# A Novel Evolutionary Algorithm-Based Method for Optimal Sample Size Estimation in Generalizability Theory: Comparison with Three Traditional Methods

**Authors:** LI Guangming, Qin Yue, Qin Yue

**Date:** 2022-05-03T13:16:33+00:00

## Abstract

Generalizability Theory is widely applied in the field of psychological and educational measurement. How to achieve optimal reliability of measurement procedures under budget constraints is an important issue that researchers need to consider, and this problem can be transformed into one of optimal sample size estimation. This study proposes a novel method for estimating optimal sample size under Generalizability Theory based on evolutionary algorithms—the constrained evolutionary algorithm—and employs simulation studies to compare the advantages and disadvantages of three traditional methods (differential optimization method, Lagrange method, and Cauchy inequality method) with the constrained evolutionary algorithm. The results demonstrate that in two-facet crossed designs, two-facet nested designs, and three-facet crossed designs, the constrained evolutionary algorithm proves superior; researchers should prioritize its use in future studies.

## Full Text

### A New Method for Estimating Optimal Sample Size in Generalizability Theory Based on Evolutionary Algorithms: Comparisons with Three Traditional Methods

**LI Guangming**[1], **QIN Yue**[1],[2]
([1] School of Psychology, Center for Studies of Psychological Application, South China Normal University, Guangzhou 510631, China)
([2] Guangzhou Marine Geological Survey, Guangzhou 511466, China)

**Abstract**

Generalizability Theory (GT) is widely applied in psychological and educational measurement. A critical consideration for researchers is how to achieve optimal measurement reliability within budget constraints, a problem that can be framed as optimal sample size estimation. This study proposes a novel method based on evolutionary algorithms—Constrained Optimization Evolutionary Algorithms (COEAs)—for estimating optimal sample size under GT, and uses simulation studies to compare its performance against three traditional methods: the differential optimization method, Lagrange method, and Cauchy-Schwarz inequality method. Results demonstrate the superiority of COEAs across two-facet crossed designs, two-facet nested designs, and three-facet crossed designs. Researchers are recommended to prioritize this method in future work.

**Keywords:** Generalizability Theory, budget constraints, optimal sample size estimation, Constrained Optimization Evolutionary Algorithms

Generalizability Theory (GT) represents a modern psychological testing framework extensively used in psychological and educational measurement (Yang & Zhang, 2003; Zhu et al., 2013; Luo & Guo, 2014; Truong et al., 2021). A key advantage of GT is its capacity to precisely identify multiple sources of measurement error (Clayson et al., 2021). The theory comprises two phases: Generalizability Study (G-study) and Decision Study (D-study) (Qi et al., 2002). The G-study aims to identify and characterize various sources of measurement error within the universe of observations, given specified measurement targets and facets (Vispoel et al., 2020). The D-study then uses these variance component estimates to explore how measurement error can be controlled and regulated by adjusting relationships among measurement facets—such as altering sample sizes, facet relationships, or variable weights—to meet specific decision-making needs (Li, 2019a).

The generalizability coefficient reflects the accuracy of estimating universe scores from observed scores for examinees under conditions equally acceptable to the measurement procedure (Zhang & Lin, 2016). It is defined as the ratio of universe score variance to the sum of universe score variance and observed score variance, expressed as:

$$E\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)}$$

In this formula, $E\rho^2$ represents the generalizability coefficient, $\sigma^2(\tau)$ denotes the variance component for universe scores (i.e., measurement target variance), and $\sigma^2(\delta)$ represents relative error variance—the sum of variances from all interactions between measurement targets and facets. The magnitude of the generalizability coefficient indicates test reliability. As shown in Formula (1), the coefficient increases when the numerator (measurement target variance $\sigma^2(\tau)$)

increases or when the denominator (relative error variance $\sigma^2(\delta)$) decreases, meaning that reduced relative measurement error yields higher reliability.

In GT, increasing facet levels improves the generalizability coefficient. However, practical constraints—such as limited human, material, and financial resources—necessitate design trade-offs. In some cases, the additional observations required to meaningfully increase the coefficient may exceed available budgets, forcing researchers to weigh the necessity of such increases (Brennan, 2001). Thus, budget and cost are critical considerations that can affect measurement validity. A primary research concern is identifying highly reliable measurement procedures under budget limitations.

Increasing sample size is the simplest and most intuitive method for reducing error variance, yet sample size is constrained by funding and measurement content and cannot expand indefinitely. Moreover, due to varying measurement designs, simply increasing facet sample sizes may fail to achieve desirable generalizability coefficients even when actual costs surpass budgets (Marcoulides, 1993). Cronbach et al. (1972) noted that while increasing measurement facets generally improves generalizability coefficients, this improvement is unstable—under certain conditions, substantial increases in facet levels produce only minimal changes. Consequently, identifying the balance between budget and reliability, and obtaining optimal generalizability coefficients under total budget constraints, becomes a complex problem. The challenge is further complicated because different facets may incur different costs across various GT designs. Therefore, selecting an appropriate method to estimate optimal sample sizes that yield maximum generalizability coefficients under budget constraints represents a significant practical issue in GT applications.

Researchers have investigated optimal sample size estimation under budget constraints for many years (Marcoulides, 1993; Meyer et al., 2014; Li et al., 2020; Liu et al., 2020). However, previous work has focused primarily on traditional methods, notably the differential optimization method, Lagrange method, and Cauchy-Schwarz inequality method. These approaches, rooted in mathematical programming theory, can produce satisfactory solutions through applicability studies. Nevertheless, all traditional methods suffer from computational complexity, poor adaptability, and stringent application conditions, making them unsuitable for guiding optimal sample size selection in complex real-world scenarios.

With the development of high-performance computing algorithms, evolutionary algorithms have emerged as competitive alternatives to traditional optimization methods, often surpassing them across multiple performance metrics. Yet no theoretical or practical research has explored their application to this specific problem. This gap warrants investigation, discovery, and promotion. Accordingly, this study proposes a novel evolutionary algorithm-based approach—Constrained Optimization Evolutionary Algorithms (COEAs)—and uses simulation studies to compare its performance against the three traditional methods.

## 2.1 Traditional Methods

Cleary and Linn (1969) first addressed the problem of identifying optimal measurement designs under budget constraints, finding that in single-facet designs, optimal generalizability coefficients could be achieved by using the maximum number of items within budget limits. However, this approach lacks generalizability and fails when designs involve more than two facets.

Woodward and Joe (1973) derived mathematical methods for calculating optimal sample sizes under budget constraints in two- and three-facet crossed designs using differential optimization. This method substitutes budget constraint expressions into the generalizability coefficient equation and solves for extrema through differentiation, then screens reasonable results and substitutes them back into constraint formulas to obtain final optimal sample sizes. While conceptually straightforward, Woodward and Joe (1973) only discussed simple budget formulas for crossed designs under two- and three-facet conditions. Due to its theoretical limitations, the method cannot be easily extended to nested designs, and formula derivation difficulty and computational complexity increase exponentially with additional facets, presenting significant limitations.

Marcoulides and Goldstein (1990) first proposed using the Lagrange method to solve optimal sample size problems under budget constraints in two- and three-facet fully crossed random designs, introducing scaling techniques to simplify calculations for three-facet models. The Lagrange method introduces a multiplier $\lambda$. Zheng and Gao (2018) constructed its general form:

$$L(x, y, \lambda) = f(x, y) - \lambda\varphi(x, y)$$

The solution process involves optimizing function $f(x, y)$ under constraint $\varphi(x, y)$ by taking partial derivatives of $L(x, y, \lambda)$ with respect to $x$, $y$, and $\lambda$ and setting them to zero:

$$f'_x(x, y) - \lambda\varphi'_x = 0 \tag{1}$$
$$f'_y(x, y) - \lambda\varphi'_y = 0 \tag{2}$$
$$L'_\lambda = \varphi(x, y) = 0 \tag{3}$$

Solving this system yields values for $x$, $y$, and $\lambda$, where $(x, y)$ represents the extremum of the optimization function under constraints. If only one extremum exists, it constitutes the optimal solution. In GT budget constraint problems, the relative error variance function containing sample variables $\sigma^2_\delta$ typically serves as the optimization function, with transformed budget formulas as constraints, ultimately yielding optimal sample sizes.

However, Marcoulides and Goldstein (1990) only presented computational procedures without demonstrating applicability or explaining operational principles. Goldstein and Marcoulides (1991) subsequently validated the Lagrange

method's correctness for three-facet random models under budget constraints and proposed a binary search method, demonstrating its application through a three-facet random crossed design example. This approach, though effective, involves cumbersome steps and difficult implementation. They also introduced preliminary formulas for a rough four-facet random model to illustrate broader applicability, but these complex formulas lacked empirical data and research on applicability in other GT designs.

Marcoulides (1993) first discussed multiple budget constraint expressions, proposing several formulations for different practical measurement contexts. Meyer et al. (2014) examined the Lagrange method's applicability in two- and three-facet nested designs through a student teaching evaluation example. Nevertheless, the Lagrange method continues to exhibit poor adaptability and high computational demands.

Stuart (1954) proposed using the Cauchy-Schwarz inequality to estimate optimal sample sizes under constraints in survey sampling, where total sample variance equals the sum of facet variances weighted by sample size ratios: $v = \sum(v_h/n_h)$, with $v$ as total variance, $v_h$ as computed facet variance components, and $n_h$ as facet sample sizes. However, this approach cannot be directly applied to GT. Sanders (1992) first adapted the Cauchy inequality method to GT budget constraint problems, proposing solutions for random two-facet designs under several budget expressions and providing a two-facet example.

The Cauchy inequality is defined as: for any real numbers $a_1, a_2, \dots, a_n$ and $b_1, b_2, \dots, b_n$,

$$\sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2 \geq \left( \sum_{i=1}^{n} a_i b_i \right)^2$$

Equality holds if and only if $a_1/b_1 = \cdots = a_n/b_n = k$ (where $k$ is a constant). To apply this to GT, Sanders (1992) defined the budget expression as the sum of products between facet quantities and fixed costs: $c = \sum c_h n_h$, where $c$ is total budget, $c_h$ is fixed cost per unit for facet $h$, and $n_h$ is sample size for facet $h$. Following the Cauchy inequality structure:

$$v = \left( \sum \sqrt{v_h} \right)^2 \left( \sum c_h n_h \right) \geq \left( \sum \sqrt{v_h c_h} \right)^2$$

Equality holds when $\sqrt{v_h}/(c_h n_h) = k$ (constant). Since $k$ is solved as a constant and parameters $v_h$, $c_h$, $c$, and $v$ are known, optimal sample sizes $n_h$ can be derived. However, each GT design requires separate formula construction, and the Cauchy inequality strictly requires additive budget expressions.

Computationally, the Cauchy inequality method presents moderate difficulty but requires constructing special inequalities, increasing complexity substantially and limiting generalizability.

## 2.2 Budget Constraint Conditions in Traditional Methods

Research on optimal sample size estimation under budget constraints in GT has produced various budget expressions. Woodward and Joe (1973) proposed a simple product constraint: $n_i n_j = L$, where $n_i$ and $n_j$ are facet sample sizes and $L$ is total budget. Marcoulides (1993) introduced a multiplicative cost-defined expression: $cn_1 n_2 \ldots n_n \leq B$, where $c$ is unit cost per test item, $n_1, n_2, \ldots, n_n$ are facet levels, and $B$ is total budget. To accommodate inconsistent costs across facets or levels, researchers developed polynomial forms such as $c_1 n_1 + c_2 n_2 + c_{12} n_1 n_2 \leq B$, $c_1 n_1 + c_2 n_2 \leq B$, and $c_1 n_1 + c_{12} n_1 n_2 \leq B$, where $c_1$, $c_2$, and $c_{12}$ represent unit costs under specific budget definitions.

Budget expression selection is critical in traditional methods. Constraints can be simple or complex depending on research design and settings (Jutkowitz et al., 2019). Different GT designs may require specific budget expressions—for instance, when costs vary across levels within a facet or when a facet's cost must be fixed. Additionally, traditional methods impose specific requirements: differential optimization depends on simple budget expressions and may fail to converge or become computationally prohibitive with complex formulas, while the Cauchy inequality method requires additive expressions and cannot handle multiplicative forms.

## 2.3 The New Method

The optimal sample size estimation problem under budget constraints in GT can be mathematically formulated as a constrained optimization problem. In recent years, evolutionary algorithms developed through high-performance computing have emerged as state-of-the-art solutions for such problems. Traditional constrained optimization methods typically rely on gradient information, such as reduced gradient methods, projected gradient methods, and interior penalty function methods (Li et al., 2021). These approaches only optimize certain problem types and fail when faced with complex non-convex, discontinuous, non-differentiable, multi-objective, or multi-constraint problems. Moreover, they lack flexibility—slight problem variations can reduce effectiveness or cause complete failure.

Constrained evolutionary algorithms simulate biological processes like crossover, mutation, and recombination to drive computational operators toward optimization, guiding populations toward optimal solutions. Early versions suffered from premature convergence, search stagnation, and low precision, but modern algorithms balance global and local search with high accuracy and effective convergence. They also offer natural parallelism, require no additional information, and produce diverse outputs—advantages absent in traditional methods. Algorithm types include classical genetic algorithms, differential evolution, improved differential evolution, ant colony optimization, and multi-objective particle swarm optimization (Diao et al., 2017), each with distinct performance characteristics and trade-offs. However, no research has yet examined the appli-

cability of constrained evolutionary algorithms in GT.

Based on the characteristics of optimal sample size estimation under budget constraints in GT, this study selected the "Improved Differential Evolution Algorithm," which demonstrates mature performance in single-objective global optimization. This method is a novel, efficient heuristic parallel search technique (Neri & Tirronen, 2010) offering high optimization efficiency, simple parameter settings, and good robustness (Ding & Yin, 2017). The algorithm drives evolution through three typical genetic operations: it uses the difference vector between two randomly selected individuals from the initial population (weighted) as a mutation factor for a third individual, then performs crossover through random parameter mixing, evaluates fitness, and iteratively selects superior solutions to converge on optimal results.

The basic steps of the improved differential evolution algorithm are: (1) select an initial solution space population; (2) check termination conditions—stop if satisfied; (3) analyze and record population fitness metrics; (4) select difference vectors to obtain differential mutation individuals; (5) perform crossover between differential mutation individuals and the population to obtain a mutated population; (6) select between initial and mutated populations based on fitness; (7) return to step 2.

## 3. p×i×r Design Simulation Study

This study compares the four estimation methods using Monte Carlo data simulation in the widely applied two-facet crossed design p×i×r.

### 3.1 Generalizability Coefficient and Relative Error in p×i×r Design

In the p×i×r design, the generalizability coefficient and relative error expressions are:

$$E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{pir,e}^2}{n_i n_r}}$$

Relative error variance is:

$$\sigma_\delta^2 = \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{pir,e}^2}{n_i n_r}$$

### 3.2 Budget Expression for p×i×r Design

The Cauchy inequality method requires additive polynomial budget expressions. For consistency across two-facet designs, this study adopts:

$$c_1 n_1 + c_{12} n_1 n_2 \leq B$$

For generalizability and practical relevance, this study uses cost parameters from Li and Ou (2020) regarding teaching evaluation questionnaires (Table 1).

**Table 1** Questionnaire Costs

| Cost Item | Unit Cost |
|---|---|
| Questionnaire printing | $c_1 = 5$ yuan/item |
| Questionnaire mailing | $c_{12} = 0.4$ yuan/item |

In this context, $n_1$ represents the number of items and $n_2$ represents the number of examinees. To focus on optimization method validity, budget units are omitted, using only $c_{12} = 0.4$ and $c_1 = 5$ for abstract analysis.

### 3.3 Optimal Sample Size Estimation Methods for p×i×r Design

**3.3.1 Differential Optimization Method** For p×i×r designs, the differential optimization method substitutes the budget expression (8) into the generalizability coefficient expression (6) to obtain:

$$E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pr}^2}{\frac{B-c_i n_i}{c}} + \frac{\sigma_{pir,e}^2}{n_i \frac{B-c_i n_i}{c}}}$$

Taking partial derivatives yields:

$$\frac{\partial E\rho^2}{\partial n_i} = -\sigma_p^2 \left[\sigma_p^2 + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pr}^2}{\frac{B-c_i n_i}{c}} + \frac{\sigma_{pir,e}^2}{n_i \frac{B-c_i n_i}{c}}\right]^{-2} \cdot \left[-\frac{\sigma_{pi}^2}{n_i^2} + \frac{c_i \sigma_{pr}^2}{(B-c_i n_i)^2} + \frac{c_i \sigma_{pir,e}^2}{n_i (B-c_i n_i)^2} - \frac{\sigma_{pir,e}^2}{n_i^2 (B-c_i n_i)}\right]$$

The final optimal sample size expressions for $n_i$ and $n_r$ are:

$$n_i = \frac{-2B \cdot \sigma_{pi}^2 \cdot c_i n_i \pm \sqrt{(2B \cdot \sigma_{pi}^2 \cdot c_i n_i)^2 - 4[B \cdot c \cdot \sigma_{pr}^2 + c_i \cdot c \cdot \sigma_{res}^2 - \sigma_{pi}^2 - \sigma_{pi}^2 \cdot c_i^2 + c_i \cdot c \cdot \sigma_{res}^2 \cdot c_i^2] \cdot \sigma_{pi}^2 \cdot B^2}}{2[B \cdot c \cdot \sigma_{pr}^2 + c_i \cdot c \cdot \sigma_{res}^2 - \sigma_{pi}^2 - \sigma_{pi}^2 \cdot c_i^2 + c_i \cdot c \cdot \sigma_{res}^2 \cdot c_i^2]}$$

$$n_r = \frac{B - c_i n_i}{c}$$

**3.3.2 Lagrange Method** In p×i×r designs, the Lagrange method substitutes the relative error expression (7) and budget expression (8) into the general Lagrange formula (2):

$$L(n_i, n_r, \lambda) = \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{pir,e}^2}{n_i n_r} - \lambda(c_1 n_1 + c_{12} n_1 n_2 - B)$$

Following the Lagrange solution process and referencing Li and Ou (2020), the optimal sample size expressions become:

$$n_i = \sqrt{\frac{\sigma_{pi}^2 + \sigma_{pir,e}^2/n_r}{c_i \lambda}}, \quad n_r = \sqrt{\frac{\sigma_{pr}^2 + \sigma_{pir,e}^2/n_i}{c\lambda}}$$

**3.3.3 Cauchy Inequality Method** Similar to the Lagrange method, the Cauchy inequality method uses relative error expression (7) and budget expression (8). Applying Cauchy inequality principles and Sanders' (1992) construction method yields constant $k$, which combined with the budget expression produces:

$$k = \frac{\sqrt{\sigma_{pi}^2}}{c_i + cn_r} = \frac{\sqrt{\sigma_{pr}^2}}{c_i n_i + c} = \frac{\sqrt{\sigma_{pir,e}^2}}{c_i n_i + cn_r}$$

The resulting optimal sample size expressions are:

$$n_i = \sqrt{\frac{\sigma_{pi}^2}{c_i k^2}}, \quad n_r = \frac{B - c_i n_i}{c}$$

**3.3.4 Constrained Evolutionary Algorithm** Unlike traditional methods requiring formula derivation, COEAs need only the budget constraint expression (8) and optimization objective (7) as inputs. After algorithm execution with appropriately tuned parameters, optimal solutions (sample sizes $n_i$ and $n_r$) are obtained directly. Key parameters for the improved differential evolution algorithm include:

1. Initialize objective dimension M = 1, decision variable dimension Dim = 2
2. Optimization objective: `pop.ObjV = evpr / x1 + evpi / x2 + evpir / (x1 * x2)`
3. Constraint: `pop.CV = np.hstack([(c1 * x1 * x2 + c * x2) - B])`
4. Tune fitness, mutation rate, recombination rate:
   - `myAlgorithm.MAXGEN = 10000` (maximum generations)
   - `myAlgorithm.mutOper.F = 0.00001` (differential evolution parameter F)
   - `myAlgorithm.recOper.XOVR = 0.7` (recombination probability)

As a probabilistic algorithm, COEA performance depends on population size, encoding scheme, fitness, and mutation rate, requiring manual tuning for optimal results.

### 3.4 Computational Tools

To handle large-scale simulation data, this study coded all computational methods. R programs were developed for differential optimization, Lagrange, and Cauchy inequality methods. Python was used with the Geatpy2 evolutionary algorithm toolkit (co-developed by South China Agricultural University and Jinan University) to implement the improved differential evolution algorithm, with interfaces retained for simulation data integration.

### 3.5 Simulation Data Generation

Following Brennan (2001), Monte Carlo simulation in GT involves:

First, converting the GT design to a mathematical model. For p×i×r:

$$X_{pir} = \mu+(\mu_p-\mu)+(\mu_i-\mu)+(\mu_r-\mu)+(\mu_{pi}-\mu_p-\mu_i+\mu)+(\mu_{pr}-\mu_p-\mu_r+\mu)+(\mu_{ir}-\mu_i-\mu_r+\mu)+(X_{pir}-\mu_{pi}-\mu_{pr}-\mu$$

Second, converting to variance component form:

$$X_{pir} = \mu + \nu_p + \nu_i + \nu_r + \nu_{pi} + \nu_{pr} + \nu_{ir} + \nu_{pir,e}$$

Third, transforming to:

$$X_{pir} = \mu + \sigma_p z_p + \sigma_i z_i + \sigma_r z_r + \sigma_{pi} z_{pi} + \sigma_{pr} z_{pr} + \sigma_{ir} z_{ir} + \sigma_{pir} z_{pir}$$

where variance parameters $\sigma_p, \sigma_i, \sigma_r, \sigma_{pi}, \sigma_{pr}, \sigma_{ir}, \sigma_{pir}$ are known, $\mu$ (typically set to 0 in simulations) is the expected score, and $z_p, z_i, z_r, z_{pi}, z_{pr}, z_{ir}, z_{pir}$ are random numbers from (0,1) generated using normal random functions (e.g., `rnorm` in R).

Finally, components are summed to produce simulated data $X_{pir}$, formatted into final matrices according to facet sample sizes.

This study used R for Monte Carlo simulation. Due to R's single-core limitations and low efficiency for large matrix operations, Docker container virtualization enabled multi-core parallel computing on Linux high-performance servers, improving efficiency by over 20× compared to standard PCs—a technique applicable to other simulation studies.

### 3.6 Simulation Data Parameter Estimation

After generating simulation matrices, the study used the mature R package `gtheory` (based on linear mixed models and restricted maximum likelihood) for parameter estimation. For the p×i×r model, variance parameters were set using values from Li (2019b) (Table 2), with measurement target $p$ sample size fixed at $n_p = 30$ based on practical experience.

**Table 2** Variance Components for p×i×r Design

| Variance Component | Value |
|---|---|
| $\sigma_p^2$ | 0.85 |
| $\sigma_i^2$ | 0.05 |
| $\sigma_r^2$ | 0.03 |
| $\sigma_{pi}^2$ | 0.04 |
| $\sigma_{pr}^2$ | 0.02 |
| $\sigma_{ir}^2$ | 0.01 |
| $\sigma_{pir,e}^2$ | 0.05 |

To examine method effectiveness across different facet levels, multiple sample sizes were tested: $n_i = (20, 40)$ and $n_r = (5, 10)$, creating four design combinations. Budget constraints were set at four levels: $B = (100, 150, 200, 250)$. Each design was simulated 500 times (2,000 total simulations). Variance components from `gtheory` were output to CSV files for analysis.

### 3.7 Simulation Results

For analysis, average optimal generalizability coefficients were computed across four methods × four designs (Table 3).

### 3.8 Analysis and Discussion

Table 3 shows that all four methods produced excellent generalizability coefficients meeting typical research needs in p×i×r designs. Notably, three traditional methods yielded nearly identical results, likely due to computational precision and estimation similarities, which also validates the derived differential optimization formulas under the new budget expression.

COEAs achieved comparable coefficients while demonstrating superior budget compliance. In computational time, traditional methods appeared faster since calculations occurred outside programs, but this difference is negligible given that design selection occurs only once per study.

## 4. (r:p)×i Design Simulation Study

This study also simulated the two-facet nested design (r:p)×i. Mathematical derivations were similar to p×i×r, with results showing comparable algorithm

performance. Due to space limitations, details are omitted.

## 5. p×i×r×o Design Simulation Study

### 5.1 Generalizability Coefficient and Relative Error in p×i×r×o Design

The three-facet crossed design p×i×r×o is a classic three-facet GT design and foundation for others. In practice, $p$ may represent examinees, $i$ test items, $r$ raters, and $o$ test occasions.

The generalizability coefficient and relative error expressions are:

$$E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{po}^2}{n_o} + \frac{\sigma_{pir}^2}{n_i n_r} + \frac{\sigma_{pio}^2}{n_i n_o} + \frac{\sigma_{pro}^2}{n_r n_o} + \frac{\sigma_{piro,e}^2}{n_i n_r n_o}}$$

Relative error variance:

$$\sigma_\delta^2 = \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{po}^2}{n_o} + \frac{\sigma_{pir}^2}{n_i n_r} + \frac{\sigma_{pio}^2}{n_i n_o} + \frac{\sigma_{pro}^2}{n_r n_o} + \frac{\sigma_{piro,e}^2}{n_i n_r n_o}$$

### 5.2 Budget Expression for p×i×r×o Design

Three-facet designs present substantially greater derivation and computational challenges, making budget expression requirements more stringent. Differential optimization and Lagrange methods require multiplicative budget expressions:

$$cn_1n_2n_3 \leq B$$

The Cauchy inequality method still demands polynomial expressions:

$$c_1n_1 + c_2n_2 + c_3n_3 \leq B$$

Only COEAs accommodate all expression types, serving as a comparative bridge in three-facet designs. For simulation, multiplicative expression uses $c = 1$, while polynomial uses $c_1 = 20$, $c_2 = 30$, $c_3 = 40$.

### 5.3 Optimal Sample Size Estimation Methods for p×i×r×o Design

**5.3.1 Differential Optimization Method** Woodward and Joe (1973) proposed a complex sixth-degree polynomial solution for three-facet designs by substituting multiplicative budget expression (21) into the generalizability coefficient formula (19) and taking partial derivatives to obtain $n_r$:

$$0 = An_r^6 + Bn_r^5 + Cn_r^4 + Dn_r^3 + En_r^2 + Fn_r + G$$

where coefficients $A$ through $G$ are complex functions of variance components. After solving for $n_r$, $n_i$ and $n_o$ are obtained through:

$$n_i = \sqrt{\frac{n_r\sigma_{pi}^2 L + \sigma_{pir}^2 + \sigma_{pio}^2}{\sigma_{pr}^2 L + \sigma_{pir}^2 + \sigma_{pro}^2}}, \quad n_o = \frac{B}{cn_i n_r}$$

**5.3.2 Lagrange Method** Marcoulides and Goldstein (1990) introduced scaling to obtain upper bounds for relative error variance. This study applies similar scaling to the three-facet relative error expression (20):

$$\sigma_\delta^2 \leq \frac{\sigma_{pi}^2 + \sigma_{pio}^2 + \sigma_{pir}^2}{n_i} + \frac{\sigma_{pr}^2 + \sigma_{pro}^2 + \sigma_{pir}^2}{n_r} + \frac{\sigma_{po}^2 + \sigma_{pio}^2 + \sigma_{pro}^2}{n_o}$$

Substituting into the Lagrange formula with budget expression (21) yields optimal sample sizes:

$$n_i = \sqrt{\frac{c(\sigma_{pi}^2 + \sigma_{pio}^2 + \sigma_{pir}^2)(\sigma_{pr}^2 + \sigma_{pro}^2)}{\sigma_{po}^2 + \sigma_{pio}^2 + \sigma_{pro}^2}}, \quad n_r = \sqrt{\frac{c(\sigma_{pr}^2 + \sigma_{pro}^2 + \sigma_{pir}^2)(\sigma_{pi}^2 + \sigma_{pio}^2)}{\sigma_{po}^2 + \sigma_{pio}^2 + \sigma_{pro}^2}}, \quad n_o = \sqrt{\frac{c(\sigma_{po}^2 + \sigma_{pio}^2 + \sigma_{p}^2}{\sigma_{pr}^2 + \sigma_{pr}^2}}$$

where $c$ corresponds to $B$ in this study.

**5.3.3 Cauchy Inequality Method** No prior research has applied Cauchy inequality to three-facet designs, likely due to construction complexity. This study extends the method using scaling:

First, scale the p×i×r×o relative error expression (20). Then construct polynomial budget expression (22). Finally, applying Cauchy inequality yields:

$$n_i = \sqrt{\frac{(\sigma_{pi}^2 + \sigma_{pio}^2 + \sigma_{pir}^2) \cdot c_i}{c_i + c_o\sqrt{(\sigma_{po}^2 + \sigma_{pro}^2 + \sigma_{pir}^2) \cdot c_i^2 + \sigma_{res}^2 + \sigma_{pir}^2) \cdot c_o^2 + \sigma_{pro}^2 + \sigma_{pir}^2 + \sigma_{res}^2) \cdot c_r}}}, \quad n_r = \sqrt{\frac{(\sigma_{pr}^2 + \sigma_{pro}^2 + \sigma_{p}^2}{c_i -}}$$

**5.3.4 Constrained Evolutionary Algorithm** COEAs require no formula derivation and can accommodate both budget expression types (21) and (22), serving as a comparative bridge. Key parameters for the improved differential evolution algorithm include:

1. Initialize M = 1, Dim = 2
2. Optimization objective: `pop.ObjV = evc_p/(evc_p + evc_{pi}/x1) + evc_{pj}/x2 + evc_{pk}/x3 + evc_{pij}/(x1*x2) + evc_{pik}/(x1*x3) + evc_{pjk}/(x2*x3) + evc_{res}/(x1*x2*x3)`

3. Constraints: `pop.CV = np.hstack([(c1*x1 + c2*x2 + c3*x3) - B])` and `pop.CV = np.hstack([x1*x2*x3 - B])`
4. Tune parameters:
   - `myAlgorithm.MAXGEN = 40000`
   - `myAlgorithm.mutOper.F = 0.00001`
   - `myAlgorithm.recOper.XOVR = 0.8`

### 5.4 Computational Tools

Tools mirror those used in the p×i×r study.

### 5.5 Simulation Data Generation

Based on the p×i×r×o model:

$$X_{piro} = \mu+\nu_p+\nu_i+\nu_r+\nu_o+\nu_{po}+\nu_{io}+\nu_{ro}+\nu_{pi}+\nu_{pr}+\nu_{ir}+\nu_{pir}+\nu_{pio}+\nu_{pro}+\nu_{iro}+\nu_{piro,e}$$

Variance parameters from Meyer et al. (2014) were used (Table 4). Measurement target $p$ sample size was fixed at $n_p = 30$. Facet levels were $n_i = (15, 25)$, $n_r = (5, 10)$, $n_o = (5, 10)$, creating eight design combinations. Budget levels were $B = (800, 900, 1000, 1100, 1200, 1300, 1400, 1500)$. Each design was simulated 200 times (1,600 total simulations). Results were analyzed using `gtheory` and output to CSV files.

### 5.6 Simulation Data Parameter Estimation

The four computational methods were applied to the eight variance component datasets using custom R and Python programs, producing 200 × 8 sets of generalizability coefficients, optimal sample sizes, design costs, and analysis times, saved to CSV files.

### 5.7 Simulation Results

Average optimal generalizability coefficients were computed across 5 methods × 8 designs (COEAs appears twice as a bridge). For brevity, 5 × 4 results are presented (Tables 5 and 6).

### 5.8 Analysis and Discussion

Table 5 shows that in p×i×r×o designs, the Lagrange method produces higher generalizability coefficients but substantially exceeds budget limits, severely reducing its practical guidance value. Differential optimization strictly adheres to budgets but yields poorer optimization due to cumbersome formulas. COEAs demonstrates excellent performance in both budget compliance and optimization results.

Table 6 reveals that the Cauchy inequality method achieves good coefficients under specific budget expressions but depends strictly on them. COEAs outperforms Cauchy inequality even under its required expression conditions, showing greater robustness and staying closer to budget limits.

COEAs performs well under both expression types, demonstrating unmatched extensibility and robustness compared to traditional methods. The Cauchy inequality method's effectiveness also validates this study's first derivation of three-facet Cauchy inequality application.

Computational time differences are negligible for practical research purposes.

## 6.1 Impact of Budget Rounding

All four methods produce continuous optimization results with non-integer sample sizes. In practice, integer sample sizes are required, typically obtained through rounding. However, rounded integers may not be truly optimal and often cause budget overruns, as confirmed in multiple studies. In large-scale testing, these overruns can be substantial. A practical solution involves evaluating all possible rounding combinations to select the best option. Saunders et al. (1989) proposed integer programming methods that avoid rounding but suffer from poor feasibility, high algorithmic demands, and heavy computation, making them less practical than the four algorithms discussed here.

## 6.2 Impact of Negative Variance Components

Some GT designs may produce negative variance component estimates due to data issues. Generally, negative components can be treated as zero (Brennan, 2001). However, substituting zeros directly may cause unsolvable equations in traditional methods requiring extensive fractional operations. COEAs does not encounter this problem.

## 6.3 Impact of Fixed Facets

This study examined random facet designs. In fixed facet designs where certain facet levels are predetermined or immutable, error sources decrease as fixed facet levels increase, improving generalizability coefficients. Meyer et al. (2014) discussed solutions for fixed-facet three-facet designs, which are relatively simpler simplifications of random facet problems.

## 6.4 Impact of Unbalanced Designs

Unbalanced designs may occur in practice, such as the (r:p)×i design where facet $p$ has inconsistent level counts. In GT variance component estimation, missing data are typically imputed to the maximum level standard. When level sample size differences are large, substantial missing value imputation dramatically in-

creases data volume, potentially distorting variance component estimates. COEAs effectively resolves unbalanced design issues.

## 6.5 Impact of Computer Programs on Traditional Methods

Given the complexity and data volume, this study programmed all three traditional methods for batch computation. While enabling rapid calculation, computer precision and computational principles may introduce minor effects. For instance, the differential optimization method sometimes failed to find solutions when substituting certain variance components, requiring complex workarounds. These R-language implementations may yield slight differences in other languages due to varying computational precision or principles.

## 6.6 Impact of Scaling Methods on Traditional Methods

To address three-facet design requirements, Marcoulides and Goldstein (1990) introduced scaling to obtain upper bounds for relative error variance. This study also applied scaling to extend Cauchy inequality to three-facet designs. However, scaling inherently reduces precision. While Lagrange with scaling achieves good coefficients, it significantly exceeds budgets—acceptable in low-budget scenarios but problematic for large-scale testing. This study's novel three-facet Cauchy inequality scaling application produced good results within budget, but scaling's imprecision necessitates searching for optimal solutions within a range around results.

## 6.7 Impact of COEA Performance

As a computational algorithm, COEA speed depends on hardware. This study used ordinary PCs (i5 processor, 8GB RAM) rather than high-performance Linux workstations, yielding second-level computation times fully meeting research needs. The improved differential evolution algorithm represents a mature single-objective optimization method, but numerous other constrained evolutionary algorithms exist (e.g., multi-chromosome genetic algorithms) that may perform even better for GT sample size estimation.

## Conclusion

Comparing three traditional methods with COEAs across p×i×r, (r:p)×i, and p×i×r×o designs reveals:

1. **Two-facet designs**: COEAs slightly outperforms traditional methods, which show equivalent performance. All three traditional methods produce nearly identical results under identical budget expressions, demonstrating good performance but occasional budget overruns. COEAs matches their precision while showing superior budget compliance.

2. **Three-facet designs**: COEAs significantly outperforms traditional methods, which show divergent performance. Differential optimization yields the poorest coefficients but best budget compliance; Lagrange produces the best coefficients but substantially exceeds budgets; Cauchy inequality performs well under specific budget expressions. COEAs achieves relatively optimal coefficients while maintaining good budget compliance, outperforming all traditional methods.
3. **Complexity**: COEAs demonstrates dramatically lower complexity. Traditional methods become increasingly difficult with more facets, relying heavily on scaling that reduces precision. COEAs remains simple and consistent across designs, requiring only empirical parameter tuning.
4. **Applicability**: COEAs is significantly superior. Traditional methods have narrow applicability, depending on specific designs and budget expressions. COEAs works robustly across various designs and expressions.
5. **Generalizability**: COEAs shows clear advantages. Traditional methods' mathematical limitations hinder extension to complex designs, while COEAs easily generalizes to four or more facets and can extend to multivariate GT using multi-objective constrained evolutionary algorithms.
6. **Optimal solution probability**: COEAs' probabilistic nature allows multiple runs to discover better solutions that traditional methods cannot find.

**In summary, COEAs is superior to traditional methods for optimal sample size estimation under budget constraints in GT and should be prioritized in future research.**

*Note: Figure translations are in progress. See original paper for figures.*