TPM Vol. 32, No. R2, 2025
ISSN: 1972-6325
https://www.tpmap.org/

OPTIMIZING SCORING RELIABILITY FOR CREATIVE MATHEMATICAL PROBLEM-SOLVING ASSESSMENTS: A GENERALIZABILITY THEORY APPROACH

AUTTAPON PALADPHOM
KHON KAEN UNIVERSITY

PRAKITTIYA TUKSINO
KHON KAEN UNIVERSITY

ANUCHA SOMABUT
KHON KAEN UNIVERSITY

ABSTACT: Rater-induced error significantly challenges the scoring reliability of creative mathematical problem-solving assessments. This study applied Generalizability Theory to analyze score variance from 140 students and 3 raters across three scoring designs. The Generalizability (G) study revealed the person-by-rater interaction as the largest error source (35.50-35.90%), highlighting inconsistent rater judgments. A Decision (D) study showed that increasing raters from one to three substantially improved reliability (relative G-coefficient: .45 to .71). Notably, a design where each rater specializes in scoring specific items (p x (i:r)) yielded the highest absolute reliability (.69). These findings provide empirical guidance for designing effective scoring procedures to enhance the reliability of complex skill assessments.
Keywords: Generalizability Theory, creative problem-solving, constructed-response test, inter-rater reliability, essay test.

INTRODUCTION:

Creative problem-solving is recognized as a critical 21st-century competency, essential for navigating the complex and uncertain challenges of the future (OECD, 2018). As a discipline that fosters logical and systematic thinking, mathematics plays a pivotal role in developing this skill (Ministry of Education, 2017). Assessing such a complex, process-oriented skill requires instruments that allow learners to demonstrate their thought processes, analysis, and synthesis of ideas. Constructed-response tests are widely considered the most suitable tool for this purpose (Kanjanawasee, 2013).

However, despite their advantage in measuring higher-order thinking skills, constructed-response tests have a significant vulnerability that affects the reliability and fairness of the assessment: scoring error. This error can stem from numerous sources, particularly from the raters themselves. Factors such as inconsistency among different raters, or even within the same rater over time, fatigue, personal bias, and varied interpretations of scoring criteria all undermine the precision of the evaluation (Hoyt, 2000).

Classical Test Theory (CTT), the conventional framework for analyzing instrument quality, is limited in its ability to address these complex error sources. CTT aggregates all sources of error into a single value, making it impossible to identify the specific contribution of raters, items, or other facets to the total error variance. To overcome this limitation, Cronbach et al. (1972) developed Generalizability Theory (G-Theory), a more powerful approach to analyzing measurement reliability. G-Theory allows for the simultaneous estimation of error variance from multiple sources in a single analysis. It partitions the total score variance into components attributable to persons (the object of measurement), items, raters, and the interactions among these facets (Brennan, 2001; Shavelson & Webb, 1991).

While the issue of rater-induced error is widely acknowledged, research applying G-Theory to systematically compare the effectiveness of different scoring designs for creative problem-solving assessments remains limited. Most studies focus merely on increasing the number of raters without considering how the allocation of scoring tasks might impact reliability. This study, therefore, aims to apply Generalizability Theory to analyze the variance components of scores from a constructed-response test of creative mathematical problem-solving. It further seeks to identify the optimal scoring conditions—in terms of both the number of raters and the scoring design—to provide empirical guidance for enhancing the reliability of assessing this critical skill.

METHOD

Participants

The sample consisted of two groups: 1) a student sample of 140 high school students from northeastern Thailand during the 2023 academic year, selected via multi-stage sampling, and 2) a rater sample of three mathematics teachers, each with at least three years of experience teaching at the high school level, selected based on predefined qualifications.

Instrument

The primary instrument was a researcher-developed constructed-response test designed to measure creative mathematical problem-solving, comprising three items. Each item presented a problem scenario requiring the integration of mathematical knowledge in real-life contexts. An analytic scoring rubric was developed with four dimensions based on a synthesis of creative problem-solving models (Creative Education Foundation, 2015; OECD, 2012; Parnes, 1967; Torrance, 1962; Treffinger et al., 2004) to capture both convergent and divergent thinking: 1) Exploring and Understanding, 2) Generating Ideas, 3) Formulating Solutions, and 4) Verifying Solutions. The instrument underwent content validation by five experts (Item-Objective Congruence [IOC] index values ranged from 0.60 to 1.00) and a try-out phase to select items with appropriate difficulty and discrimination indices. The overall reliability of the test (Cronbach's Alpha) was .857.

Data Collection

The test was administered to the 140 student participants. All completed answer sheets were then duplicated and distributed to the three raters. Each rater scored every student's response on all three items according to the provided rubric, yielding data for a fully crossed person (p) x item (i) x rater (r) measurement design.

Data Analysis

Generalizability Theory (G-Theory) was employed for data analysis using the EduG software (Kanjanawasee, 2020). The analysis was conducted in two stages:

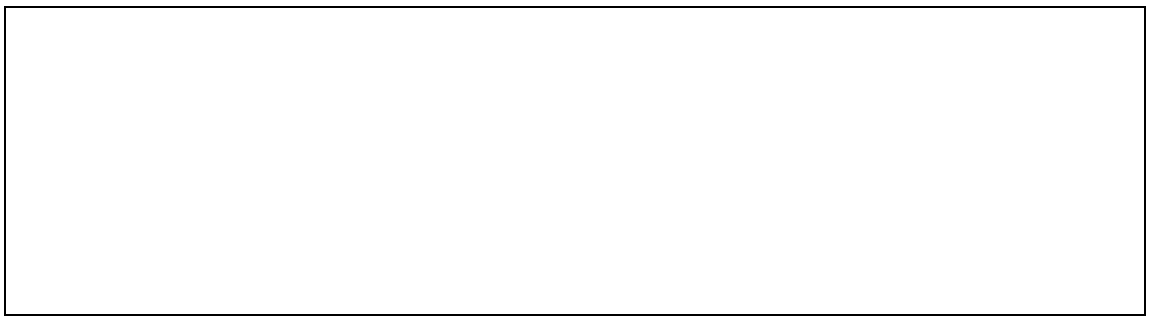
1. **Generalizability Study (G-Study):** This stage focused on estimating the variance components attributable to different sources of variation under three distinct scoring designs: 1) a fully-crossed design where every rater scores every item for every person (p x i x r), 2) a nested design where each person is scored by a different set of raters ((r:p) x i), and 3) a nested design where each rater specializes in scoring specific items (p x (i:r)). This analysis identified the proportion of total score variance contributed by each source.
2. **Decision Study (D-Study):** In this stage, the variance components estimated from the G-Study were used to calculate Generalizability coefficients (G-coefficients) under various measurement conditions. This allowed for an examination of how changing the number of raters (n'r = 1, 2, 3) and the scoring design would affect score reliability for both relative (norm-referenced) and absolute (criterion-referenced) decisions.

This study was approved by the Khon Kaen University approval no. HE663309 Written informed consent was obtained from parents/guardians, and assent was obtained from all student participants. All procedures complied with the Declaration of Helsinki and relevant institutional guidelines. Data were collected anonymously and stored securely."

RESULTS

G-Study: Sources of Score Variation

The estimation of variance components under the three scoring designs revealed the proportional contribution of each source to the total score variance, as summarized in Figure 1.



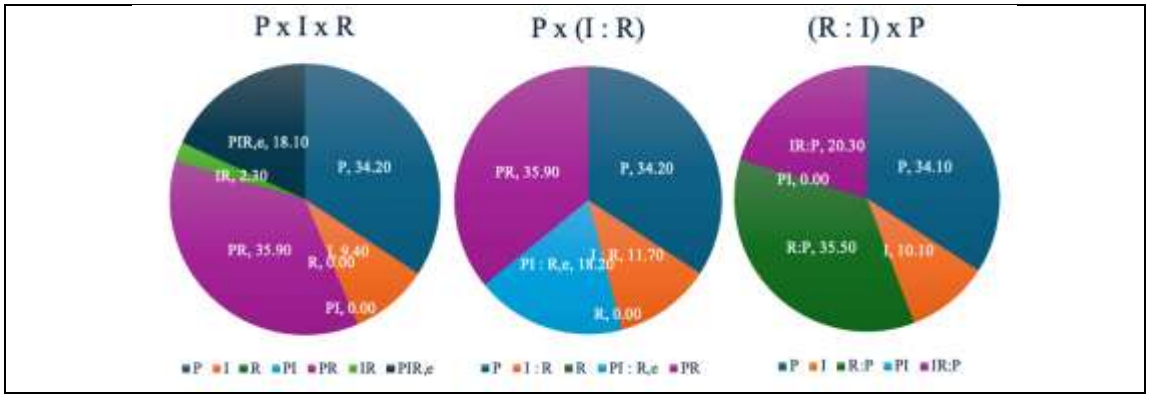


FIGURE 1 : Percentage of Variance Component Estimates by Scoring Design

Across all three designs, the variance component attributable to persons (P) was substantial and stable, ranging from 34.10% to 34.20%. This indicates that the test effectively differentiated among students based on their true ability levels. However, the most significant source of error variance was the person-by-rater interaction (PR).

- In the p x i x r design, the PR interaction accounted for 35.90% of the total variance.
- In the (r:p) x i design, the R:P component (rater nested within person) accounted for 35.50%.
- Similarly, in the p x (i:r) design, the PR interaction remained the largest component at 35.90%.

D-Study: Optimizing Measurement Conditions

The results of calculating G-coefficients under varying numbers of raters and scoring designs are presented in Figures 2, 3, and 4.

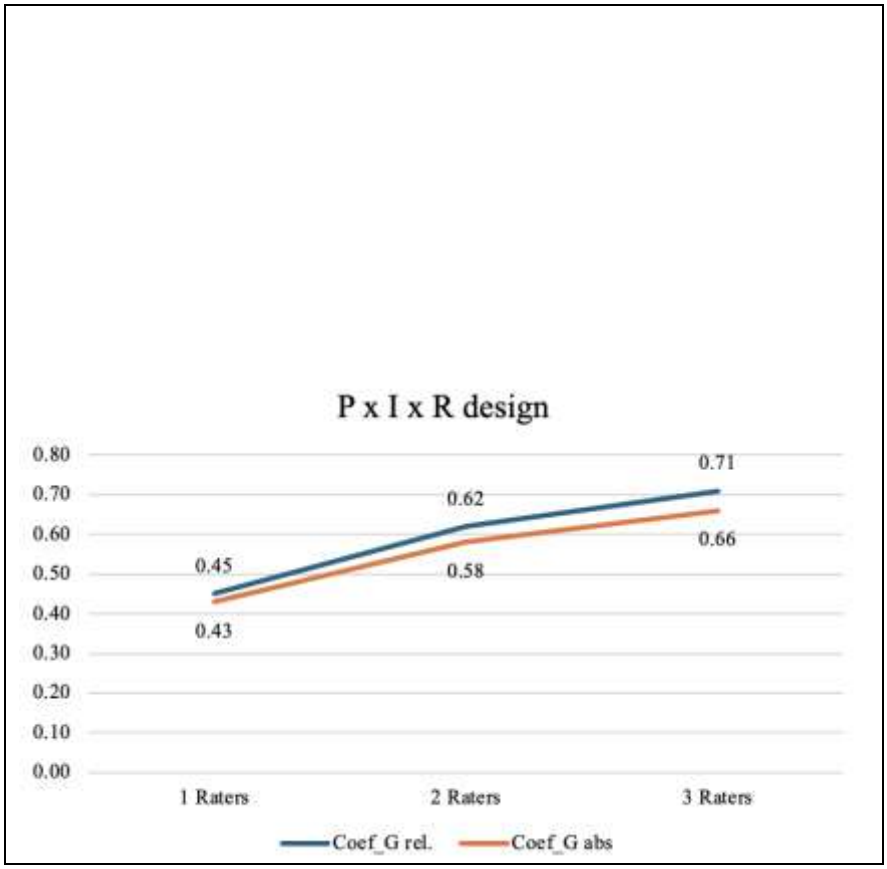


FIGURE 2 : D-Study Results for the p x i x r Design

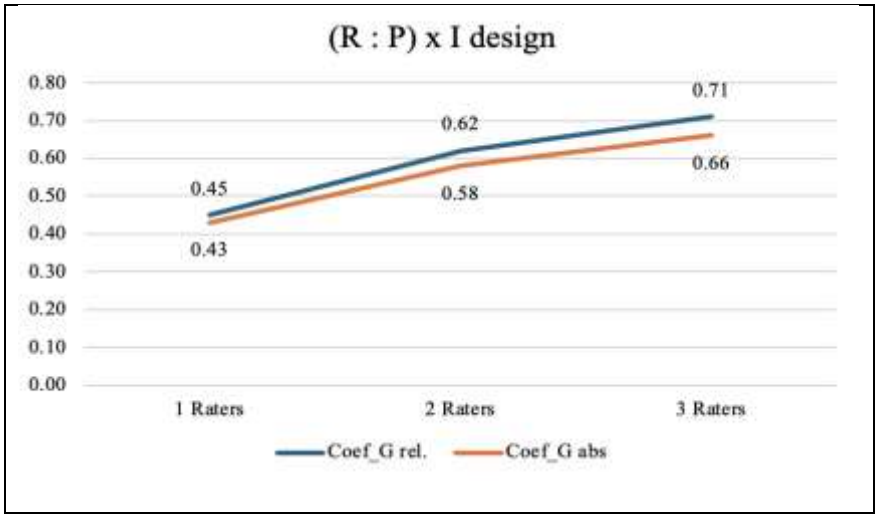


FIGURE 3 : D-Study Results for the (r:p) x i Design

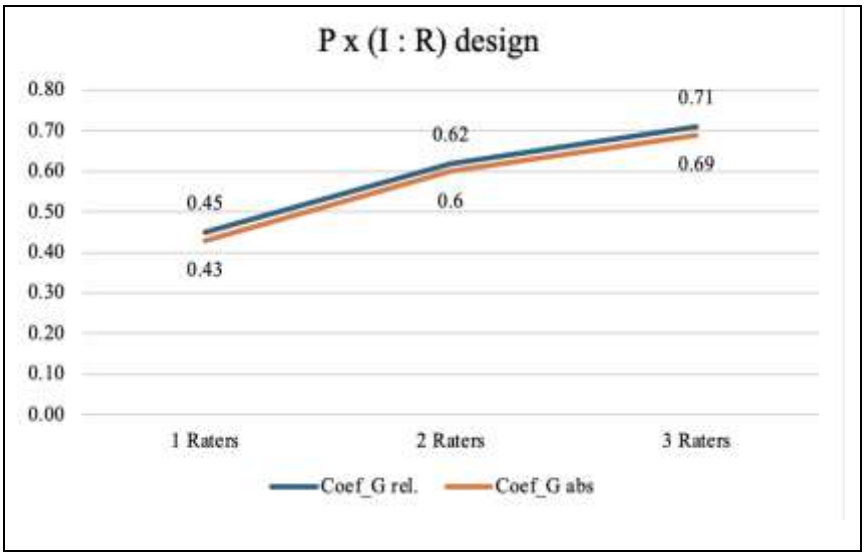


FIGURE 4 : D-Study Results for the p x (i:r) Design

The analysis demonstrated that increasing the number of raters from one to three substantially improved the G-coefficients across all designs. The relative G-coefficient, used for ranking students, increased from a low of .45 to an acceptable level of .71. Concurrently, the absolute G-coefficient, crucial for criterion-referenced decisions, rose from .43 to a range of .66–.69. This finding confirms that employing multiple raters is a highly effective strategy for mitigating rater-related error variance, which was identified as the primary issue in the G-Study.

When comparing designs with three raters, a notable difference emerged for absolute decisions. While all three designs yielded an identical relative G-coefficient of .71, the p x (i:r) design (where raters specialize in specific items) produced the highest absolute G-coefficient at .69, which was markedly higher than the values for the p x i x r design (.66) and the (r:p) x i design (.66).

DISCUSSION

This research aimed to analyze variance components and identify optimal scoring conditions for a constructed-response test of creative mathematical problem-solving. The findings provide two key empirical insights: 1) the person-by-rater interaction is the largest source of measurement error, and 2) increasing the number of raters while implementing an item-specialist scoring model is the most effective strategy for enhancing score reliability.

The Pervasive Influence of Person-by-Rater Interaction

The most striking finding from the G-study is that person-by-rater interaction variance constituted over one-third of the total variance, significantly overshadowing other error sources. This result aligns strongly with

a large body of research identifying rater inconsistency as a major challenge in performance-based assessments (Jonsson & Svingby, 2007). This issue is particularly acute when measuring creative problem-solving, where responses are diverse, lack a single correct answer, and require subjective interpretation based on a rubric, thereby increasing the potential for discrepancies in rater judgments (Kaufman et al., 2008). The high PR interaction variance provides quantitative evidence that the primary problem is not that some raters are systematically more lenient or severe (a main effect of Rater), but rather that raters apply scoring standards inconsistently across different persons, directly threatening the fairness of the assessment (Brennan, 2001).

Strategies for Enhancing Reliability: From Rater Numbers to Scoring Design

The D-study offers a concrete solution to the problem identified in the G-study. Increasing the number of raters from one to three elevates G-coefficients from a low level (< .50) to an acceptable one (> .70), providing clear evidence for the necessity of multiple raters in high-stakes assessments, as advocated by Shavelson and Webb (1991).

However, the more novel contribution of this research is the finding that the p x (i:r) design—assigning each rater to become an "item specialist"—yields the highest absolute reliability. This result can be interpreted from the perspective of cognitive load theory (Van Merriënboer & Sweller, 2005). When a rater focuses on applying the criteria for a single item repeatedly across many scripts, they develop expertise and apply the rubric more consistently. This reduces the cognitive burden and error associated with switching between the distinct criteria of multiple items. This aligns with previous findings by Apaikawee (2019) and Sanguanwai (2016). This discovery highlights that beyond the *number* of raters, the *method* of allocating scoring tasks is a critical factor influencing score quality, echoing the principles of efficient measurement design (Marcoulides, 1999).

Implications for Practice and Theory

Practically, this study provides clear guidance for educators and testing agencies: for assessments requiring high precision for criterion-referenced decisions (e.g., grading), using three raters with each assigned to a specific item is the most effective configuration. It also underscores the critical importance of intensive rater training and calibration as an essential prerequisite for reducing the problematic PR interaction variance (Stemler, 2004).

Theoretically, this research demonstrates the utility of G-Theory in dissecting complex measurement problems that CTT cannot address. By meticulously partitioning error variance, G-Theory provides targeted, data-driven insights for improving the measurement process.

CONCLUSION AND RECOMMENDATIONS

This study successfully applied Generalizability Theory to investigate scoring error in a complex, constructed-response assessment. It provided empirical evidence that 1) inconsistent rater judgments are the most significant source of error, and 2) a systematic approach involving an increased number of raters and an optimized scoring design can significantly enhance measurement reliability.

Practical Recommendations

1. For high-stakes assessments utilizing constructed-response items, institutions should employ at least two, and preferably three, raters to score each response.
2. To enhance efficiency and reduce cognitive load, an item-specialist model, where each rater is responsible for scoring only a subset of items, should be considered, particularly when absolute scores are of primary importance.
3. Mandatory and rigorous rater training sessions should be conducted prior to operational scoring to ensure a shared understanding of the rubric and to minimize the person-by-rater interaction variance.

Recommendations for Future Research

1. Future studies should incorporate additional facets, such as testing occasions, to examine the stability of student performance over time.

2. A comparative study analyzing the same dataset with Many-Facet Rasch Measurement (MFRM) could provide a more comprehensive understanding of the measurement characteristics.
3. Research should explore the cost-benefit trade-off between the number of raters employed and the resulting gain in reliability to establish practical guidelines for resource allocation in different assessment contexts.

REFERENCES

- Apaikawee, D. (2019). Improving the efficiency of constructed-response scoring. *Journal of Education, Mahasarakham University, 13*(4), 100–111.
- Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.

- Creative Education Foundation. (2015). *Educating for creativity: Level 1 resource guide*. https://www.creativeeducationfoundation.org/wp-content/uploads/2015/06/EFC-Level-1-FINALelectronic.pdf
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. John Wiley & Sons.
- Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*(2), 121–135. https://doi.org/10.1037/1082-989X.5.2.121
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: A review of the research. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.01.003
- Kanjanawasee, S. (2013). *Classical test theory* (7th ed.) [in Thai]. Chulalongkorn University Press.
- Kanjanawasee, S. (2020). *Modern test theory* (5th ed.) [in Thai]. Chulalongkorn University Press.
- Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. John Wiley & Sons.
- Marcoulides, G. A. (1999). Generalizability theory: A powerful methodology for examining reliability in measurement. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 227–254). Lawrence Erlbaum Associates.
- Ministry of Education. (2017). *Indicators and core learning standards for the Mathematics learning area (revised edition B.E. 2560) according to the Basic Education Core Curriculum B.E. 2551* [in Thai] (1st ed.). Printing House of the Agricultural Cooperative Federation of Thai land, Ltd.
- Office of the Basic Education Commission. (2018). *Guidelines for the development of learning measurement and evaluation according to the Basic Education Core Curriculum B.E. 2551* [in Thai]. https://academic.obec.go.th/images/document/1580786328_d_1.pdf
- OECD. (2012). *PISA 2012 creative problem solving framework*. https://www.oecd.org/pisa/innovation/creative-problem-solving/
- OECD. (2018). *The future of education and skills: Education 2030*. https://www.oecd.org/education/2030-project/contact/E2030%20Position%20Paper%20(05.04.2018).pdf
- Parnes, S. J. (1967). *Creative behavior guide book*. Charles Scribner's Son.
- Sanguanwai, C. (2016). *A comparison of test reliability for measuring mathematics creative problem solving abilities: Application of generalizability theory* [Master's thesis, Chulalongkorn University]. Chulalongkorn University Institutional Repository. https://cuir.car.chula.ac.th/handle/123456789/51081
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications.
- Stemler, S. E. (2004). A systematic approach to instrument development. *Practical Assessment, Research, and Evaluation, 9*(1), Article 4. https://doi.org/10.7275/2p5j-5b53
- Torrance, E. P. (1962). *Guiding creative talent*. Prentice-Hall.
- Treffinger, D. J., Isaksen, S. G., & Dorval, K. B. (2004). *Creative Problem Solving (CPS Version 6.1™): A contemporary framework for managing change*. Center for Creative Learning.
- Van Merriënboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*(2), 147–177. https://doi.org/10.1007/s10648-005-3951-0

APPENDIX A

Table 1 Generalizability Study (G-Study) Results: Estimated Variance Components and Percentage of Total Variance by Scoring Design

Design	Source of Variance	df	SS	MS	Estimated Variance Components	% of total Variance
P x I x R	P	139	1861.20	13.39	1.06	34.20
	I	2	263.86	131.93	0.29	9.40
	R	2	22.50	11.25	-0.01	0.00
	PI	278	156.59	0.56	0.00	0.00
	PR	278	1079.28	3.88	1.11	35.90
	IR	4	42.28	10.57	0.07	2.30
	PIR,e	556	310.61	0.56	0.56	18.10
	Total	1259	3736.31	172.14	3.08	100
P x (I : R)	P	139	1861.20	13.39	1.06	34.20
	I : R	6	306.14	51.02	0.36	11.70
	R	2	22.50	11.25	-0.10	0.00
	PI : R,e	834	467.20	0.56	0.56	18.20
	PR	278	1079.28	3.88	1.11	35.90

Design	Source of Variance	df	SS	MS	Estimated Variance Components	% of total Variance
	Total	1259	3736.31	80.10	2.99	100
(R : I) x P	P	139	1861.20	13.39	1.06	34.10
	I	2	263.86	131.93	0.31	10.10
	R:P	280	1101.78	3.93	1.10	35.50
	PI	278	156.59	0.56	-0.02	0.00
	IR:P	560	352.89	0.63	0.63	20.30
	Total	1259	3736.31	150.45	3.08	100

Table 2 Generalizability Coefficients and Error Variances from the Decision (D) Study by Scoring Design and Number of Raters

		ESTIMATED VARIANCE COMPONENTS IN D-STUDY								
	Design	**P x I x R**			**(R : P) x I**			**P x (I : R)**		
	n'_r	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Coef_G rel.		0.45	0.62	0.71	0.45	0.62	0.71	0.45	0.62	0.71
Coef_G abs		0.43	0.58	0.66	0.43	0.58	0.66	0.43	0.60	0.69
Rel. Err. Var.		1.29	0.65	0.43	1.31	0.66	0.44	1.29	0.65	0.43
Abs. Err. Var.		1.41	0.76	0.53	1.42	0.76	0.54	1.41	0.71	0.47