variance	Occasion 1	0.00463	0.00491	0.00412	0.00166
	Occasion 2	0.00356	0.00404	0.00368	0.00142
Generalizability	Occasion 1	0.64590	0.66872	0.64885	0.84422
	Occasion 2	0.68660	0.69558	0.62388	0.85839
Phi	Occasion 1	0.59529	0.61527	0.60451	0.81419
	Occasion 2	0.65074	0.65450	0.58490	0.83634

Looking at the results, one can see that the second occasion yielded data that tended to have a higher universe score variance, slightly higher relative error variance, and slightly lower absolute error variance. The generalizability coefficients for occasion #1 was $E\rho^2 = 0.844$ and $E\rho^2 = 0.858$ for occasion #2. These coefficients are similar to the coefficient of $E\rho^2 = 0.836$ that was estimated for both occasions taken together. This taken with the fact that the G-Study variance component estimates are similar for each occasion seems to indicate that it wasn't too bad of a decision to combine the data into one analysis.

Further Studies

While this study has concluded that this placement test will need to consist of at least 8 items in each of the three content areas, further studies could be conducted on this data. For example, conditional standard errors of measurement could be calculated via Generalizability Theory for individual students. [redacted].

References

Brennan, R.L. (2001a). *Generalizability Theory*. New York: Springer-Verlag.

Brennan, R.L. (2001b). mGENOVA 2.1.

Brennan, R.L. (2001c). urGENOVA 2.1.

Crick, J.E. and Brennan, R.L. (1983). GENOVA 2.1.