Using Ranked Set Sampling with Cluster Randomized Designs for Improved Inference on Treatment Effects

Xinlei Wang, Johan Lim, and Lynne Stokes*

Abstract

This article examines the use of ranked set sampling (RSS) with cluster randomized designs (CRDs), for potential improvement in estimation and detection of treatment or intervention effects. Outcome data in cluster randomized studies typically have nested structures, where hierarchical linear models (HLMs) become a natural choice for data analysis. However, nearly all theoretical developments in RSS to date are within the structure of one-level models. Thus, implementation of RSS at one or more levels of an HLM will require development of new theory and methods. Under RSS-structured CRDs developed to incorporate RSS at different levels, a nonparametric estimator of the treatment effect is proposed; and its theoretical properties are studied under a general HLM that has almost no distributional assumptions. We formally quantify the magnitude of the improvement from using RSS over SRS (simple random sampling), investigate the relationship between design parameters and relative efficiency, and establish connections with one-level RSS under completely balanced CRDs, as well as studying the impact of clustering and imperfect ranking (under a familiar linear ranking error model). Further, based on the proposed RSS estimator, a new test is constructed to detect treatment

*Xinlei Wang is Associate Professor and Lynne Stokes is Professor, Department of Statistical Science, Southern Methodist University, 3225 Daniel Avenue, P O Box 750332, Dallas, Texas 75275-0332 (swang@smu.edu and slstokes@smu.edu). Johan Lim is Professor, Department of Statistics, Seoul National University (yohanlim@gmail.com).

effects, which is distribution-free and easy to use. Simulation studies confirm that in general, the proposed test is more powerful than the conventional F-test for the original CRDs, especially for small or medium effect sizes. Two empirical studies, one using data from educational research (i.e., the motivating application) and the other using human dental data, show that our methods work well in real world settings and our theories provide useful predictions at the stage of experimental design; and that substantial gains may be obtained from the use of RSS at either level.

***Keywords*:** asymptotic pivotal method; cost efficiency; distribution-free; hierarchical linear models; ICC; imperfect judgement ranking; least squares estimator; nonparametric inference; order statistics; ranking error; relative efficiency; two-stage designs.

1 Introduction

Ranked set sampling (RSS) has been known as a cost-efficient sampling method for many years (Wolfe 2004, Chen et al. 2006). It was first introduced by McIntyre (1952) in forestry and has been applied in fields in which precise measurement of outcomes is expensive, but assessment of relative sizes of outcomes (ranks) of a few units at a time is feasible. The method uses ranking, usually at the time of sampling, to provide auxiliary information to select a sample, and offers a way to achieve economy by reducing the number of measurements required for estimation. In theory, RSS or its variants have been shown to improve inference for many types of parameters, including mean, variance, quantiles, correlation coefficients, and distribution functions (e.g., Takahashi and Wakimoto 1968, Stokes and Sager 1988, Bohn and Wolfe 1992, Kvam and Samaniego 1994, Ozturk 2002, MacEachern et al. 2002, Fligner and MacEachern 2006, Ozturk and Balakrishnan 2009, etc.). In various applications, RSS has been found to greatly enhance the efficiency of estimation and power of statistical inference (e.g., Nussbaum and Sinha 1997, Mode et al. 1999, Murray et al. 2000, Kvam 2003). Nowadays, RSS remains an active research area and recent work includes Frey et al. (2007), Ghosh and Tiwari (2007), Frey (2007), Ozturk (2008), Balakrishnan and Li (2008), Chen and Lim (2011), Frey and Ozturk (2011), Wang et al. (2012), Ozturk (2012), Frey and Feeman (2013), Ozturk (2013), Hatefi et al. (2013). For a detailed review, see Wolfe (2012).

In this article, we examine the application of RSS in cluster randomized designs (CRDs, Hayes and Moulton 2009). Though well established in Statistics, RSS has not yet been applied in cluster randomized experiments, which are widely used in educational, social and medical studies to assess treatment or intervention effects. One reason may be that data from such experiments typically have nested structures, requiring hierarchical modeling methods for correct analysis. The theoretical developments in RSS to date have included designs and analyses for estimation and testing of various parameters, but all within the structure of one-level models. Implementation of RSS at one or more levels of a hierarchical model will require not only adaptation of the one-level analysis methods that currently exist, but also new theoretical and methodological developments to bridge the gap.

Our work is partially motivated by the need of improving statistical power/cost efficiency in educational experiments, which often involve CRDs due to nested data structures (e.g., schools, teachers, and students). During the past decade, educational research has been galvanized by new legislation such as the No Child Left Behind Act. Increasing emphasis is being placed on accurately quantifying the success of intervention programs, where the improvement can be measured by comparing the mean or change in scores of students receiving/not receiving treatment through CRDs. There are factors that greatly favor the use of RSS in such studies. First, implementing intervention programs in a large number of schools is often a difficult and/or expensive task. Obtaining outcomes for students involved in an educational experiment also may be costly, especially when assessment must be carried out individually. Thus a data collection method that can increase statistical power without adding additional sites or students would be useful to reduce cost. In other situations, reduction of sample size is preferable so as to expose the minimum possible number of students to an intervention that may be controversial, or whose benefit to risk ratio is not known precisely in advance. The use of RSS can mitigate this ethical concern by reducing the sample size needed while still providing acceptably precise estimates. Secondly, the educational setting is one in which ranking of units by judgment is feasible. RSS provides a way to exploit purely judgment based opinions about the nature of sampling units, without biasing the resulting inference. For example, teachers having daily interactions with students may have valuable “soft” information that would allow them to rank a small number of students at a time with respect to their likely relative performance. There are school rankings published by various sources which could be used for ranking schools. Ranking may also be accomplished among schools using expert opinions or prior knowledge. Stovall (2012) investigated the use of RSS in educational statistics, which focused on improving estimation of intervention effects by incorporating ranking information through covariates. She showed via a simulation-based approach that the application is promising. Unlike Stovall (2012), we focus on pure judgment ranking so that no covariate information is needed for data collection or analysis.

In the literature, examples of incorporating RSS into existing sampling designs are scant; to our best knowledge, there are only four. Muttlak and McDonald (1992) proposed using RSS with a line intercept method and Sroka (2008) proposed using RSS with stratified sampling designs, both with RSS at the final stage. Sud and Mishra (2006) and Nematollahi et al. (2008) considered two-stage RSS designs for clustered data. All four focused on estimation of population mean/sum and developed their methodology from the point of view of finite population sampling theory; more specifically, they used a Horvitz-Thompson type approach for deriving mean and variance. Each argued the advantages based on an empirical example or simulation, without the benefit of a theoretical development showing efficiency's relationship to the design parameters. In addition, Sud and Mishra (2006) and Nematollahi et al. (2008) relied on restrictive assumptions for their methodological development: they both required N (the total number of primary units in the population) and all M_is (the total number of secondary units within each primary unit) to be known; and Sud and Mishra (2006) further required that M_is be equal across all primary units. By contrast, we focus on both estimating and testing treatment effects, rather than estimating population mean. Thus, the scope of our applications would be much wider. Our approach is model based, which entails employing hierarchical linear models (HLMs, Bryk and Raudenbush 2002), a standard framework for analyzing data collected from CRDs; and except for the assumed linear structure, our approach is nonparametric, imposing the minimum level of assumptions. Further, our work generates theoretical results that allow us to quantify the magnitude of the gain from integrating RSS at different levels, and examine the impact of design parameters analytically so that useful guidelines and predictions can be available at the design stage.

Proofs of all stated theorems are available in Sections S1-S5 of Supplementary Material (SM).

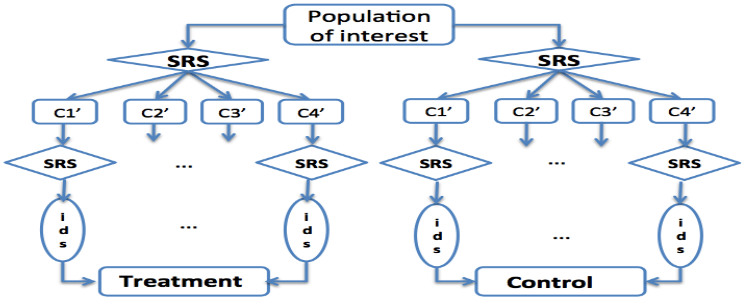


Figure 1: An illustration of a cluster randomized design with two-stage sampling. “C” stands for “cluster” and “ids” stands for “individuals”.

2 Design, Data and Model

Consider a traditional CRD with two-stage sampling in Figure 1, where SRS is used to select clusters and then individuals within each selected cluster (e.g., students within classrooms, patients within hospitals). Outcomes are measured at the individual level (say Y) and the mean scores from the treatment and control will be formally compared via statistical analysis. Throughout this paper, we use k to index individuals, j to index clusters, and i to index groups ($i = 1$ for treatment and 0 for control).

2.1 Design

We consider procedures that incorporate balanced RSS into the CRD using three ranking schemes, in which RSS is used to select clusters only, individuals only, and both, respectively.

Cluster-level ranking (Scheme i): Here, we use RSS instead of SRS to select clusters in the CRD. We first specify the set size H_i^c and the number of cycles m_i^c for group i (the superscript “c” means “cluster”). Note that H_i^c is typically chosen to be small, say $2 \sim 10$, since ranking a larger number of units cheaply with reasonable accuracy is difficult. Next, in each cycle, we repeat the following procedure for $h = 1, \dots, H_i^c$ to select H_i^c clusters for group i $(i = 0, 1)$: (i) Randomly select H_i^c clusters from the population. Without yet knowing any values of the outcome Y, rank clusters within the set based on perception of relative values of its cluster mean. (ii) Identify the cluster with rank h and assign it to group i. Discard the other $H_i^c - 1$ clusters. At the end of the repetition, $J_i = m_i^c \times H_i^c$ clusters are selected to enter each group i. Next, within each selected cluster j in group i, randomly select $K_{j(i)}$ individuals to receive treatment i. Note that we use "()" in subscripts to clearly indicate the nested structure, so $j(i)$ means the jth cluster nested in treatment i. Finally, all the individuals receiving the treatment or control are evaluated, and outcomes recorded for each.

By incorporating auxiliary information about clusters obtainable in the form of judgement ranks, we may increase the chance that the sampled clusters accurately represent the true population. This results in an increase in information for a given sample size (or alternatively a decrease in cost for the same information) that is similar in its source to that from blocking, except that no auxiliary information about the clusters external to the sample is required.

Individual-level ranking (Scheme ii): Another way to integrate RSS into the CRD is to rank individuals within the SRS of clusters. We can proceed as follows. First determine the set size $H^{id}_{j(i)}$ and number of cycles $m^{id}_{j(i)}$ for cluster j selected to enter group i (the superscript "id" means "individual"). Again, the set sizes are chosen not to be large, to control ranking error. Next, repeat in each cycle the following procedure for $h = 1, \ldots, H^{id}_{j(i)}$ to select $H^{id}_{j(i)}$ individuals for cluster j: (i) Randomly select $H^{id}_{j(i)}$ individuals from cluster j and rank them based on perception of relative values of Y. (ii) Identify the individual with rank h and assign him to group i if cluster j is in group i. Discard the other $H^{id}_{j(i)} - 1$ individuals. At the end of the repetition, $K_{j(i)} = m^{id}_{j(i)} \times H^{id}_{j(i)}$ individuals are selected from cluster j to receive treatment i. Here, it is not necessary to use the same ranker for different clusters. Finally, all individuals receiving the treatment or control are evaluated, and outcomes recorded for each.

As with cluster-level RSS, this design provides a way to obtain a more representative sample, but of individuals within each cluster. This is another approach to improve estimation efficiency and statistical power to detect the treatment effect.

Both-level ranking (Scheme iii): In certain situations, both clusters and individuals can be ranked inexpensively. This allows use of RSS at both levels for further improvement in statistical inference. The cluster-level ranking requires information about clusters while the individual-level ranking requires knowledge about individuals' traits, which may be complementary. There is no need for the same rankers to be used at both levels. Note that it is difficult to make use of multiple rankers in the original one-level design of RSS. Here, the hierarchical data structure, combined with ranking at different levels, may allow us to incorporate information from multiple rankers naturally.

Flexibility of RSS-structured CRDs: No matter at which level(s) RSS is implemented, the new designs allow for varying sample sizes in different groups and within different clusters, as shown above. This implies that RSS can be implemented in only one of the two groups, or in some but not all selected clusters. Further, the number of individuals per cluster is not required to be equal. This is indeed flexible in practice. As will be shown in later sections, all the theories and methods we develop allow for the same flexibility.

2.2 Model and Data

Let $Y_{k(ij)}$ denote the measured outcome score of individual k in cluster j under treatment i, for $k = 1, \ldots K_{j(i)}$, $j = 1, \ldots J_i$, $i = 0, 1$, where $K_{j(i)}$ is the number of sampled individuals in cluster j under treatment i, and J_i is the number of clusters selected for treatment i. To model Y under the traditional CRD in Figure 1, it is natural to adopt a HLM to reflect the nested data structure, namely

$$Y_{k(ij)} = \mu + a_i + b_{j(i)} + r_{k(ij)}, \tag{1}$$

where μ is the mean score of the control group; a_i is the fixed effect of treatment i, with $a_0 \equiv 0$; $b_{j(i)}$ is the random effect of cluster j; and $r_{k(ij)}$ is the random error, reflecting the effect of individual k that has not been systematically accounted for by other terms in the model. The cluster effects $b_{j(i)}$'s are assumed to be identically distributed (i.i.d), following some unknown continuous distribution with mean $\mu_b = 0$ and finite variance σ_b^2; the errors $r_{k(ij)}$'s are assumed to be i.i.d from some unknown continuous distribution with mean $\mu_r = 0$ and finite variance σ_r^2. All $b_{j(i)}$'s and $r_{k(ij)}$'s are assumed independent. Under (1), individual scores from the same cluster are dependent, and the intra-class correlation (ICC) is given by $\sigma_b^2/(\sigma_b^2 + \sigma_r^2)$, while scores from different clusters are independent.

For data collected using a RSS-structured CRD, ranking information, in addition to outcomes, becomes available for each cluster or individual measured, depending on the ranking scheme. For scheme (i) ranking at the cluster level only, let $O^c_{j(i)}$ denote the (judgement) order of cluster j under treatment i among its own comparison set; then the data can be expressed by $\mathbf{D}^c = \{Y_{k(ij)}, O^c_{j(i)}\}$ given the design parameters $\{H^c_i, m^c_i, K_{j(i)}\}$ for $k = 1, \dots K_{j(i)}$, $j = 1, \dots J_i$ $(J_i = H^c_i \times m^c_i)$, $i = 0, 1$. For scheme (ii) ranking at the individual level only, let $O^{id}_{k(ij)}$ denote the (judgement) order of individual k in cluster j under treatment i among its own comparison set; then the data can be expressed by $\mathbf{D}^{id} = \{Y_{k(ij)}, O^{id}_{k(ij)}\}$ given the design parameters $\{J_i, H^{id}_{j(i)}, m^{id}_{j(i)}\}$ for $k = 1, \dots K_{j(i)}$ $(K_{j(i)} = m^{id}_{j(i)} \times H^{id}_{j(i)})$, $j = 1, \dots J_i$, $i = 0, 1$. For scheme (iii) ranking at both levels, the data can be expressed by $\mathbf{D}^b = \{Y_{k(ij)}, O^c_{j(i)}, O^{id}_{k(ij)}\}$ (the superscript "b" means "both") given the design parameters $\{H^c_i, m^c_i, H^{id}_{j(i)}, m^{id}_{j(i)}\}$ for $k = 1, \dots K_{j(i)}$ $(K_{j(i)} = m^{id}_{j(i)} \times H^{id}_{j(i)})$, $j = 1, \dots J_i$ $(J_i = m^c_i \times H^c_i)$, $i = 0, 1$. Note that ranking at a single level is a special case of ranking at both levels: $H^{id}_{j(i)} = 1$ and $m^{id}_{j(i)} = K_{j(i)}$ for ranking at the cluster level; $H^c_i = 1$ and $m^c_i = J_i$ for ranking at the individual level. To avoid cumbersome notations, the superscripts "c", "id", "b" are dropped when there is no ambiguity.

The HLM in (1) indicates that within the same treatment, the mean difference in Y among clusters is reflected through the cluster effect b; and the difference among individuals from the same cluster is reflected through the individual effect r. Thus, when ranking clusters within a treatment, ranking by Y is equivalent to ranking by b; and when ranking individuals within a cluster, ranking by Y is equivalent to ranking by r. That's why we assume that judgement ranking is done based on b or r, although their values are not directly observable. In the case of perfect ranking, this would be equivalent to ranking clusters/individuals by b/r or any monotone transformation of b/r. We also assume that the ranking mechanisms in the RSS-structured CRDs are consistent (Chap. 2, pg. 12 in Chen et al. 2006).

We note that in our subsequent sections, the model (1) is used in conjunction with the ranking information, available through the added structure of RSS, for both methodological and theoretical developments (see technical detail in our supplemental material). For example, under scheme (i), given $O^c_{j(i)} = h$, we have $Y_{k(ij)} = \mu + a_i + b^*_{j(i)} + r_{k(ij)}$, where $b^*_{j(i)}$, defined as $b_{j(i)}|O^c_{j(i)} = h$, follows the distribution of the hth (judgement) order statistic of b.

3 RSS Estimator of Treatment Effect

Under the model (1), the treatment effect Δ is given by $\Delta = \mu_1 - \mu_0 = a_1$, where $\mu_i = \mu + a_i$ is the mean score of the treatment/control group for $i = 0/1$.

An advantage of the original (one-level) balanced RSS design is that the RSS estimator of the population mean has the same simple form as the sample mean, as an average of the quantified observations. This estimator is unbiased and distribution-free; and it is at least as efficient as the sample mean from a SRS of the same size, even with imperfect ranking. Thus, for RSS-structured CRDs, it is natural to consider a nonparametric estimator in the form of

$$\hat{\Delta}_{RSS} \quad = \hat{\mu}_1 - \hat{\mu}_0 = \quad \frac{1}{J_1}\sum_{j=1}^{J_1} \bar{Y}_{j(1)} - \frac{1}{J_0}\sum_{j=1}^{J_0} \bar{Y}_{j(0)} \tag{2}$$

where $\hat{\mu}_i$ denotes the RSS estimator of μ_i, and $\bar{Y}_{j(i)}$ is the average score of individuals in cluster j receiving treatment i. The subscript "RSS" in $\hat{\Delta}_{RSS}$ indicates that the estimator uses data collected from a RSS-structured CRD. Correspondingly, we use $\hat{\Delta}_{SRS}$ to denote the estimator that has the same form as (2), but uses data collected from the original CRD.

3.1 Unbiasedness, variance and efficiency

Under schemes (i) and (iii), for each rank stratum h in treatment i, we define the index set $\mathcal{J}_i(h)$=$\{j$: cluster j in treatment i has rank $h\}$, where $h = 1, \ldots H_i$; further, we let $\mu_{b.ih} \equiv E\left[b_{j(i)} \mid O_{j(i)} = h\right]$ and $\sigma^2_{b.ih} \equiv Var\left[b_{j(i)} \mid O_{j(i)} = h\right]$, the mean and variance of the hth judgment order statistic of the cluster effect b, respectively. Similarly, under schemes (ii) and (iii), for each rank stratum h' within cluster j of treatment i, we define the index set $\mathcal{K}_{j(i)}(h')$=$\{k$: individual k within cluster j of treatment i has rank $h'\}$, where $h' = 1, \ldots H_{j(i)}$; further, we let $\mu_{r.ijh'} \equiv E\left[r_{k(ij)} \mid O_{k(ij)} = h'\right]$, and $\sigma^2_{r.ijh'} \equiv Var\left[r_{k(ij)} \mid O_{k(ij)} = h'\right]$, the mean and variance of the hth judgment order statistic of the individual effect r, respectively.

The following theorems describe some optimality and finite-sample properties of $\hat{\Delta}_{RSS}$.

Theorem 1. *The estimator $\hat{\Delta}_{RSS}$ is a least squares estimator, which, combined with $\hat{\mu} \equiv \hat{\mu}_0 =$*

$\sum_{j=1}^{J_0} \bar{Y}_{j(0)}/J_0$, *minimizes the sum of weighted squared distances from each observed* $Y_{k(ij)}$ *to its conditional mean given the available ranking information, with weights* $\{1/K_{j(i)}\}$. *That is, under the full ranking scheme (iii) (without loss of generality),*

$$\left(\hat{\mu}, \hat{\Delta}_{RSS}\right) = arg\min_{\mu,\Delta} \sum_{i=0}^{1} \sum_{h=1}^{H_i} \sum_{j \in \mathcal{J}_i(h)} \frac{1}{K_{j(i)}} \sum_{h'=1}^{H_{j(i)}} \sum_{k \in \mathcal{K}_{j(i)}(h')} \left[Y_{k(ij)} - E\left(Y_{k(ij)} \mid O_{j(i)} = h, O_{k(ij)} = h'\right)\right]^2,$$

where $E\left(Y_{k(ij)} \mid O_{j(i)} = h, O_{k(ij)} = h'\right) = \mu + a_i + \mu_{b.ih} + \mu_{r.ijh'}$.

Theorem 2. *The estimator* $\hat{\Delta}_{RSS}$ *is unbiased, i.e.,* $E\hat{\Delta}_{RSS} = \Delta$ *under all three ranking schemes. For scheme (i) ranking at the cluster level,*

$$Var(\hat{\Delta}^{c}_{RSS}) = \sum_{i=0}^{1} \left[\frac{1}{J_i H_i} \sum_{h=1}^{H_i} \sigma^2_{b.ih} + \frac{1}{J_i^2} \sum_{j=1}^{J_i} \frac{1}{K_{j(i)}} \sigma^2_r\right], \tag{3}$$

where $J_i = m_i \times H_i$; *for scheme (ii) ranking at the individual level,*

$$Var(\hat{\Delta}^{id}_{RSS}) = \sum_{i=0}^{1} \left[\frac{1}{J_i} \sigma^2_b + \frac{1}{J_i^2} \sum_{j=1}^{J_i} \frac{1}{K_{j(i)} H_{j(i)}} \sum_{h'=1}^{H_{j(i)}} \sigma^2_{r.ijh'}\right], \tag{4}$$

where $K_{j(i)} = m_{j(i)} \times H_{j(i)}$; *and for scheme (iii) ranking at both levels,*

$$Var(\hat{\Delta}^{b}_{RSS}) = \sum_{i=0}^{1} \left[\frac{1}{J_i H_i} \sum_{h=1}^{H_i} \sigma^2_{b.ih} + \frac{1}{J_i^2} \sum_{j=1}^{J_i} \frac{1}{K_{j(i)} H_{j(i)}} \sum_{h'=1}^{H_{j(i)}} \sigma^2_{r.ijh'}\right], \tag{5}$$

where $J_i = m_i \times H_i$ *and* $K_{j(i)} = m_{j(i)} \times H_{j(i)}$.

Since $\hat{\Delta}_{RSS}$ and $\hat{\Delta}_{SRS}$ are both unbiased, Theorem 3 indicates that the relative efficiency (RE) for $\hat{\Delta}_{RSS}$ versus $\hat{\Delta}_{SRS}$, defined as the ratio of Mean Square Errors (i.e., $RE \equiv MSE(\hat{\Delta}_{SRS})/MSE(\hat{\Delta}_{RSS})$), is the same as the relative precision (defined as the ratio of the variances), which is always no less than 1, regardless of ranking error.

Theorem 3. $\hat{\Delta}_{RSS}$ *is at least as efficient as* $\hat{\Delta}_{SRS}$, *i.e.,* $Var\left(\hat{\Delta}_{RSS}\right) \leq Var\left(\hat{\Delta}_{SRS}\right)$ *under all three ranking schemes, where*

$$Var\left(\hat{\Delta}_{SRS}\right) = \sum_{i=0}^{1} \left[\frac{1}{J_i} \sigma^2_b + \frac{1}{J_i^2} \sum_{j=1}^{J_i} \frac{1}{K_{j(i)}} \sigma^2_r\right]. \tag{6}$$

3.2 Impacts of design parameters and ranking schemes

We illustrate how much gain can be achieved through the incorporation of RSS into CRDs and how design parameters/ranking schemes affect the relative efficiency, through examining completely balanced designs (i.e., balance in both treatments and clusters).

For ranking at the cluster level, complete balance means that $H_i \equiv H^c$ and $m_i \equiv m^c$ so that $J_i \equiv H^c \times m^c$ for $i = 0, 1$ (i.e., the same RSS design is used in both treatments), and the number of individuals $K_{j(i)} \equiv K$ is equal for all the selected clusters. Then from (3) and (6),

$$RE^c = \left(\sigma_b^2 + \frac{\sigma_r^2}{K} \right) / \left(\frac{\sum_{h=1}^{H^c} \sigma_{b.h}^2}{H^c} + \frac{\sigma_r^2}{K} \right), \tag{7}$$

where $\sigma_{b.h}^2$ is the variance of the hth judgment order statistic (relative to H^c ordered observations in total) of the cluster effect b. For ranking at the individual level, complete balance means $J_i \equiv J$ for $i = 0, 1$ (the number of clusters is equal in both treatments), and for all selected clusters, $H_{j(i)} \equiv H^{id}$, and $m_{j(i)} \equiv m^{id}$ so that $K_{j(i)} \equiv K = H^{id} \times m^{id}$ (i.e., the same RSS design is used in the clusters). Then from (4) and (6),

$$RE^{id} = \left(\sigma_b^2 + \frac{\sigma_r^2}{K} \right) / \left(\sigma_b^2 + \frac{\sum_{h=1}^{H^{id}} \sigma_{r.h}^2}{H^{id} K} \right), \tag{8}$$

where $\sigma_{r.h}^2$ is the variance of the hth judgment order statistic (relative to H^{id} ordered observations in total) of the individual effect r.

For completely balanced designs, RE has an intuitive interpretation – it shows reduction in the number of clusters achievable from RSS. For example, if RE=2, then SRS requires twice as many clusters as RSS to achieve the same precision. The following propositions investigate the impact of design parameters on RE when the distributions of b and r are both fixed.

Proposition 1. *Suppose a CRD that incorporates RSS at the cluster level is completely balanced with design parameters* (H^c, m^c, K).

(i) RE^c does not depend on the number of cycles m^c.

(ii) For perfect ranking, as $H^c \uparrow$ (increases), $RE^c \uparrow$ for constant K. This relationship also holds for imperfect ranking under the assumption of the linear ranking error model in (11), as will be described in Section 3.4.

(iii) As $K \uparrow$, $RE^c \uparrow$ for constant H^c.

(iv) RE^c does not depend on the distribution of the individual effect r, other than through its variance σ_r^2.

Proposition 2. *Suppose a CRD that incorporates RSS at the individual level is completely balanced with design parameters* (J, H^{id}, m^{id}).

(i) RE^{id} does not depend on the number of clusters J.

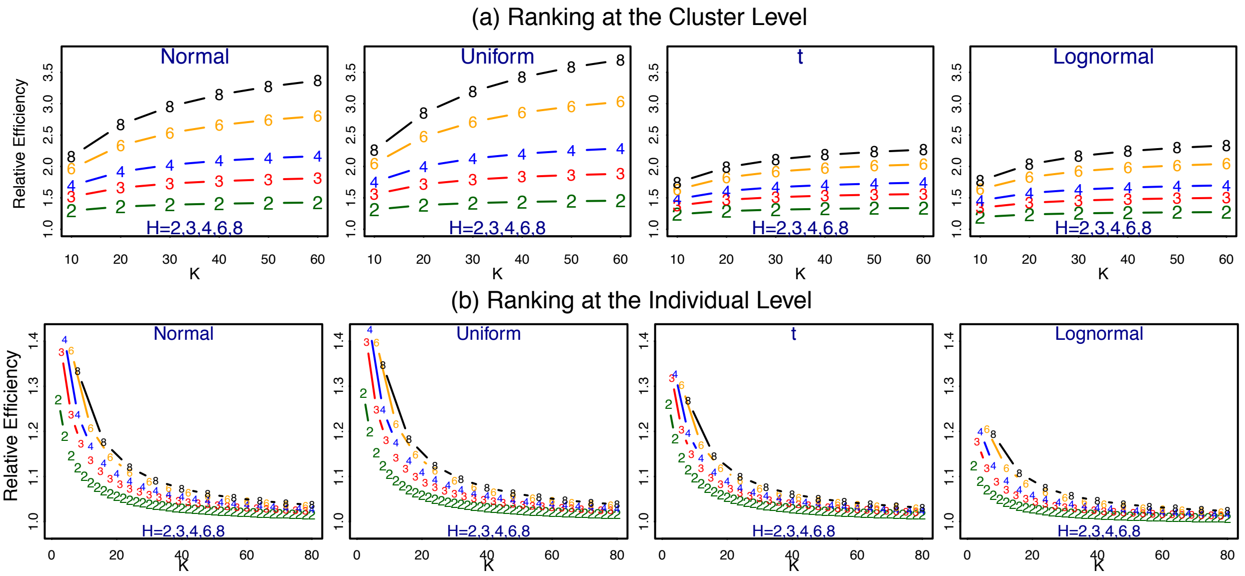
(ii) For perfect ranking, as $H^{id} \uparrow$, $RE^{id} \uparrow$ for constant K (so that $m^{id} \downarrow$). This relationship also holds for imperfect ranking under the linear ranking error model in (14), as will be described in Section 3.4.

(iii) As $K \uparrow$ (or equivalently $m^{id} \uparrow$), $RE^{id} \downarrow$ for constant H^{id}; and as $K \to +\infty$ (or equivalently $m^{id} \to +\infty$), $RE^{id} \to 1$.

(iv) RE^{id} does not depend on the distribution of the cluster effect b other than through its variance σ_b^2.

The top and bottom panels of Figure 2 present theoretical values of RE for (perfect) ranking at the cluster and individual levels, respectively, for completely balanced CRDs. We consider different values of (H, K) and four different distributions: normal, uniform, t and lognormal (lognormal is shifted so that the mean is zero). Each subplot, which corresponds to one of the distributions, contains five lines for $H = 2, 3, 4, 6, 8$; and the number connecting each line indicates the value of H. For applications in education, ICC is typically between 0.15 and 0.25 (Hedges and Hedberg 2007), so we set $\sigma_b^2 = 1$ and $\sigma_r^2 = 4$ ($ICC = 0.2$) for all cases. For the t distribution, we set df=3, so that it has very heavy tails. Note that in Figure 2(a) the distributions are for the cluster effect b, since Proposition 1(iv) states that the distribution of r is irrelevant to RE^c, meaning no specification on the distribution of r is necessary. Similarly, in Figure 2(b), based on Proposition 2(iv), the distributions are for the individual effect r and there is no specification necessary for the distribution of b.

Figure 2(a) shows results that are consistent with Proposition 1 (i.e., $RE^c \uparrow$ as $H^c \uparrow$ or $K \uparrow$ while keeping the other constant) and that the improvement can be substantial, even


(a) Ranking at the Cluster Level

(b) Ranking at the Individual Level

Figure 2: Theoretical values of relative efficiency of $\hat{\Delta}_{RSS}$ versus $\hat{\Delta}_{SRS}$ under completely balanced CRDs, with perfect ranking done at the cluster level and at the individual level. K represents the number of individuals per cluster. In Panel (a), $N(0,1^2)$, $U(-1.74,1.74)$, $t_3/\sqrt{3}$, and $LN(0,0.481)-1.27$ were considered for the cluster effect b, all with mean zero and unit variance; in Panel (b), $N(0,2^2)$, $U(-3.47,3.47)$, $t_3\times 2/\sqrt{3}$, and $LN(0,0.941)-1.6$ were considered for the individual effect r, all with mean zero and variance 4.

for the lognormal case. Figure 2(b) shows results that are consistent with Proposition 2 (i.e., $RE^{id}\uparrow$ as $H^{id}\uparrow$ for constant K, but $RE^{id}\downarrow$ as $K\uparrow$ for constant H^{id}) and that the benefit from RSS disappears more quickly for small than large H^{id}, as the sample size per cluster increases. In both panels, we can observe that the improvement is largest for the uniform distribution, followed by the normal distribution; and it becomes smaller for the t and lognormal distributions, perhaps due to the heavy tail of the t distribution and the skewness of the lognormal distribution.

One observation from comparison of Panel (a) to (b) in Figure 2 is that RE for individual-level ranking is generally smaller than that for cluster-level ranking. This is not surprising. Comparing (7) and (8) side by side, we find that the role that σ_b^2 plays in (7) is equivalent to the role that σ_r^2/K plays in (8), and the gain from the use of RSS is reflected through replacing the variance terms by the corresponding average variability of their (judgement) order statistics. When ICC is within the interval [0.15, 0.25], as typical in educational applications, $3\sigma_b^2\le\sigma_r^2\le$

$6\sigma_b^2$. So when K is larger than 6, the terms related to the cluster effect dominate those of the individual effect in both (7) and (8). This leads to less significant improvement for ranking at the individual level. Nevertheless, when K is small (e.g., ≤ 10), noticeable improvement can be achieved, which could prove useful for educational studies where the number of potential subjects per cluster is small (e.g., Project Maximize, Allor et al. 2010).

We proceed to examine the impact of design parameters in the case of ranking at both levels and further compare the relative efficiency for the three different ranking schemes.

Proposition 3. *Suppose a CRD that incorporates RSS at both levels is completely balanced with design parameters* $(H^c, m^c, H^{id}, m^{id})$.

(i) RE^b does not depend on the number of cycles m^c.

(ii) For perfect ranking, as $H^c \uparrow$, $RE^b \uparrow$ for constant H^{id} and m^{id}; and as $H^{id} \uparrow$, $RE^b \uparrow$ for constant H^c and K (so that $m^{id} \downarrow$). This relationship also holds for imperfect ranking under the assumption of the linear ranking error models in (11) and (14).

(iii) Suppose H^c and H^{id} are both held constant. Then as $K \uparrow$ (or equivalently $m^{id} \uparrow$), $RE^b \uparrow$ when $RE_r^{RSS(H^{id})} < RE_b^{RSS(H^c)}$; $RE^b \downarrow$ when $RE_r^{RSS(H^{id})} > RE_b^{RSS(H^c)}$; and RE^b remains constant C_0 if $RE_r^{RSS(H^{id})} = RE_b^{RSS(H^c)} \equiv C_0$. Further, as $K \to +\infty$ (or equivalently $m^{id} \to +\infty$), $RE^b \to RE_b^{RSS(H^c)}$ (see Section 3.3 for the definitions of $RE_r^{RSS(H^{id})}$ and $RE_b^{RSS(H^c)}$).

(iv) Consider three completely balanced designs with the same K but different ranking schemes, indexed by parameters (H^c, m_1^c, K), (J, H^{id}, m^{id}), and $(H^c, m_2^c, H^{id}, m^{id})$, respectively, where $K = H^{id}m^{id}$. Then $RE^b(H^c, m_2^c, H^{id}, m^{id}) \geq \max[RE^c(H^c, m_1^c, K),\ RE^{id}(J, H^{id}, m^{id})]$.

3.3 Connections with one-level RSS, role of ICC, etc.

We define $RE_b^{RSS(H^c)}$ to be the relative efficiency of a (one-level) balanced RSS of b's with set size H^c, which is given by $RE_b^{RSS(H^c)} \equiv H^c\sigma_b^2 / \sum_{h=1}^{H^c} \sigma_{b.h}^2$; and $RE_r^{RSS(H^{id})}$ is defined in the same manner. Under completely balanced CRDs, we establish connections with one-level RSS, investigate the relationship between RE and ICC, and provide the maximum value of

RE. It is also shown that RE is no larger than the one-level REs, due to the hierarchical nature of data involved.

Proposition 4. *Suppose a CRD that incorporates RSS at the cluster level is completely balanced with design parameters* (H^c, m^c, K)*. Then* RE^c *is connected to* $RE_b^{RSS(H^c)}$ *through the following relationship:*

$$RE^c = \left(K + \frac{1}{ICC} - 1\right) / \left(\frac{K}{RE_b^{RSS(H^c)}} + \frac{1}{ICC} - 1\right). \tag{9}$$

RE^c *is an increasing function of both* $RE_b^{RSS(H^c)}$ *and* ICC*; and* $RE^c \leq RE_b^{RSS(H^c)}$*. Further, for continuous* b*,*

$$RE^c \leq \left(K + \frac{1}{ICC} - 1\right) / \left(\frac{2K}{H^c + 1} + \frac{1}{ICC} - 1\right),$$

where the maximum is attainable with perfect ranking when b *is uniformly distributed.*

Proposition 5. *Suppose a CRD that incorporates RSS at the individual level is completely balanced with design parameters* (J, H^{id}, m^{id})*. Then* RE^{id} *is connected to* $RE_r^{RSS(H^{id})}$ *through the following relationship:*

$$RE^{id} = \left(K + \frac{1}{ICC} - 1\right) / \left[K + \frac{1}{RE_r^{RSS(H^{id})}} \cdot \left(\frac{1}{ICC} - 1\right)\right]. \tag{10}$$

RE^{id} *is an increasing function of* $RE_r^{RSS(H^{id})}$*, but a decreasing function of* ICC*; and* $RE^{id} \leq RE_r^{RSS(H^{id})}$*. Further, for continuous* r*,*

$$RE^{id} \leq \left(K + \frac{1}{ICC} - 1\right) / \left[K + \frac{2}{H^{id} + 1} \cdot \left(\frac{1}{ICC} - 1\right)\right],$$

where the maximum is attainable with perfect ranking when r *is uniformly distributed.*

As indicated in Proposition 5, the clustering effect, reflected through ICC, has a negative impact on RE when ranking is done at the individual level. This agrees with Ridout and Cobby (1987), who showed that correlation within ranked sets reduces RE. Proposition 4, however, indicates a positive impact of clustering on RE when ranking is done at the cluster level, an intriguing result not yet noticed in the literature.

It is easy to extend the results to the case of ranking at both levels. For example, $RE^b \leq \max(RE_b^{RSS(H^c)}, RE_r^{RSS(H^{id})})$; and RE^b is an increasing function of the one-level efficiencies

$RE_b^{RSS(H^c)}$ and $RE_r^{RSS(H^{id})}$. Also, given $(H^c, m^c, H^{id}, m^{id})$, we only need to know ICC from data to obtain the upper bound of RE^b, given by $(K + 1/ICC - 1)/\{2K/(H^c+1) + [2/(H^{id}+1)] \cdot (1/ICC - 1)\}$. If $H^c = H^{id} = H$, the upper bound is simply $(H+1)/2$, which is independent of both K and ICC. However, RE^b is not a monotone function of ICC. If the change in ICC does not affect either $RE_b^{RSS(H^c)}$ or $RE_r^{RSS(H^{id})}$ (e.g., the distributions of b and r are from scale families and ranking is perfect or follows (11) and (14)), then RE^b is an increasing function of ICC when $RE_r^{RSS(H^{id})} < RE_b^{RSS(H^c)}$, a decreasing function of ICC when $RE_r^{RSS(H^{id})} > RE_b^{RSS(H^c)}$, and constant C_0 if $RE_r^{RSS(H^{id})} = RE_b^{RSS(H^c)} \equiv C_0$.

3.4 Impact of Ranking Error

We model imperfect ranking in a manner similar to that of Dell and Clutter (1972). For ranking at the cluster level, we assume that ranking is carried out through a cluster-level latent variable X that is an imperfect assessment of b:

$$b = \beta_x (X - \mu_x) + \epsilon_x, \tag{11}$$

where β_x is the regression coefficient, μ_x is the mean of X; and ϵ_x is the error term, independent of X, with mean 0 and variance $\tau_{b|x}^2$ that reflects the remaining variability of clusters after taking X into account. Using Theorem 1 in Wang et al. (2006), we obtain

$$\sigma_{b.ih}^2 \quad = \quad \tau_{b|x}^2 + \beta_x^2 \sigma_{x.(ih)}^2 \tag{12}$$

where $\sigma_{x.(ih)}^2$ is the variance of the hth order statistic (relative to H_i ordered observations in total) of $X_{j(i)}$. Further, based on (11), we have

$$\sigma_{b.ih}^2 \quad = \quad \sigma_b^2 \left[1 - \rho^2 \left(1 - \frac{\sigma_{x.(ih)}^2}{\sigma_x^2}\right)\right], \tag{13}$$

where ρ is the correlation coefficient between b and X and σ_x^2 is the variance of X. In the case of perfect ranking, $\rho = 1$. From (13), we know that when the distribution of X and H_i are fixed, increasing ρ would decrease $\sigma_{b.ih}^2$ so that $Var(\hat{\Delta}_{RSS})$ in (3) decreases. This indicates better ranking accuracy (measured by ρ) would lead to higher efficiency in estimation.

Similarly, we can model imperfect ranking at the individual level, by assuming the exis-

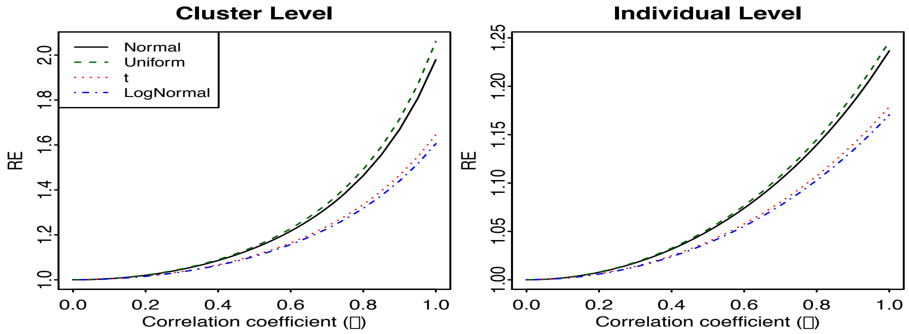


Figure 3: Theoretical values of relative efficiency of $\hat{\Delta}_{RSS}$ versus ρ under completely balanced CRDs for four different distributions of the ranking variables, including $N(0,1^2)$, $U(-1.74,1.74)$, $t_3/\sqrt{3}$, $LN(0,0.481)-1.27$.

tence of an individual-level variable Z that is an imperfect assessment of r:

$$r = \beta_z\,(Z-\mu_z)+\epsilon_z. \tag{14}$$

Let ρ be the correlation coefficient between r and Z. Then

$$\sigma^2_{r.ijh} = \sigma^2_r\left[1-\rho^2\left(1-\frac{\sigma^2_{z.(ijh)}}{\sigma^2_z}\right)\right], \tag{15}$$

where β_z , μ_z , ϵ_z, $\tau^2_{r|z}$, $\sigma^2_{z.(ijh)}$, and σ^2_z are defined similarly as in (11)-(13). As with the ranking of clusters, $\rho = 1$ represents the case of perfect ranking; and (15), combined with (4), indicates that better ranking accuracy leads to higher efficiency.

To show the impact of imperfect ranking, we plot theoretical values of RE versus the correlation coefficient ρ in Figure 3 for the four different distributions under completely balanced CRDs. The left panel is for ranking at the cluster level, where we set $H = 4$, $K = 25$, $\sigma^2_b = \sigma^2_x = 1$ and $\sigma^2_r = 4$; and the right panel is for ranking at the individual level, where we set $H = 4$, $m = 2$, $\sigma^2_b = \sigma^2_z = 1$ and $\sigma^2_r = 4$. Note that the distributions are for the ranking variable X or Z; the distributions of the ranking error terms e_x and e_z are irrelevant (other than through their variances) to RE, as indicated in (13) and (15).

Figure 3 shows that as $\rho\uparrow$, $RE\uparrow$. For ranking at the cluster level, even when ρ decreases to 0.6 (i.e., the ranking variable X or Z only explains 36% of the variation in b or r), the improvement over SRS is still sizable. As for the different distributions, the improvement follows the order that we observed in Section 3.3: Uniform>Normal >t>Lognormal, where the first two show quite obvious difference from the last two.

3.5 Asymptotic properties

Let $J. = J_0 + J_1$ be the total number of clusters in a RSS-structured CRD, and $\pi_i = \lim_{J.\to+\infty} J_i/J.$, $i = 0, 1$. In the next theorem, we discuss the asymptotic properties of $\hat{\Delta}_{RSS}$ as $J. \to +\infty$. Here, $J.$ goes to infinity in such a way that the π_i's exist and are bounded away from zero, and m_i's (no. of cycles) go to infinity (so that the J_i's go to infinity), but the set sizes H_i's are predetermined and finite. Note that for ranking at the individual level, H_i is fixed at 1 so that $J_i = m_i$ goes to infinity.

Theorem 4. *Assume that (i) the latent random variables b (the cluster effect) and r (the individual effect) satisfy that for some $\delta > 0$, $E\left|b\right|^{2+\delta} < +\infty$ and $E\left|r\right|^{2+\delta} < +\infty$; and (ii) π_is exist and are bounded away from zero.*

1. *Under scheme (i) ranking at the cluster level, as $J. \to +\infty$,*
$$\sqrt{J.}(\hat{\Delta}_{RSS} - \Delta) \to N\left(0, \sum_{i=0}^{1} \frac{1}{\pi_i}\left[\frac{1}{H_i}\sum_{h=1}^{H_i} \sigma^2_{b.ih} + w_i\sigma^2_r\right]\right) \tag{16}$$
where $w_i \equiv \lim_{J_i\to+\infty} \frac{1}{J_i}\sum_{j=1}^{J_i} \frac{1}{K_{j(i)}}$ that is assumed to exist.

2. *Under scheme (ii) ranking at the individual level, as $J. \to +\infty$,*
$$\sqrt{J.}(\hat{\Delta}_{RSS} - \Delta) \to N\left(0, \sum_{i=0}^{1} \frac{1}{\pi_i}\left[\sigma^2_b + \tilde{\sigma}^2_{r.i}\right]\right) \tag{17}$$
where $\tilde{\sigma}^2_{r.i} \equiv \lim_{J_i\to+\infty} \frac{1}{J_i}\sum_{j=1}^{J_i} \frac{1}{H^2_{j(i)} m_{j(i)}} \sum_{h'=1}^{H_{j(i)}} \sigma^2_{r.ijh'}$ that is assumed to exist.

3. *Under scheme (iii) ranking at both levels, as $J. \to +\infty$,*
$$\sqrt{J.}(\hat{\Delta}_{RSS} - \Delta) \to N\left(0, \sum_{i=0}^{1} \frac{1}{\pi_i}\left[\frac{1}{H_i}\sum_{h=1}^{H_i} \sigma^2_{b.ih} + \tilde{\sigma}^2_{r.i}\right]\right) \tag{18}$$
where $\tilde{\sigma}^2_{r.i}$, as defined above, is assumed to exist.

It follows directly from Theorem 4 that $\hat{\Delta}_{RSS}$ is a consistent estimator of Δ.

Finally, we illustrate the asymptotic properties in Theorem 4 using the completely balanced CRDs, where all the limits exist as $J. = 2J \to +\infty$. For scheme (i), we have $\pi_i \equiv 0.5$ and $w_i \equiv 1/K$, and (16) becomes $\sqrt{J}(\hat{\Delta}_{RSS} - \Delta) \to N\left(0, 2\sum_{h=1}^{H^c} \sigma^2_{b.h}/H^c + 2\sigma^2_r/K\right)$. For scheme (ii), we have $\tilde{\sigma}^2_{r.i} = \frac{1}{H^{id}K}\sum_{h=1}^{H^{id}} \sigma^2_{r.h}$, and (17) becomes $\sqrt{J}(\hat{\Delta}_{RSS} - \Delta) \to N\left(0,\, 2\sigma_b^2 + 2\sum_{h=1}^{H^{id}} \sigma^2_{r.h}/(H^{id}K)\right)$. Further, for scheme (iii), (18) becomes $\sqrt{J}(\hat{\Delta}_{RSS} - \Delta) \to N\left(0,\, 2\sum_{h=1}^{H^c} \sigma^2_{b.h}/H^c + 2\sum_{h=1}^{H^{id}} \sigma^2_{r.h}/(H^{id}K)\right)$.

4 Hypothesis Testing

4.1 Asymptotic pivotal method

For the unbiased estimator $\hat{\Delta}_{RSS}$ that was developed nonparametrically, it is natural to consider the asymptotic pivotal method for significance testing and confidence interval construction (Chen et al. 2006). To do so, we must obtain a reasonable estimator for the variance of $\hat{\Delta}_{RSS}$.

We first note a fact that is used in our derivation of such a variance estimator. Suppose Z_l, $l = 1, \dots L$, are independent random variables with mean 0 and variance σ_l^2. Let $\bar{Z}$ denotes the average, $\bar{Z} = \sum_{l=1}^{L} Z_l / L$. Then for $l = 1, \dots L$,

$$\sum_{l=1}^{L} E\left(Z_l - \bar{Z}\right)^2 = \left(1 - \frac{1}{L}\right) \sum_{j=1}^{L} \sigma_j^2. \tag{19}$$

Next, let

$$SSB(i,h) \equiv \sum_{j \in \mathcal{J}_i(h)} \left(\bar{Y}_{j(i)} - \hat{\mu}_{ih}\right)^2,$$

where $\hat{\mu}_{ih} = \sum_{j \in \mathcal{J}_i(h)} \bar{Y}_{j(i)}/m_i$. Note that

$$\begin{aligned} \bar{Y}_{j(i)} &= \mu + a_i + b_{j(i)} + \bar{r}_{j(i)}, \\ \hat{\mu}_{ih} &= \mu + a_i + \frac{1}{m_i} \sum_{j \in \mathcal{J}_i(h)} b_{j(i)} + \frac{1}{m_i} \sum_{j \in \mathcal{J}_i(h)} \bar{r}_{j(i)}, \end{aligned}$$

where $\bar{r}_{j(i)} = \sum_{k=1}^{K_{j(i)}} r_{k(ij)}/K_{j(i)}$.

So we have

$$SSB(i,h) = \sum_{j \in \mathcal{J}_i(h)} \left\{ \left[b_{j(i)} - \frac{1}{m_i} \sum_{j \in \mathcal{J}_i(h)} b_{j(i)} \right] + \left[\bar{r}_{j(i)} - \frac{1}{m_i} \sum_{j \in \mathcal{J}_i(h)} \bar{r}_{j(i)} \right] \right\}^2.$$

Since the two bracketed terms are independent, each with a mean of zero, we have

$$E\left[SSB(i,h)\right] = \sum_{j \in \mathcal{J}_i(h)} E\left[b_{j(i)} - \frac{1}{m_i} \sum_{j \in \mathcal{J}_i(h)} b_{j(i)} \right]^2 + \sum_{j \in \mathcal{J}_i(h)} E\left[\bar{r}_{j(i)} - \frac{1}{m_i} \sum_{j \in \mathcal{J}_i(h)} \bar{r}_{j(i)} \right]^2.$$

Applying the fact (19), we have

$$E\left[SSB(i,h)\right] = (m_i - 1)\,\sigma^2_{b.ih} + \left(1 - \frac{1}{m_i}\right) \sum_{j \in \mathcal{J}_i(h)} \frac{1}{H^2_{j(i)}} \sum_{h'=1}^{H_{j(i)}} \frac{\sigma^2_{r.ijh'}}{m_{j(i)}}.$$

So

$$E\left[\frac{SSB(i,h)}{(m_i - 1)\,m_i}\right] = \frac{1}{m_i}\sigma^2_{b.ih} + \frac{1}{m_i^2} \sum_{j \in \mathcal{J}_i(h)} \frac{1}{K_{j(i)}H_{j(i)}} \sum_{h'=1}^{H_{j(i)}} \sigma^2_{r.ijh'},$$

which, combined with (5), leads to the following theorem. Note that for ranking at the individual level only, $H_0 = H_1 = 1$ so that $h \equiv 1$ and $J_i = m_i$. Then $SSB(i,h) \equiv SSB(i) = \sum_{j=1}^{J_i} \left(\bar{Y}_{j(i)} - \hat{\mu}_i\right)^2$.

Theorem 5. *An unbiased estimator for $Var(\hat{\Delta}_{RSS})$ for all the three ranking schemes is given by*

$$\hat{V}\left(\hat{\Delta}_{RSS}\right) = \sum_{i=0}^{1} \frac{\sum_{h=1}^{H_i} SSB(i,h)}{H_i^2\,(m_i - 1)\,m_i}. \tag{20}$$

Further, for scheme (ii) ranking at the individual level only, it reduces to

$$\hat{V}\left(\hat{\Delta}^{id}_{RSS}\right) = \sum_{i=0}^{1} \frac{SSB(i)}{(J_i - 1)\,J_i}. \tag{21}$$

Now, based on the asymptotic normality of $\hat{\Delta}_{RSS}$ established in Theorem 4, we can construct a pivotal test statistic, $Z_{RSS} = (\hat{\Delta}_{RSS} - \Delta)/\hat{\sigma}_{\hat{\Delta}_{RSS}}$, where $\hat{\sigma}_{\hat{\Delta}_{RSS}} = \sqrt{\hat{V}\left(\hat{\Delta}_{RSS}\right)}$. Further, an equally tailed $100(1-\alpha)\%$ confidence interval of Δ is given by $[\hat{\Delta}_{RSS} - z_{1-\alpha/2}\hat{\sigma}_{\hat{\Delta}_{RSS}}$, $\hat{\Delta}_{RSS} - z_{\alpha/2}\hat{\sigma}_{\hat{\Delta}_{RSS}}]$, where $z_{\alpha/2}$ denotes the $(\alpha/2)$th quantile of $N(0,1)$. It is straightforward to test a hypothesis for Δ, either one-sided or two-sided, based on the pivot Z_{RSS}.

Finally, we mention that the extremely simple forms of $\hat{\Delta}_{RSS}$ and its variance estimator given in (20) or (21), which are both distribution-free, make the testing procedure easy to implement and widely applicable. Theoretically, the testing procedure requires large J_is. However, our simulations suggest that $J_i \geq 10$ is generally acceptable.

4.2 Power comparison

To examine the size and power of the proposed test based on $\hat{\Delta}_{RSS}$ for testing $H_0 : \Delta = 0$, we conducted simulation studies, where we performed the test using simulated data with different effect sizes from completely balanced RSS-structured CRDs, and compared the performance with the traditional F-test based on $\hat{\Delta}_{SRS}$ using data simulated from the corresponding original CRDs. Under (1), the effect size can be defined by $\delta \equiv (\mu_1 - \mu_0)/\sigma_Y =$

$\Delta/\sqrt{\sigma_b^2+\sigma_r^2}$. We considered five different values for δ: 0, 0.15, 0.25, 0.45 and 0.8, representing no effect (i.e. H_0 holds), a very small, small, medium and large effect, respectively. The first scenario is to examine the unbiasedness of the test size (i.e., whether the test achieves the nominal significance level α) and the others are to assess the test power. Throughout this paper, all tests were performed at $\alpha = 0.05$. The F-test (based on SRS) uses the test statistic $MS_{treatment}/MS_{cluster}$ that (asymptotically) follows an $F_{1,2J-2}$ distribution under H_0.

In our first study, we simulated ranking at the cluster level, with different (m, H) combinations ($m = 4, 6, 8, 10$ and $H = 2, 4, 6, 8$) and the number of individuals per cluster K fixed at 25. Imperfect ranking was simulated using the linear model (11), where ρ was set to 0.7, 0.9 and 1, the ranking variable $X \sim N(0, \sigma_b^2)$, $\beta_x = \rho$, and the error term $\epsilon_x \sim N(0, (1-\rho^2)\sigma_b^2)$. Again, we set $\sigma_b^2 = 1$ and $\sigma_r^2 = 4$ so that the cluster effect $b \sim N(0, 1)$ and the individual effect $r \sim N(0, 4)$. Outcome data were generated using the HLM (1), where we chose Δ according to the value of δ. For each setting, we generated 10,000 samples using the RSS-structured CRD, and 10,000 samples using the original CRD. The power (or size) of each test was computed as the proportion of times H_0 was rejected among the 10,000 replicates.

From Table 1, we can see that, to detect very small or small effects, the proposed test based on $\hat{\Delta}_{RSS}$ is consistently better than the traditional F-test, regardless of ranking errors. The improvement in power increases with either H or m, but the increase with H is more dramatic. For the medium-level effect size, the proposed test is more efficient than the F-test for $J < 30$; as J increases, the improvement diminishes since both tests work well. For the large effect size (results not reported in the table), both tests have power close to 1 except for very small J, where the proposed test shows a sizable improvement. As to the impact of ranking errors, the power of the proposed test increases as $\rho \uparrow 1$, as we expect. But even with $\rho = 0.7$, where X only explains 49% of the variability in b, the proposed test is always at least as powerful as the F-test, and the improvement over the F-test is notable in most settings. When H_0 holds, we find that the size of the traditional F-test seems to be unbiased, which maintains the nominal level 0.05 very well. The test based on $\hat{\Delta}_{RSS}$ rejects slightly more often than it should for small J, but as J increases, its test size becomes unbiased.

For a fair comparison in power, we obtained the 2.5th and 97.5th percentiles from the empirical distribution of Z_{RSS} using the data sets generated with $\delta = 0$ under each (H, m, ρ) setting, and used them instead of ± 1.96 as the critical values to reject $H_0 : \Delta = 0$, to control the size of the proposed test to be 0.05 exactly. Results are reported in Table S1 in SM. Although the size of gain becomes smaller in general after matching the type I error rates, all the above conclusions made from Table 1 remain valid.

		$\delta = 0$				$\delta = 0.15$				$\delta = 0.25$				$\delta = 0.45$			
		SRS	RSS			SRS	RSS			SRS	RSS			SRS	RSS		
H	m		0.7	0.9	1		0.7	0.9	1		0.7	0.9	1		0.7	0.9	1
2	4	0.04	0.07	0.07	0.07	0.09	0.13	0.13	0.14	0.16	0.23	0.25	0.26	0.42	0.54	0.57	0.60
	6	0.05	0.06	0.07	0.06	0.11	0.14	0.15	0.16	0.23	0.30	0.32	0.34	0.59	0.69	0.73	0.76
	8	0.05	0.06	0.06	0.06	0.14	0.17	0.18	0.19	0.29	0.36	0.40	0.42	0.73	0.81	0.85	0.87
	10	0.05	0.06	0.06	0.06	0.16	0.20	0.21	0.22	0.37	0.43	0.47	0.50	0.81	0.88	0.92	0.93
4	4	0.05	0.06	0.06	0.06	0.13	0.19	0.22	0.26	0.30	0.40	0.48	0.54	0.71	0.85	0.92	0.96
	6	0.05	0.05	0.06	0.06	0.19	0.25	0.30	0.34	0.42	0.54	0.65	0.72	0.89	0.96	0.99	1.00
	8	0.05	0.05	0.06	0.06	0.24	0.30	0.36	0.42	0.52	0.66	0.77	0.83	0.96	0.99	1.00	1.00
	10	0.05	0.05	0.05	0.06	0.28	0.37	0.45	0.50	0.64	0.77	0.85	0.91	0.98	1.00	1.00	1.00
6	4	0.05	0.06	0.06	0.06	0.18	0.25	0.33	0.41	0.42	0.57	0.71	0.80	0.89	0.97	0.99	1.00
	6	0.05	0.05	0.05	0.06	0.25	0.36	0.45	0.56	0.58	0.74	0.86	0.93	0.97	1.00	1.00	1.00
	8	0.05	0.05	0.05	0.05	0.33	0.44	0.56	0.67	0.71	0.85	0.94	0.98	0.99	1.00	1.00	1.00
	10	0.05	0.05	0.05	0.05	0.40	0.53	0.66	0.76	0.80	0.92	0.97	0.99	1.00	1.00	1.00	1.00
8	4	0.05	0.06	0.06	0.06	0.24	0.33	0.44	0.56	0.54	0.71	0.85	0.94	0.96	0.99	1.00	1.00
	6	0.05	0.05	0.05	0.05	0.32	0.46	0.60	0.73	0.72	0.87	0.95	0.99	1.00	1.00	1.00	1.00
	8	0.05	0.05	0.05	0.06	0.41	0.56	0.73	0.85	0.83	0.95	0.99	1.00	1.00	1.00	1.00	1.00
	10	0.05	0.05	0.05	0.05	0.49	0.66	0.81	0.91	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00

Table 1: Ranking at the cluster level: comparison of computed size and power for the tests based on the RSS and SRS designs under different levels of the effect size δ. For RSS, imperfect ranking was simulated using (11), where the correlation ρ was set to 0.7, 0.9 and 1 (perfect ranking).

In our second study, we simulated (imperfect) ranking at the individual level using (14), with the same settings for (m, H, ρ) as in the first study. The number of clusters per treatment J was fixed at 20. All the other parameters were the same as before. Table 2 reports the results for the comparison of the two tests. The size of the proposed test is slightly biased, but there is improvement in power over the F-test for all the cases we examined, even for medium or large effect sizes. Also, it is not surprising to see that the overall improvement

		$\delta = 0$				$\delta = 0.15$				$\delta = 0.25$				$\delta = 0.45$			
		SRS	RSS			SRS	RSS			SRS	RSS			SRS	RSS		
H	m		0.7	0.9	1		0.7	0.9	1		0.7	0.9	1		0.7	0.9	1
2	4	0.06	0.06	0.06	0.06	0.14	0.15	0.16	0.17	0.29	0.32	0.34	0.35	0.72	0.76	0.77	0.78
	6	0.05	0.06	0.06	0.06	0.15	0.17	0.17	0.17	0.33	0.35	0.36	0.37	0.78	0.80	0.81	0.82
	8	0.05	0.06	0.06	0.06	0.15	0.17	0.17	0.18	0.33	0.37	0.37	0.38	0.80	0.82	0.83	0.84
	10	0.05	0.06	0.06	0.05	0.16	0.18	0.18	0.18	0.35	0.38	0.39	0.39	0.80	0.83	0.84	0.84
4	4	0.05	0.06	0.06	0.06	0.15	0.18	0.18	0.17	0.34	0.38	0.39	0.40	0.79	0.83	0.85	0.85
	6	0.05	0.06	0.06	0.05	0.16	0.19	0.18	0.18	0.36	0.39	0.40	0.39	0.82	0.85	0.86	0.87
	8	0.05	0.06	0.06	0.06	0.16	0.18	0.19	0.19	0.38	0.40	0.41	0.41	0.83	0.86	0.87	0.87
	10	0.05	0.06	0.06	0.05	0.17	0.18	0.19	0.19	0.37	0.41	0.41	0.42	0.85	0.87	0.87	0.86
6	4	0.05	0.06	0.06	0.06	0.16	0.18	0.19	0.19	0.35	0.40	0.40	0.42	0.82	0.85	0.86	0.86
	6	0.05	0.06	0.06	0.06	0.17	0.18	0.19	0.19	0.37	0.40	0.42	0.42	0.83	0.87	0.87	0.88
	8	0.05	0.06	0.06	0.06	0.17	0.19	0.19	0.19	0.38	0.42	0.42	0.42	0.84	0.87	0.88	0.88
	10	0.05	0.06	0.06	0.06	0.17	0.19	0.19	0.20	0.39	0.42	0.42	0.42	0.85	0.87	0.88	0.88
8	4	0.05	0.06	0.06	0.06	0.16	0.18	0.19	0.19	0.37	0.41	0.42	0.42	0.83	0.86	0.87	0.87
	6	0.05	0.06	0.06	0.06	0.17	0.19	0.19	0.19	0.39	0.41	0.42	0.43	0.85	0.87	0.88	0.88
	8	0.05	0.06	0.06	0.06	0.17	0.19	0.19	0.19	0.39	0.42	0.42	0.42	0.85	0.88	0.88	0.88
	10	0.05	0.06	0.06	0.06	0.17	0.19	0.19	0.19	0.39	0.42	0.42	0.42	0.86	0.88	0.88	0.89

Table 2: Ranking at the individual level: comparison of computed size and power for the tests based on the RSS and SRS designs under different levels of the effect size δ. For RSS, imperfect ranking was simulated using (14), where the correlation ρ was set to 0.7, 0.9 and 1 (perfect ranking).

in Table 2 is not as large as that in Table 1, since RE^{id} is generally smaller than RE^{c}, as discussed in Section 3.2. In addition, the impact of ranking errors in test power is smaller in general, as opposed to ranking at the cluster level. As in the first study, we obtained empirical cutoff values for Z_{RSS} to control the type I error rate at 0.05, and reported results in Table S2 in SM. Still, the proposed test is at least as powerful as the F-test in all the cases, and the power is improved generally.

Finally, we mention that in practice, the power values in Tables 1 and 2 can be used for reference, but with the caution that the type I error is actually controlled at ~ 0.06 for small J. We also considered (shifted) lognormal distributions for the ranking variable X or Z in (11) or (14) (results not reported), and conclusions are essentially unchanged.

5 Examples

We conducted two empirical studies to illustrate the proposed methods. The first fits in the motivating setting described in the introduction, where we use data from an educational study HSLS09 (i.e., the High School Longitudinal Study of 2009), and examine how well our theory can predict actual gains (and their patterns) from using RSS rather than SRS and provide guidance in designing RSS-structured cluster randomized experiments. The second study provides an example using human dental data, which intends to demonstrate that the application scope of our methods is not limited to educational settings only. Further, it provides an interesting example in which, unlike educational applications, substantial gains can be achieved from ranking at the individual level rather than ranking at the cluster level. In addition, we assess gains in realistic situations by simulating an error-prone ranking process that attempts to mimic a natural one.

5.1 Educational data example

Data description and preprocessing: According to the National Center for Education Statistics (NCES), HLSL09 was conducted to monitor a national sample as the students progressed from the beginning of high school into post-secondary education, the workforce, and beyond. An important feature of this study is that the data are hierarchical by nature, and so the sampling design entailed selecting schools and then randomly selecting students from within those schools. The study involves a nationally representative sample of 944 high schools. An average of 25 ninth-graders per school were selected, for a total of ~24,000 students. These students were given a math assessment in the fall of 9th grade (2009) and again in the spring of most students' 11th grade year (2012), to gauge achievement gains in mathematics. For comprehensive information about this study, see the URL http://nces.ed.gov/surveys/hsls09/.

We downloaded the public-use student-level data of HSLS09 for the base year (2009) and the first follow-up year (2012) from the NCES website, in which all the school IDs are suppressed for confidentiality. To be able to group students by schools, we preprocessed the data and recovered student-school membership information for 763 schools and 16,975 students. This recovery was done based on two continuous variables (X2SchoolCli and C2CaseLoad) whose values, if not missing, are supposed to be the same for all students from the same school. For illustrative purposes, we removed schools with number of students < 20 and treated the remaining 472 schools and 12,533 students as the population(s) from which samples were drawn using either SRS or RSS. The students' 2012 math theta scores (X2TXMTH) were thought of as the response. Though our methodological development focuses on pure judgment ranking without using any covariate information, we generated rankings using the base-year number correct scores of 72 math items (X1TXMSCR), in order to simulate a large number of samples from the proposed designs easily. For cluster-level ranking, the school mean scores were used. The correlation between X2TXMTH and X1TXMSCR is 0.78 at the student level, and 0.89 at the school level. Further, treatment and control labels were randomly assigned to sampled schools. Since none of the students really received treatment, the means of the two groups should be the same; therefore, the parameter of interest, the treatment effect, was initially set to zero. We also imputed missing values of both variables based on a simple regression model between the two variables. ICC is about 0.25 for the preprocessed data.

Performance assessment: Suppose we are interested in estimating and testing the effect of an educational intervention program that is to enhance high school students' math ability. Before implementing the program, we want to investigate the performance of different (completely balanced) designs using the HSLS09 data. We fix J (the number of schools per group) at 15 and K (the number of students per school) at 6. Table 3 lists all possible RSS designs for the set size $H \in \{2, \ldots 10\}$, including two for scheme (i) ranking at the cluster level, three for scheme (ii) ranking at the individual level, and six for scheme (iii) ranking at both levels.

For each design, we generated 100,000 samples from the "population" (i.e., the entire dataset). We also generated 100,000 samples using the original CRD. Schools and students were drawn both with replacement in each simulated sample. Then we computed the approximate MSE for $\hat{\Delta}_{RSS}$ and $\hat{\Delta}_{SRS}$ based on these RSS and SRS samples, respectively, and their ratio was recorded as the empirical RE in Table 3. Further, to evaluate the performance in testing, we added $\Delta = \delta \cdot \sqrt{\sigma_b^2 + \sigma_r^2}$ to the treatment group, where δ was set to 0.15, 0.25, 0.45 and 0.8 as in Section 4.2, and σ_b^2 and σ_r^2 were estimated from the entire data set. The power (or size) values of the proposed test and F-test are reported in Table 3.

							Size/Power Comparison($\alpha = 0.05$)									
					Estimation		δ=0		δ=0.15		δ=0.25		δ=0.45		δ=0.8	
DesignID	H^c	m^c	H^{id}	m^{id}	RE	TRE	SRS	RSS	SRS	RSS	SRS	RSS	SRS	RSS	SRS	RSS
i-1	3	5	1	6	1.35	1.35	0.05	0.06	0.10	0.14	0.19	0.27	0.49	0.65	0.93	0.98
i-2	5	3	1	6	1.50	1.54	0.05	0.06	0.10	0.15	0.19	0.30	0.50	0.70	0.93	0.99
ii-1	1	15	2	3	1.05	1.07	0.05	0.06	0.10	0.12	0.19	0.22	0.49	0.55	0.94	0.96
ii-2	1	15	3	2	1.12	1.11	0.05	0.06	0.10	0.12	0.19	0.23	0.49	0.57	0.93	0.96
ii-3	1	15	6	1	1.18	1.16	0.05	0.06	0.10	0.12	0.19	0.24	0.50	0.59	0.93	0.97
iii-1-1	3	5	2	3	1.48	1.48	0.05	0.06	0.10	0.14	0.19	0.29	0.50	0.69	0.93	0.99
iii-1-2	3	5	3	2	1.54	1.55	0.05	0.06	0.10	0.15	0.19	0.30	0.50	0.70	0.93	0.99
iii-1-3	3	5	6	1	1.66	1.66	0.05	0.06	0.10	0.15	0.19	0.31	0.49	0.74	0.93	0.99
iii-2-1	5	3	2	3	1.69	1.70	0.05	0.06	0.10	0.16	0.19	0.32	0.49	0.74	0.93	1.00
iii-2-2	5	3	3	2	1.77	1.80	0.05	0.06	0.10	0.17	0.19	0.34	0.50	0.76	0.93	1.00
iii-2-3	5	3	6	1	1.92	1.95	0.05	0.06	0.10	0.17	0.19	0.36	0.49	0.79	0.93	1.00

Table 3: HSLS09 example: comparing performance of different (completely balanced) designs in estimation and testing. The 1st part of DesignID indicates the ranking scheme used; the 2nd part indicates which setting is used for the school level, where 1 stands for $(3, 5)$ and 2 for (5,3); and the 3rd part indicates which setting is used for the student level, where 1 stands for (2,3), 2 for (3,2) and 3 for (6,1). The power of the F-test based on SRS in each column should be constant, but subject to Monte Carlo errors.

In fact, our analytical results in Section 3.3 enable us to analyze the completely balanced designs without using actual data. For the two designs with ranking scheme (i), we predict that Design i-2 is better than Design i-1, since Proposition 1 indicates that larger H^c leads to higher efficiency while m^c is irrelevant. Among the three designs with scheme (ii), we predict that Design ii-3 is the best, followed by ii-2, and last ii-1, since Proposition 2 indicates that when K is fixed, increasing H^c increases efficiency. Similarly, according to Proposition 3, Design iii-s-3 should be the best among Designs iii-s-1, iii-s-2 and iii-s-3 that are all better than Design i-s, for $s = 1, 2$; and Design iii-s-t is better than both i-s and ii-t for all s, t. Finally, Design iii-2-3 should be the best among all the 11 designs, since iii-2-3 is better than iii-1-3, due to larger H^c. The RE and power values in Table 3 confirmed all the predictions based on our theories.

We computed theoretical RE values (TRE in Table 3) based on the linear ranking error models and normal distributions, using estimated ICC and correlation coefficients (for measuring the ranking quality) as reported before. TRE predicts the empirical RE surprisingly well. On one hand, the normality assumption roughly holds for the data. On the other hand, our formulas were all developed based on the HLM. The data in this example show obvious deviation from the model assumptions. For example, the variances of math scores from different schools range from 0.16 to 2.8 with mean 0.95, while the HLM requires a constant variance across schools. The close match between theoretical and empirical values is consistent with the finding that HLMs are robust for modeling nested data in the literature (e.g., Bryk and Raudenbush 2002); and it further suggests that our theories based on the HLM are resistant to violations of model assumptions, too.

As to testing, we again observe that the size of the proposed test is slightly biased while the size of the F-test is unbiased. But the power of the proposed test is consistently higher, for all the settings we examined. This remains true when empirical critical values were used to match the type I error rates (see Table S3 in SM). As δ gets larger, the improvement over the F-test first increases and then decreases for all the designs. The gain is quite large for small or moderate effect sizes, especially for the winning design iii-2-3.

In summary, this example has provided a successful proof of concept for the proposed methods using educational data. It also shows that our theories can be very helpful in guiding experiment design and predicting gains without using actual data. In fact, users only need design parameters, ICC and correlation values to get a reasonable estimate of RE based on normal distributions; to get an upper bound for RE, they only need design parameters and ICC. Besides the simplicity offered, this theoretical approach is useful when researchers only have a restricted access to complete data, as is typical in educational research.

5.2 Dental data example

Data description: Human tooth-size analysis plays an important role in orthodontics and forensic identification (Mitsea et al. 2014, Othman and Harradine 2007). Past studies have presented normative data with sex differences, and shown variations in tooth sizes among racial groups (e.g., Garn et al. 1967, Buschang et al. 1988, Lee et al. 2006). We consider a data set containing measurements on tooth widths for 296 subjects with normal occlusion, including 179 men and 117 women, who were selected from over 15,000 young Korean adults through a community dental health survey conducted in 1997-2005 (Lee et al. 2006). Typically, human adults have 32 teeth, among which there are 8 incisors, 4 canines, 8 premolars, and 12 molars (including 4 wisdom teeth). For each of the subjects, the widths of all the non-wisdom teeth were measured by digital Vernier calipers, a process that requires a 3-week training period to master; and these measurements are available in the data set. In the numerical study, we treated the 296 subjects as the population of interest, from which we simulated samples using both RSS-structured and original CRDs. Here, each person can be treated as a cluster, and each of his/her teeth as an individual within the cluster. So a (completely balanced) CRD entails selecting J persons in group i, and then selecting K teeth from the set of the 28 non-wisdom teeth for each selected person j (using either RSS or SRS). For each generated sample, the analysis goal was set to investigate whether there exists any gender difference in tooth sizes given the specific ethnicity (i.e., a Korean population); and so the comparison based on the HLM (1) was made between male and female groups. The width difference between Korean men and women estimated from the entire data set, $\hat{\Delta} \approx 0.28$mm (with p-value $< 10^{-11}$), was thought of as the true value of Δ; and ICC is about 0.04, indicating $\sigma_r^2 \gg \sigma_b^2$ (i.e., the within-cluster variability is much larger than the between-cluster variability).

Performance assessment: For RSS-structured CRDs, we considered all the three ranking schemes with K fixed at 4: (i) ranking at the cluster level only with $H^c \equiv 2$, $m^c = 5$ or 10 (so that $J = 10$ or 20 correspondingly); (ii) ranking at the individual level with $H^{id} = 2$ or 4 (so that $m^{id} = 2$ or 1, correspondingly), and $J = 10$ or 20; and (iii) ranking at both levels with all the four combinations in the first two schemes. In the real application, perfect ranking at the cluster level is not easy to obtain inexpensively, which requires ranking different persons based on the total width of all the 28 non-wisdom teeth. However, ranking with a decent quality can be easily done based on the width of one of the front teeth via visual inspection or other cheap methods without using digital calipers. For example, one can use a simple tool, similar to a compass with two sharp pointers, whose distance can be adjusted, to copy the width of a front tooth and then mark the width on a straight line for subsequent comparison. In this study, ranking at the cluster level was simulated using $N3L$, the width of the left mandibular cuspid (the third tooth from the center, one of the canines). The estimated correlation between $N3L$ and the total width of the 28 teeth is 0.78. Ranking at the individual level requires comparing different teeth of the same person. At either level that employs RSS, we simulated imperfect ranking in the following way: for any pairwise comparison during the ranking process, if the width difference between two teeth is larger than γ millimeters, we then use the actual rankings of the two teeth; otherwise, we randomly assign the ranks. This is to resemble a realistic situation when the difference is indistinguishable (without accurate measurement), but a ranker is forced to assign exact ranks. Obviously, doing so would induce nonlinear ranking errors that do not follow the model discussed in Section 3.4. We let γ vary from 0 to 2, with step size $\frac{1}{3}$ that is about $0.2\sigma_r$ (σ_r was estimated from the entire data set).

We generated 50,000 samples from the "population" for each setting considered. Again, samples were drawn with replacement at both stages. We computed the empirical RE and power from these samples as before. The top panels of Figure 4 compare the performance of different designs in estimating Δ. The RE values in all the settings are no less than 1, meaning that the RSS designs achieve better or comparable efficiency than the corresponding SRS designs, regardless of ranking errors. Among the three schemes, ranking at both levels, as expected, achieves higher efficiency than ranking at either level. However, we find that the gain is mostly from ranking at the individual level while ranking at the cluster level is only slightly better than SRS (and as γ increases, the gain diminishes). This seems to be interesting – as we observe the opposite from Figure 2, simulation results in Section 4.2 and the educational example in Section 5.1, where ranking at the cluster level is more effective. It is simply because that in this example, ICC is close to zero, much smaller than the typical

ICC values in educational studies (i.e, 0.15–0.25). As $ICC \to 0$, we have $RE^c \to 1$ from (9) in Proposition 4; and $RE^{id} \to RE_r^{RSS(H^{id})}$, the maximum value it can attain, from (10) in Proposition 5. As to the impact of the nonlinear ranking errors, $RE \downarrow$ as $\gamma \uparrow$ under each design. The RE curves for $H^{id} = 4$ are steeper than those for $H^{id} = 2$. As γ becomes sufficiently large, ranking is close to random so that all the RE curves approach 1. Note that even for $\gamma = 1$ (1mm is a quite distinguishable difference even when a cheap method is used for ranking), there are considerable gains in efficiency when ranking at the individual level or both levels.

Besides the observations made above, the RE patterns agree very well with our analytical results in Section 3.2. For example, our theory suggests that when H^c is fixed, RE does not depend on J, regardless of ranking schemes. This has been confirmed by the observation that in each upper panel, the black lines (for $J = 10$) overlap almost completely with the corresponding red lines (for $J = 20$) for designs with the same H^c and H^{id}. Also, the middle and right upper panels of Figure 4 show that for designs with the same H^c and K, larger H^{id} yields larger RE. This is, again, consistent with our theory. In addition, we compared the relative efficiency attained to what theory would have predicted under the designs with perfect ranking at both levels. Table 4 shows that they matched well.

The bottom panels of Figure 4 compare the performance of different designs in testing $H_0 : \Delta = 0$, where the black/red dotted straight line shows the power of the test based on the original CRDs (using SRS) for $J = 10/20$. For the same J, the power curves are always above the dotted lines, meaning that the RSS designs are more powerful than the corresponding SRS design in testing, and the gains from the designs with $H^{id} = 4$ are sizable. The patterns of the power are quite similar to those observed for RE, except for the following: (i) larger J leads to larger power so that we have two separate power curves for designs with the same H^c and H^{id}; and (ii) generally, the power curves decrease slower than the corresponding RE curves as γ increases.

To assess the size of the proposed test in this example, we deliberately simulated all the samples from the 179 men and randomly assigned the gender labels so that $H_0 : \Delta = 0$ holds

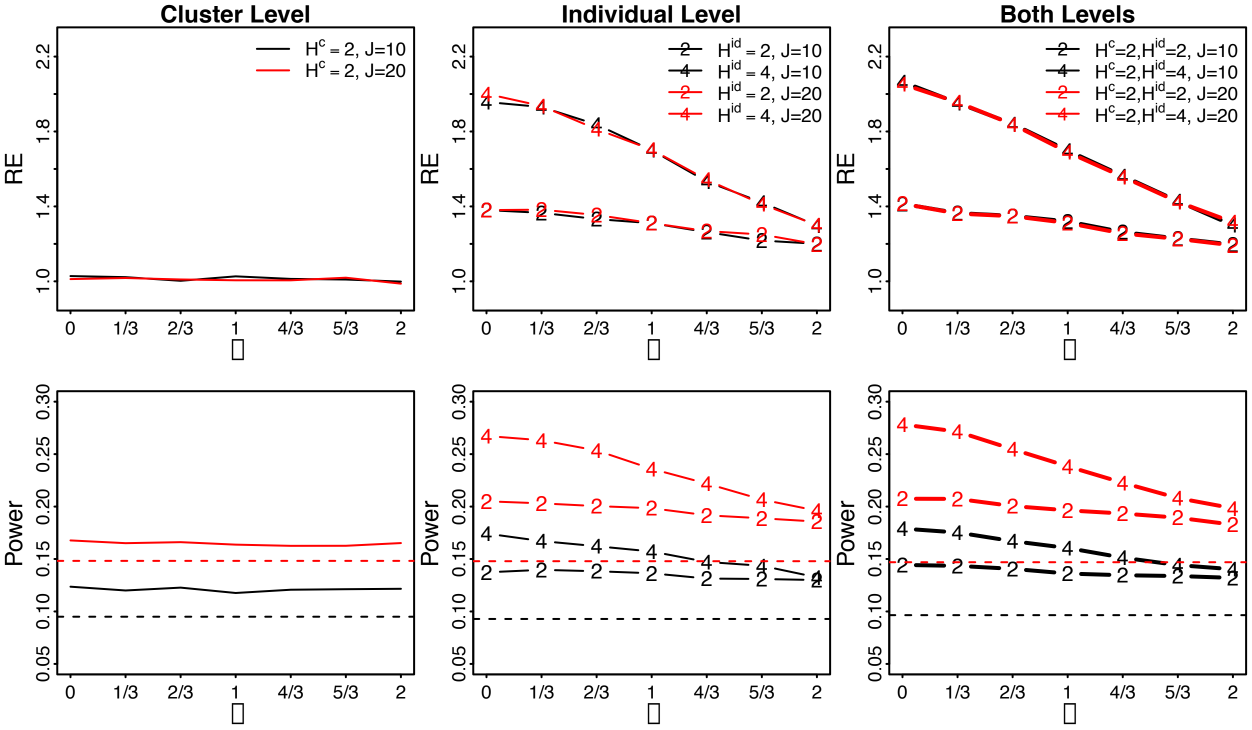


Figure 4: Dental example: empirical relative efficiency of $\hat{\Delta}_{RSS}$ versus $\hat{\Delta}_{SRS}$ (the top panels) and power of the tests based on the RSS and SRS designs (the bottom panels) as γ, a measure of the size of judgment error, increases. The left, middle and right panels are for ranking at the cluster level, individual level and both levels, respectively. The dotted straight lines in the bottom panels show the power of the test based on the original CRDs (using SRS). The lines in black are for $J = 10$ and lines in red are for $J = 20$.

				RE		TRE
Scheme	H^c	H^{id}	m^{id}	$J = 10$	$J = 20$	
i	2	1	4	1.03	1.03	1.04
ii	1	2	2	1.38	1.38	1.38
	1	4	1	1.99	1.97	1.99
iii	2	2	2	1.48	1.45	1.47
	2	4	1	2.11	2.20	2.17

Table 4: Dental example: comparing empirical relative efficiency (RE) with theoretical relative efficiency (TRE) under perfect ranking. The TRE values were calculated based on normal distributions.

for all the designs considered in Figure 4. We find that the F-test holds the size at 0.05 very well; and the size of the proposed test for the RSS designs is slightly biased (i.e., 0.05 or 0.06). This is exactly the same as what we observed before.

The results reported in Figure 4 were based on samples selected with replacement. We also simulated samples without replacement from the “population”. The RE and power values are larger than those for samples with replacement, due to the finite-population effect. However, all the findings from Figure 4 remain valid.

6 Discussion

We first mention that among the research that proposed using RSS with existing sampling designs, the contribution of our work is significantly different from that of the four papers mentioned in the introduction (i.e., Muttlak and McDonald 1992, Sud and Mishra 2006, Sroka 2008, and Nematollahi et al. 2008). In each case, the authors showed that an improvement would occur (although always with RSS at the final stage of sampling), and then produced an example or simulation to show how much. However, none of them produced theory that allowed them to quantify the improvement. In fact, the comment in Nematollahi et al. (2008) was typical: “Comparing relative precisions shows that increasing the number of cycles does not increase the relative precision in most cases. Although the effect of increasing n (# of primary sampling units) and m' (number ranked) on the relative precision (RP) is not clear, the results show that the maximum RP is obtained when n and m' are large.” In other words, these papers used empirical results to speculate on the effect of changes to the various ranked set sampling design parameters. Our work provides definite answers to this question, as well as other important questions like how the inferential procedures about testing the treatment effect should be constructed, what assumptions about the two-level model are required, what are the connections with the one-level RSS, and what are the role of ICC and the impact of imperfect ranking, for all possible combinations of RSS integration, leaving no doubt to practitioners. No previous work has addressed these questions systematically and comprehensively, especially under a model-based framework that imposes the minimum level of (distributional) assumptions.

We next point out two potential contributions of our work in other research settings. One is related to Ridout and Cobby (1987), where the authors investigated how much ranking errors and what they call non-random selection, which is really the effect of clustering within ranked sets, hurts RSS. The problem can be viewed as a special case of ranking at the individual level, and so our work provides a different angle on their work. In fact, our expressions (8), (14), (15) could be used to investigate this analytically, and explain observations they made from an example. The other is about the tradeoffs between stratification at different levels in a multistage design. In the sampling literature, there is not a well-known answer to address this. Since RSS can be viewed as a stratified sample (Stokes and Sager 1988), our work, which considers all possible combinations within two stages, may provide analytical insights on the tradeoffs.

Since this is the first study to investigate the use of RSS with CRDs to assess treatment effects under the HLM, our development focuses on the designs involving two-level data. This helps us avoid unnecessarily complex notations and technical details, and allows us to concentrate on developing basic approaches that can be adapted for higher-level data. In fact, no matter how many levels are in a sequential design, the essential idea for the RSS-integrated designs remains the same: use RSS in some stage(s) when feasible. Practical implementation of RSS requires that approximate ranking of outcomes be possible before measurement by some inexpensive means. As long as this requirement is met, RSS can potentially be used for selecting sampling units of any level. It would therefore be of interest to consider the extension to multi-site cluster randomized designs (MSCRDs), which have become very common in recent years (Spybrook 2008).

The proposed test relies on the asymptotic normality of $\hat{\Delta}_{RSS}$, which requires $J \to +\infty$. However, as long as J is not too small (e.g., $J \geq 10$), the gain in power is significantly larger than the small bias in test size when H_0 holds (test size 0.05–0.06 at the nominal level of 0.05), as suggested by our simulation. On the other hand, for very small J, the size of the proposed test tends to be above the nominal level. Thus, one interesting topic for future research is to develop an improved test for small J. Other potential topics include estimating variance components (σ_b^2, σ_r^2), testing the existence of the cluster effect (i.e., $\sigma_b^2 = 0$), and allowing the use of unbalanced RSS in CRDs.

References

Allor, J. H., Mathes, P. G., Roberts, J. K., Jones, F. G., and Champlin, T. M. (2010). Teaching students with moderate intellectual disabilities to read: An experimental examination of a comprehensive reading intervention. *Education and Training in Autism and Developmental Disabilities*, 45:3–22.

Balakrishnan, N. and Li, T. (2008). Ordered ranked set samples and application to inference. *Journal of Statistical Planning and Inference*, 138:3512–3524.

Bohn, L. L. and Wolfe, D. A. (1992). Nonparametric two-sample procedures for ranked-set samples data. *Journal of the American Statistical Association*, 87:552–561.

Bryk, A. S. and Raudenbush, S. W. (2002). *Hierarchical linear models: Applications and Data Analysis Methods*. Advanced qualitative techniques in the social sciences, 1. Sage Publications, Thousand Oaks, CA, US, 2nd edition.

Buschang, P. H., Demirjian, A., and Cadotte, L. (1988). Permanent mesiodistal tooth size of french-canadians. *J Can Dent Assoc*, 54(6):441–444.

Chen, M. and Lim, J. (2011). Estimating variances of strata in ranked set sampling. *Journal of Statistical Planning and Inference*, 141:2513–2518.

Chen, Z., Bai, Z., and Sinha, B. K. (2006). *Ranked Set Sampling: Theory and Applications*. Springer.

Dell, T. and Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28(2):545–555.

Fligner, M. A. and MacEachern, S. N. (2006). Nonparametric two-sample methods for ranked-set sample data. *Journal of the American Statistical Association*, 101:1107–1118.

Frey, J. (2007). Distribution-free statistical intervals via ranked-set sampling. *Canadian Journal of Statistics*, 35:585–596.

Frey, J. and Feeman, T. G. (2013). Variance estimation using judgment post-stratification. *Annals of the Institute of Statistical Mathematics*, 65:551–569.

Frey, J. and Ozturk, O. (2011). Constrained estimation using judgment post-stratification. *Annals of the Institute of Statistical Mathematics*, 63:769–789.

Frey, J., Ozturk, O., and Deshpande, J. V. (2007). Nonparametric tests for perfect judgment rankings. *Journal of the American Statistical Association*, 102:708–717.

Garn, S. M., Lewis, A. B., Swindler, D. R., and Kerewsky, R. S. (1967). Genetic control of sexual dimorphism in tooth size. *J Dent Res*, 46(5):963–972.

Ghosh, K. and Tiwari, R. C. (2007). Empirical process approach to some two-sample problems based on ranked set samples. *Annals of the Institute of Statistical Mathematics*, 59:757–787.

Hatefi, A., Jafari Jozani, M., and Ziou, D. (2013). Statistical inference for finite mixture models based on ranked set samples. *Statistica Sinica. Accepted.DOI:10.5705/ss.2012.178.*

Hayes, R. J. and Moulton, L. H. (2009). *Cluster Randomised Trials*. Chapman & Hall/CRC, Taylor & Francis Groups.

Hedges, L. V. and Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis*, 29:60–87.

Kvam, P. H. (2003). Ranked set sampling based on binary water quality data with covariates. *Journal of Agricultural, Biological and Environmental Science*, 8(3):271–279.

Kvam, P. H. and Samaniego, F. J. (1994). Nonparametric maximum likelihood estimation based on ranked set samples. *Journal of the American Statistical Association*, 89:526–537.

Lee, S.-J., Lee, S., Lim, J., Ahn, S.-J., and Kim, T.-W. (2006). Cluster analysis of human tooth size in subjects with normal occlusion. *American Journal of Orthodontics & Dentofacial Orthopedics*, To appear.

MacEachern, S. N., Ozturk, O., Wolfe, D. A., and Stark, G. (2002). A new ranked set sample estimator of variance. *Journal of the Royal Statistical Society, Series B*, 64:177–188.

McIntyre, G. A. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*, 3(4):385 – 390.

Mitsea, A., Moraitis, A., Leon, G., Nicopoulou-Karayianni, K., and Spiliopoulou, C. (2014). Sex determination by tooth size in a sample of greek population. *Journal of Comparative Human Biology*, 65:322–329.

Mode, N. A., Conquest, L. L., and Marker, D. A. (1999). Ranked set sampling for ecological research: Accounting for the total costs of sampling. *Environmetrics*, 10:179–194.

Murray, R., Ridout, M., and Cross, J. (2000). The use of ranked set sampling in spray deposit assessment. *Aspect of Applied Biology*, 57:141–146.

Muttlak, H. A. and McDonald, L. L. (1992). Ranked set sampling and the line intercept method: a more efficient procedure. *Biometrical Journal*, 34:329–346.

Nematollahi, N., Salehi, M. M., and Aliakbari Saba, R. (2008). Two-stage cluster sampling with ranked set sampling in the secondary sampling frame. *Communications in Statistics -Theory and Methods*, 37:2402–2415.

Nussbaum, B. D. and Sinha, B. K. (1997). Cost effective gasoline sampling using ranked set sampling. In *American Statistical Association 1997 Proceedings of the Section on Statistics and the Environment*, pages 83–87. The American Statistical Association.

Othman, S. and Harradine, N. (2007). Tooth size discrepancies in an orthodontic population. *The Angle Orthodontist*, 77:668–674.

Ozturk, O. (2002). Rank regression in ranked-set samples. *Journal of the American Statistical Association*, 97:1180–1191.

Ozturk, O. (2008). Statistical inference in the presence of ranking error in ranked set sampling. *Canadian Journal of Statistics*, 36:577–594.

Ozturk, O. (2012). Combining ranking information in judgment post stratified and ranked set sampling designs. *Environmental and Ecological Statistics*, 19:73–93.

Ozturk, O. (2013). Combining multi-observer information in partially rank-ordered judgment post-stratified and ranked set samples. *Canadian Journal of Statistics*, 41:304–324.

Ozturk, O. and Balakrishnan, N. (2009). An exact control-versus-treatment comparison test based on ranked set samples. *Biometrics*, 65:1213–1222.

Ridout, M. S. and Cobby, J. M. (1987). Ranked set sampling with non-random selection of sets and errors in ranking. *Applied Statistics*, 36:145–152.

Spybrook, J. (2008). Are power analyses reported with adequate detail: Findings from the first wave of group randomized trials funded by the Institute of Education Sciences. *Journal of Research on Educational Effectiveness*, 1:215–235.

Sroka, C. J. (2008). *Extending Ranked Set Sampling to Survey Methodology*. PhD thesis, Department of Statistics, Ohio State University.

Stokes, S. L. and Sager, T. W. (1988). Characterization of a ranked-set sample with application to estimating distribution functions. *Journal of the American Statistical Association*, 83:374–381.

Stovall, H. (2012). *Ranked Set Sampling and Its Applications in Educational Statistics*. Ph.d. thesis, Southern Methodist Unviersity.

Sud, V. and Mishra, D. C. (2006). Estimation of finite population mean using ranked set two stage sampling designs. *Journal of the Indian Society of Agricultural Statistics*, 60:108–117.

Takahashi, K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of Institute of Statistical Mathematics*, 20:1–31.

Wang, X., Stokes, L., Lim, J., and Chen, M. (2006). Concomitants of multivariate order statistics with application to judgment poststratification. *Journal of the American Statistical Association*, 101(476):1693–1704.

Wang, X., Wang, K., and Lim, J. (2012). Isotonized CDF estimation from judgment poststratification data with empty strata. *Biometrics*, 68:194–202.

Wolfe, D. A. (2004). Ranked set sampling: An approach to more efficient data collection. *Statistical Science*, 19:636–643.

Wolfe, D. A. (2012). Ranked set sampling: Its relevance and impact on statistical inference. *ISRN Probability and Statistics*, 2012:Article ID 568385, 32 pages.

Supplementary Material

Using Ranked Set Sampling with Cluster Randomized Designs for Improved Inference on Treatment Effects

Throughout this material, all the formulas cross-referenced are in the main body of the article.

S1 Proof of Theorem 1

Without loss of generality, we derive the weighted least square estimator under the full ranking scheme (i.e., ranking at both cluster and individual levels). Noting that $a_1 \equiv \Delta$ and $a_0 \equiv 0$, we can write the minimization problem in Theorem 1 as $\min_{\mu,a_1} Q(\mu, a_1)$, where

$$Q(\mu, a_1) = \sum_{i=0}^{1} \sum_{h=1}^{H_i} \sum_{j \in \mathcal{J}_i(h)} \frac{1}{K_{j(i)}} \sum_{h'=1}^{H_{j(i)}} \sum_{k \in \mathcal{K}_{j(i)}(h')} \left[Y_{k(ij)} - (\mu + a_i + \mu_{b.ih} + \mu_{r.ijh'}) \right]^2 ;$$

and $\mathcal{J}_i(h)$, $\mathcal{K}_{j(i)}(h')$, $\mu_{b.ih}$ and $\mu_{r.ijh'}$ are defined in Section 3.1.

To solve the above problem, we first set $\partial Q/\partial \mu = 0$ and $\partial Q/\partial a_1 = 0$, and obtain the following

$$\begin{cases} \sum_{i=0}^{1} \sum_{h=1}^{H_i} \sum_{j \in \mathcal{J}_i(h)} \frac{1}{K_{j(i)}} \sum_{h'=1}^{H_{j(i)}} \sum_{k \in \mathcal{K}_{j(i)}(h')} \left[Y_{k(ij)} - (\mu + a_i + \mu_{b.ih} + \mu_{r.ijh'}) \right] & = 0 \\ \sum_{h=1}^{H_1} \sum_{j \in \mathcal{J}_1(h)} \frac{1}{K_{j(1)}} \sum_{h'=1}^{H_{j(1)}} \sum_{k \in \mathcal{K}_{j(1)}(h')} \left[Y_{k(1j)} - (\mu + a_1 + \mu_{b.ih} + \mu_{r.ijh'}) \right] & = 0 \end{cases} .$$

Based on the consistency of ranking, $\sum_{h=1}^{H_{(i)}} \mu_{b.ih}/H_{(i)} = \mu_b = 0$ and $\sum_{h=1}^{H_{j(i)}} \mu_{r.ijh}/H_{j(i)} = \mu_r = 0$ hold. Then after some algebra, we have the following linear equations

$$\begin{cases} \sum_{i=0}^{1} \sum_{j=1}^{J_i} \bar{Y}_{j(i)} - (J_0 + J_1)\,\mu - J_1 a_1 & = 0 \\ \sum_{j=1}^{J_0} \bar{Y}_{j(0)} - J_1 \mu - J_a a_1 & = 0 \end{cases},$$

whose solution is unique, given by

$$\begin{cases} \hat{\mu} & = \frac{1}{J_0} \sum_{j=1}^{J_0} \bar{Y}_{j(0)} \\ \hat{a}_1 & = \frac{1}{J_1} \sum_{j=1}^{J_1} \bar{Y}_{j(1)} - \frac{1}{J_0} \sum_{j=1}^{J_0} \bar{Y}_{j(0)} \end{cases}.$$

It is easy to verify that $\nabla^2 Q$ is positive definite so that $(\hat{\mu}, \hat{a}_1)$ minimizes the objective function Q. It is also straightforward to show that under completely balanced RSS-structured CRDs, as described in Section 3.2, $(\hat{\mu}, \hat{a}_1)$ is the ordinary least square estimator of (μ, a_1), which minimizes the sum of unweighted squared distances, namely

$$\min_{\mu, a_1} \sum_{i=0}^{1} \sum_{h=1}^{H_i} \sum_{j \in \mathcal{J}_i(h)} \sum_{h'=1}^{H_{j(i)}} \sum_{k \in \mathcal{K}_{j(i)}(h')} \left[Y_{k(ij)} - E\left(Y_{k(ij)} \mid O_{j(i)} = h, O_{k(ij)} = h' \right) \right]^2.$$

S2 Proof of Theorem 2

We first derive the mean and variance of $\hat{\Delta}_{RSS}$ for the full ranking scheme with the design parameters $\{H_i^c, m_i^c, H_{j(i)}^{id}, m_{j(i)}^{id}\}$ and data $\mathbf{D}^b = \{Y_{k(ij)}, O_{j(i)}^c, O_{k(ij)}^{id}\}$.

Let $\hat{\mu}_{ih} = \sum_{j \in \mathcal{J}_i(h)} \bar{Y}_{j(i)}/m_i$ so that we can reorganize the expression of $\hat{\mu}_i$,

$$\hat{\mu}_i = \frac{1}{J_i} \sum_{j=1}^{J_i} \bar{Y}_{j(i)} = \frac{1}{H_i m_i} \sum_{h=1}^{H_i} \sum_{j \in \mathcal{J}_i(h)} \bar{Y}_{j(i)} = \frac{1}{H_i} \sum_{h=1}^{H_i} \hat{\mu}_{ih}.$$

Unbiasedness: From (1), we note that

$$\begin{aligned}\bar{Y}_{j(i)} &= \frac{1}{H_{j(i)}}\sum_{h'=1}^{H_{j(i)}}\frac{\sum_{k\in\mathcal{K}_{j(i)}(h')}Y_{k(ij)}}{m_{j(i)}}\\ &= \mu + a_i + b_{j(i)} + \frac{1}{H_{j(i)}}\sum_{h'=1}^{H_{j(i)}}\frac{\sum_{k\in\mathcal{K}_{j(i)}(h')}r_{k(ij)}}{m_{j(i)}}.\end{aligned}$$

Then

$$\begin{aligned}E(\hat{\mu}_{ih}) &= E\left[\frac{\sum_{j\in\mathcal{J}_i(h)}\bar{Y}_{j(i)}}{m_i}\right] = \frac{1}{m_i}\sum_{j\in\mathcal{J}_i(h)}E\left[\bar{Y}_{j(i)}\mid O_{j(i)}=h\right]\\ &= \frac{1}{m_i}\sum_{j\in\mathcal{J}_i(h)}\left\{\mu + a_i + E\left[b_{j(i)}\mid O_{j(i)}=h\right] + \frac{1}{H_{j(i)}}\sum_{h'=1}^{H_{j(i)}}\frac{\sum_{k\in\mathcal{K}_{j(i)}(h')}E\left[r_{k(ij)}\mid O_{k(ij)}=h'\right]}{m_{j(i)}}\right\}\\ &= \frac{1}{m_i}\sum_{j\in\mathcal{J}_i(h)}\left\{\mu + a_i + \mu_{b.ih} + \frac{1}{H_{j(i)}}\sum_{h'=1}^{H_{j(i)}}\frac{\sum_{k\in\mathcal{K}_{j(i)}(h')}\mu_{r.ijh'}}{m_{j(i)}}\right\}\\ &= \mu + a_i + \mu_{b.ih} + \frac{1}{m_i}\sum_{j\in\mathcal{J}_i(h)}\frac{1}{H_{j(i)}}\sum_{h'=1}^{H_{j(i)}}\mu_{r.ijh'} = \mu + a_i + \mu_{b.ih} + \frac{1}{J_{ih}}\sum_{j\in\mathcal{J}_i(h)}\mu_r\\ &= \mu + a_i + \mu_{b.ih}.\end{aligned}$$

So

$$\begin{aligned}E\left(\hat{\Delta}^b_{RSS}\right) &= \frac{\sum_{h=1}^{H_1}E\left(\hat{\mu}_{1h}\right)}{H_1} - \frac{\sum_{h=1}^{H_0}E\left(\hat{\mu}_{0h}\right)}{H_0},\\ &= \frac{\sum_{h=1}^{H_1}\left(\mu + a_1 + \mu_{b.1h}\right)}{H_1} - \frac{\sum_{h=1}^{H_0}\left(\mu + a_0 + \mu_{b.0h}\right)}{H_0}\\ &= = \Delta + \frac{\sum_{h=1}^{H_1}\mu_{b.1h}}{H_1} - \frac{\sum_{h=1}^{H_1}\mu_{b.0h}}{H_0} = \Delta + \mu_b - \mu_b = \Delta.\end{aligned}$$

It is obvious that the unbiasedness also holds for $\hat{\Delta}^c_{RSS}$ and $\hat{\Delta}^{id}_{RSS}$, since the first two ranking schemes are just reduced cases of the third.

Variance:

$$
\begin{aligned}
Var(\hat{\mu}_{ih}) &= Var\left[\frac{\sum_{j\in\mathcal{J}_i(h)} \bar{Y}_{j(i)}}{m_i}\right] = \frac{1}{m_i^2}\sum_{j\in\mathcal{J}_i(h)} Var\left[\bar{Y}_{j(i)} \mid O_{j(i)} = h\right] \\
&= \frac{1}{m_i^2}\sum_{j\in\mathcal{J}_i(h)} Var\left[\mu + a_i + b_{j(i)} + \frac{1}{H_{j(i)}}\sum_{h'=1}^{H_{j(i)}} \frac{\sum_{k\in\mathcal{K}_{j(i)}(h')} r_{k(ij)}}{m_{j(i)}} \mid O_{j(i)} = h\right] \\
&= \frac{1}{m_i^2}\sum_{j\in\mathcal{J}_i(h)} \left\{Var\left[b_{j(i)} \mid O_{j(i)} = h\right] + \frac{1}{H_{j(i)}^2} Var\left[\sum_{h'=1}^{H_{j(i)}} \frac{\sum_{k\in\mathcal{K}_{j(i)}(h')} r_{k(ij)}}{m_{j(i)}}\right]\right\} \\
&= \frac{1}{m_i^2}\sum_{j\in\mathcal{J}_i(h)} \left\{\sigma_{b.ih}^2 + \frac{1}{H_{j(i)}^2}\sum_{h'=1}^{H_{j(i)}} \frac{\sum_{k\in\mathcal{K}_{j(i)}(h')} Var\left[r_{k(ij)} \mid O_{k(ij)} = h'\right]}{m_{j(i)}^2}\right\} \\
&= \frac{1}{m_i}\sigma_{b.ih}^2 + \frac{1}{m_i^2}\sum_{j\in\mathcal{J}_i(h)} \frac{1}{m_{j(i)}H_{j(i)}^2}\sum_{h'=1}^{H_{j(i)}} \sigma_{r.ijh'}^2.
\end{aligned}
$$

Then

$$
\begin{aligned}
Var(\hat{\Delta}_{RSS}^b) &= \sum_{i=0}^{1} Var\left(\frac{\sum_{h=1}^{H_i} \hat{\mu}_{ih}}{H_i}\right) = \sum_{i=0}^{1} \frac{1}{H_i^2}\sum_{h=1}^{H_i} Var\left(\hat{\mu}_{ih}\right), \\
&= \sum_{i=0}^{1} \left[\frac{1}{m_i H_i^2}\sum_{h=1}^{H_i} \sigma_{b.ih}^2 + \frac{1}{m_i^2 H_i^2}\sum_{j=1}^{J_i} \frac{1}{m_{j(i)}H_{j(i)}^2}\sum_{h'=1}^{H_{j(i)}} \sigma_{r.ijh'}^2\right],
\end{aligned}
$$

which gives the formula in (5) after inserting $J_i = m_i H_i$ and $K_{j(i)} = m_{j(i)}H_{j(i)}$.

Finally, we set $H_{j(i)} = 1$, $m_{j(i)} = K_{j(i)}$ and $\sum_{h'=1}^{H_{j(i)}} \sigma_{r.ijh'}^2 = \sigma_r^2$ in (5) to get (3), and set $H_i = 1$, $m_i = J_i$ and $\sum_{h=1}^{H_i} \sigma_{b.ih}^2 = \sigma_b^2$ in (5) to get (4).

S3 Proof of Theorem 3

It is easy to verify the variance formula (6) for $\hat{\Delta}_{SRS}$. Let $\mu_{b.ih}$ be the mean of the hth judgment order statistic (relative to H_i ordered observations in total) of the cluster effect b. Let $\mu_{r.ijh}$ be the mean of the hth judgment order statistic (relative to $H_{j(i)}$ ordered observations in total) of the individual effect r. From the equalities $\sum_{h=1}^{H_i} \sigma_{b.ih}^2 = H_i\sigma_b^2 - \sum_{h=1}^{H_i} \mu_{b.ih}^2$ and

$\sum_{h'=1}^{H_{j(i)}} \sigma^2_{r.ijh'} = H_{j(i)}\sigma^2_r - \sum_{h=1}^{H_i} \mu^2_{r.ijh}$ (Dell and Clutter 1972), we have

$$\frac{1}{H_i}\sum_{h=1}^{H_i} \sigma^2_{b.ih} \leq \sigma^2_b, \quad \frac{1}{H_{j(i)}}\sum_{h'=1}^{H_{j(i)}} \sigma^2_{r.ijh'} \leq \sigma^2_r$$

for all i, j, which leads to $Var\left(\hat{\Delta}_{RSS}\right) \leq Var\left(\hat{\Delta}_{SRS}\right)$ under all three ranking schemes.

S4 Proof of Propositions

Proposition 1

Results (i) and (iv) follow directly from (7). Based on Theorem 2 in Takahashi and Wakimoto (1968), as $H^c \uparrow$, $\sum_{h=1}^{H^c} \sigma^2_{b.h}/H^c \downarrow$ (decreases) under perfect ranking and so $RE^c \uparrow$. Using the relationship in (13), (ii) holds under the linear ranking error model, too. To establish (iii), it is sufficient to rewrite RE^c in (7) as

$$RE^c = 1 + \left(\sigma^2_b - \frac{\sum_{h=1}^{H^c} \sigma^2_{b.h}}{H^c}\right) / \left(\frac{\sum_{h=1}^{H^c} \sigma^2_{b.h}}{H^c} + \frac{\sigma^2_r}{K}\right).$$

Proposition 2

The proof is similar to that of Proposition 1, and so omitted for brevity.

Proposition 3

Result (i) follows directly from the expression

$$RE^b = \left(\sigma^2_b + \frac{\sigma^2_r}{K}\right) / \left(\frac{\sum_{h=1}^{H^c} \sigma^2_{b.h}}{H^c} + \frac{\sum_{h=1}^{H^{id}} \sigma^2_{r.h}}{H^{id}K}\right).$$

For result (ii), the proof is similar to that for (ii) in Proposition 1. To prove (iii), we re-express RE^b by

$$RE^b = \left(K + \frac{1}{ICC} - 1\right) / \left[\frac{K}{RE_b^{RSS(H^c)}} + \frac{1}{RE_r^{RSS(H^{id})}} \cdot \left(\frac{1}{ICC} - 1\right)\right],$$

where $RE_r^{RSS(H^{id})}$ and $RE_b^{RSS(H^c)}$ are defined in Section 3.3. Now let's treat K as a continuous variable so that

$$\frac{d}{dK} RE^b = \left(\frac{1}{RE_r^{RSS(H^{id})}} - \frac{1}{RE_b^{RSS(H^c)}}\right)\left(\frac{1}{ICC} - 1\right) / \left[\frac{K}{RE_b^{RSS(H^c)}} + \frac{1}{RE_r^{RSS(H^{id})}} \cdot \left(\frac{1}{ICC} - 1\right)\right].$$

Clearly, when $RE_r^{RSS(H^{id})} < RE_b^{RSS(H^c)}$, $dRE^b/dK > 0$ so that RE^b is an increasing function of K; when $RE_r^{RSS(H^{id})} > RE_b^{RSS(H^c)}$, $dRE^b/dK < 0$ so that RE^b is a decreasing function of K; and when $RE_r^{RSS(H^{id})} = RE_b^{RSS(H^c)} \equiv C_0$, then $RE^b \equiv C_0$. The above results hold no matter whether K is continuous or discrete.

Finally, result (iv) simply follows from the fact that $RE_r^{RSS(H^{id})} \geq 1$ and $RE_b^{RSS(H^c)} \geq 1$.

Proposition 4

To establish (9) in Proposition 4, we simply note that in (7), $RE_b^{RSS(H^c)} = H^c\sigma_b^2 / \sum_{h=1}^{H^c} \sigma_{b.h}^2$ and $\sigma_r^2/\sigma_b^2 = (1 - ICC)/ICC$. It is trivial to show RE^c is an increasing function of $RE_b^{RSS(H^c)}$ using (9). Re-organizing (9) into

$$RE^c = 1 + K\left(1 - \frac{1}{RE_b^{RSS(H^c)}}\right) / \left(\frac{K}{RE_b^{RSS(H^c)}} + \frac{1}{ICC} - 1\right)$$

reveals that RE^c is an increasing function of ICC since $RE_b^{RSS(H^c)} \geq 1$. The maximum of RE^c follows from the fact that $RE_b^{RSS(H^c)} \leq (H+1)/2$ (Takahashi and Wakimoto, 1968).

Proposition 5

The proof is similar to that of Proposition 4, and so omitted for brevity.

S5 Proof of Theorem 4

We first introduce a lemma on which the asymptotic theorem is built.

Lemma 1. *Let $X_1, \dots, X_H$ be i.i.d continuous random variables with mean 0 and $\mathrm{E}\left|X_1\right|^{2+\delta} < \infty$ for some $\delta > 0$. Then, for every $h = 1, \dots, H$, $\mathrm{E}|X_{[h]}|^{2+\delta} \leq H \cdot \mathrm{E}\left|X_1\right|^{2+\delta} < \infty$.*

Proof. Let $f(x)$ denote the pdf of X, with the support $\mathcal{X}$. For $h \in \{1, \dots, H\}$, let $f_{[h]}(x)$ be the pdf of the hth judgment order statistic (relative to H observations) of X. Under the assumption of a consistent ranking mechanism (Chap. 2, pg. 12 in Chen et al. 2006), we know

$$f(x) = \frac{f_{[1]}(x) + \cdots + f_{[H]}(x)}{H}.$$

Thus, for any $h \in \{1, \dots, H\}$, $f_{[h]}(x) \leq H \cdot f(x)$. Then, we have

$$\begin{aligned}
\mathrm{E}\left|X_{[h]}\right|^{2+\delta} &= \int_{\mathcal{X}} |x|^{2+\delta} f_{[h]}(x)dx \\
&\leq H \cdot \int_{\mathcal{X}} |x|^{2+\delta} f\left(x\right) dx \\
&= H \cdot \mathrm{E}\left|X_1\right|^{2+\delta} < \infty.
\end{aligned}$$

□

We present the proofs to Theorem 4 for the first two ranking schemes. The proof for the third is basically a combination of the first two but much involved in notations and hence omitted for brevity.

Ranking at the cluster level

Let $W_{j(i)} = \bar{Y}_{j(i)} - \mu_{ih}$ and $U_{j(i)} = \frac{1}{\sqrt{m_i}} W_{j(i)}$ for $j \in \mathcal{J}_i(h)$ and $i = 0, 1$, where $\mu_{ih} = \mu + a_i + \mu_{b.ih}$. Then $E\left(W_{j(i)}\right) = 0$ and $Var\left(W_{j(i)}\right) = \sigma^2_{b.ih} + \frac{\sigma^2_r}{K_{j(i)}}$. We will apply the Lindberg-Feller theorem to show the asymptotic normality of $\sum_{j \in \mathcal{J}_i(h)} U_{j(i)}$. To do so, we need to check the two conditions of the theorem.

First, note that for $j \in \mathcal{J}_i(h)$ and $i = 0, 1$, $E\left(U_{j(i)}\right) = 0$ and as $m_i \to +\infty$,

$$\sum_{j \in \mathcal{J}_i(h)} E\left(U^2_{j(i)}\right) = \sigma^2_{b.ih} + \frac{1}{m_i} \sum_{j \in \mathcal{J}_i(h)} \frac{\sigma^2_r}{K_{j(i)}} \to \sigma^2_{b.ih} + w_{ih}\sigma^2_r > 0$$

where $w_{ih} \equiv \lim_{m_i \to +\infty} \frac{1}{m_i} \sum_{j \in \mathcal{J}_i(h)} \frac{1}{K_{j(i)}}$. Note that each w_{ih} exists since $w_{ih} > 0$, $w_i = \frac{1}{H_i} \sum_{h=1}^{H_i} w_{ih}$ and w_i is assumed to exist. Second, for all $\epsilon > 0$,

$$\begin{aligned}
\sum_{j \in \mathcal{J}_i(h)} E\left(U^2_{j(i)}; \left|U_{j(i)}\right| > \epsilon\right) &= \sum_{j \in \mathcal{J}_i(h)} E\left(\frac{1}{m_i} W^2_{j(i)}; \left|W_{j(i)}\right| > \sqrt{m_i}\epsilon\right) \\
&\leq \frac{1}{\epsilon^\delta m_i^{1+\delta/2}} \sum_{j \in \mathcal{J}_i(h)} E\left(\left|W_{j(i)}\right|^{2+\delta}\right)
\end{aligned}$$

and based on Jensen's inequality and Lemma 1,

$$\begin{aligned}
E\left(\left|W_{j(i)}\right|^{2+\delta}\right) &= E\left(\left|\bar{Y}_{j(i)} - \mu_{ih}\right|^{2+\delta}\right) \\
&= E\left(\left|\frac{1}{K_{j(i)}} \sum_{k=1}^{K_{j(i)}} \left(Y_{k(ij)} - \mu_{ih}\right)\right|^{2+\delta}\right) \leq \frac{1}{K^{2+\delta}_{j(i)}} \sum_{k=1}^{K_{j(i)}} E\left(\left|Y_{k(ij)} - \mu_{ih}\right|^{2+\delta}\right) \\
&\leq \frac{1}{K^{2+\delta}_{j(i)}} \sum_{k=1}^{K_{j(i)}} \left\{E\left(\left|b_{j(i)} - \mu_{b.ih}\right|^{2+\delta}\right) + E\left(\left|r_{k(ij)}\right|^{2+\delta}\right)\right\} < +\infty.
\end{aligned}$$

These lead to $\sum_{j \in \mathcal{J}_i(h)} E\left(U^2_{j(i)}; \left|U_{j(i)}\right| > \epsilon\right) \to 0$, as as $m_i \to +\infty$.

Now based on Lindberg-Feller (L-F) theorem (see Theorem 3.4.5, pg. 129, Durrett 2010), we have

$$\sum_{j \in \mathcal{J}_i(h)} U_{j(i)} \to N\left(0, \sigma^2_{b.ih} + w_{ih}\sigma^2_r\right) \quad \text{as } m_i \to +\infty.$$

Further noting that $\sqrt{m_i}(\hat{\mu}_{ih} - \mu_{ih}) = \sum_{j \in \mathcal{J}_i(h)} U_{j(i)}$ and $\lim_{m_i \to +\infty} J_i / J. = \pi_i$, we have

$$\sqrt{\frac{J.}{J_i} H_i} \cdot \frac{\sum_{h=1}^{H_i} \sqrt{m_i}\left(\hat{\mu}_{ih} - \mu_{ih}\right)}{H_i} \to N\left(0, \frac{1}{\pi_i H_i} \sum_{h=1}^{H_i} \left(\sigma_{b.ih}^2 + w_{ih}\sigma_r^2\right)\right),$$

which, combined with

$$\sqrt{J.}\left(\hat{\Delta}_{RSS} - \Delta\right) = \sqrt{J.}\left[\frac{\sum_{h=1}^{H_1}\left(\hat{\mu}_{1h} - \mu_{1h}\right)}{H_1} - \frac{\sum_{h=1}^{H_0}\left(\hat{\mu}_{0h} - \mu_{0h}\right)}{H_0}\right],$$

leads to the final result in (3).

Ranking at the individual level

Let $Q_{j(i)} = \bar{Y}_{j(i)} - \mu_i$ and $T_{j(i)} = \frac{1}{\sqrt{J_i}} Q_{j(i)}$ for $j = 1, \ldots J_i$ and $i = 0, 1$. Then $E\left(Q_{j(i)}\right) = 0$ and $Var\left(Q_{j(i)}\right) = Var\left(\bar{Y}_{j(i)}\right) = \sigma_b^2 + \frac{1}{H_{j(i)}^2 m_{j(i)}} \sum_{h'=1}^{H_{j(i)}} \sigma_{r.ijh'}^2$. We will apply the L-F theorem to show the asymptotic normality of $\sum_{j=1}^{J_i} T_{j(i)}$. To do so, we need to check the two conditions of the theorem.

First, note that for $j = 1, \ldots J_i$ and $i = 0, 1$, $E\left(T_{j(i)}\right) = 0$, and

$$\begin{aligned} \sum_{j=1}^{J_i} E\left(T_{j(i)}^2\right) &= \sigma_b^2 + \frac{1}{J_i} \sum_{h=1}^{H_i} \frac{1}{H_{j(i)}^2 m_{j(i)}} \sum_{h'=1}^{H_{j(i)}} \sigma_{r.ijh'}^2 \\ &\to \sigma_b^2 + \tilde{\sigma}_{r.i}^2 > 0 \quad \text{as } J_i \to +\infty. \end{aligned}$$

Second, for all $\epsilon > 0$,

$$\sum_{j=1}^{J_i} E\left(T_{j(i)}^2; \left|T_{j(i)}\right| > \epsilon\right) = \sum_{j=1}^{J_i} E\left(\frac{1}{J_i} Q_{j(i)}^2; \left|Q_{j(i)}\right| > \sqrt{J_i}\epsilon\right) \leq \frac{1}{\epsilon^\delta J_i^{1+\delta/2}} \sum_{j=1}^{J_i} E\left(\left|Q_{j(i)}\right|^{2+\delta}\right)$$

and based on Jensen's inequality and Lemma 1,

$$
\begin{aligned}
E\left(\left|Q_{j(i)}\right|^{2+\delta}\right) = E\left(\left|\frac{1}{H_{j(i)}}\sum_{h'=1}^{H_{j(i)}}\left(b_{j(i)}+\bar{r}_{ijh'}-\mu_{r.ijh'}\right)\right|^{2+\delta}\right) &\leq \frac{1}{H_{j(i)}^{2+\delta}}\sum_{h'=1}^{H_{j(i)}} E\left|b_{j(i)}+\bar{r}_{ijh'}-\mu_{r.ijh'}\right|^{2+\delta} \\
&\leq \frac{1}{H_{j(i)}^{2+\delta}}\sum_{h'=1}^{H_{j(i)}}\left(E\left(\left|b_{j(i)}\right|^{2+\delta}\right)+E\left(\left|\bar{r}_{ijh'}-\mu_{r.ijh'}\right|^{2+\delta}\right)\right) \\
&\leq \frac{1}{H_{j(i)}^{2+\delta}}\sum_{h'=1}^{H_{j(i)}}\left(E\left(\left|b_{j(i)}\right|^{2+\delta}\right)+\sum_{k\in\mathcal{K}_{j(i)}(h')} E\left(\left|r_{k(ij)}-\mu_{r.ijh'}\right|^{2+\delta}\right)\right) < \infty,
\end{aligned}
$$

where $\bar{r}_{ijh'} = \frac{1}{m_{j(i)}}\sum_{k\in\mathcal{K}_{j(i)}(h')} r_{k(ij)}$. These lead to $\sum_{j=1}^{J_i} E\left(T_{j(i)}^2; \left|T_{j(i)}\right| > \epsilon\right) \to 0$ as $J_i \to +\infty$. Now based on the L-F theorem, we have $\sum_{j=1}^{J_i} T_{j(i)} \to N\left(0, \sigma_b^2 + \tilde{\sigma}_{r.i}^2\right)$, as $J_i \to +\infty$. Further note that $\sqrt{J_i}(\hat{\mu}_i - \mu_i) = \sum_{j=1}^{J_i} T_{j(i)}$, $\lim_{m_i\to+\infty} J_i/J. = \pi_i$, which, combined with $\sqrt{J.}\left(\hat{\Delta}_{RSS} - \Delta\right) = \sqrt{J.}\left[(\hat{\mu}_1 - \mu_1) - (\hat{\mu}_0 - \mu_0)\right]$, leads to the final result in (4).

S6 Additional Simulation Results for Power Comparison

In Section 4.2, we present two tables to compare the tests based on the RSS and SRS designs, where the original critical values were used. Here, Tables S1 and S2 present results for power comparison, where empirical critical values were obtained for Z_{RSS} to control the type I error rate at 0.05.

		$\delta = 0.15$				$\delta = 0.25$				$\delta = 0.45$			
H	m	SRS	RSS			SRS	RSS			SRS	RSS		
	ρ		0.7	0.9	1		0.7	0.9	1		0.7	0.9	1
2	4	0.09	0.10	0.10	0.10	0.16	0.18	0.20	0.20	0.42	0.47	0.50	0.52
	6	0.11	0.12	0.13	0.13	0.23	0.26	0.27	0.30	0.59	0.65	0.69	0.72
	8	0.14	0.14	0.16	0.16	0.29	0.31	0.36	0.37	0.73	0.77	0.82	0.84
	10	0.16	0.18	0.19	0.20	0.37	0.40	0.45	0.46	0.81	0.86	0.91	0.92
4	4	0.13	0.16	0.19	0.23	0.30	0.35	0.44	0.51	0.71	0.82	0.90	0.95
	6	0.19	0.24	0.28	0.30	0.42	0.53	0.62	0.68	0.89	0.95	0.98	0.99
	8	0.24	0.29	0.34	0.39	0.52	0.64	0.74	0.80	0.96	0.99	1.00	1.00
	10	0.28	0.36	0.43	0.47	0.64	0.76	0.84	0.89	0.98	1.00	1.00	1.00
6	4	0.18	0.23	0.30	0.37	0.42	0.54	0.67	0.77	0.89	0.96	0.99	1.00
	6	0.25	0.35	0.44	0.54	0.58	0.74	0.86	0.92	0.97	1.00	1.00	1.00
	8	0.33	0.45	0.56	0.66	0.71	0.85	0.94	0.97	0.99	1.00	1.00	1.00
	10	0.40	0.52	0.65	0.76	0.80	0.91	0.97	0.99	1.00	1.00	1.00	1.00
8	4	0.24	0.32	0.43	0.53	0.54	0.70	0.84	0.92	0.96	0.99	1.00	1.00
	6	0.32	0.44	0.58	0.71	0.72	0.85	0.95	0.99	1.00	1.00	1.00	1.00
	8	0.41	0.56	0.72	0.83	0.83	0.94	0.99	1.00	1.00	1.00	1.00	1.00
	10	0.49	0.65	0.81	0.92	0.91	0.98	1.00	1.00	1.00	1.00	1.00	1.00

Table S1: Ranking at the cluster level: power comparison after matching the type I error rates for the tests based on the RSS and SRS designs under different levels of the effect size δ. Here, the test sizes based on RSS were controlled at the nominal level 0.05 by using simulated critical values.

		$\delta = 0.15$				$\delta = 0.25$				$\delta = 0.45$			
H	m	SRS	RSS			SRS	RSS			SRS	RSS		
	ρ		0.7	0.9	1		0.7	0.9	1		0.7	0.9	1
2	4	0.14	0.14	0.14	0.15	0.29	0.30	0.31	0.32	0.72	0.74	0.75	0.76
	6	0.15	0.15	0.15	0.16	0.33	0.33	0.34	0.34	0.78	0.79	0.79	0.80
	8	0.15	0.15	0.16	0.17	0.33	0.35	0.35	0.37	0.80	0.80	0.81	0.83
	10	0.16	0.16	0.16	0.18	0.35	0.36	0.36	0.39	0.80	0.81	0.82	0.84
4	4	0.15	0.16	0.16	0.16	0.34	0.35	0.37	0.39	0.79	0.82	0.83	0.84
	6	0.16	0.17	0.17	0.17	0.36	0.37	0.37	0.36	0.82	0.83	0.84	0.85
	8	0.16	0.17	0.17	0.18	0.38	0.38	0.39	0.39	0.83	0.85	0.85	0.86
	10	0.17	0.17	0.17	0.18	0.37	0.39	0.38	0.41	0.85	0.85	0.86	0.86
6	4	0.16	0.17	0.18	0.18	0.35	0.38	0.38	0.41	0.82	0.84	0.85	0.85
	6	0.16	0.16	0.17	0.17	0.37	0.37	0.38	0.39	0.83	0.85	0.85	0.86
	8	0.17	0.17	0.17	0.18	0.38	0.39	0.39	0.40	0.84	0.85	0.86	0.88
	10	0.17	0.18	0.18	0.18	0.39	0.40	0.41	0.41	0.85	0.86	0.87	0.86
8	4	0.16	0.17	0.18	0.18	0.37	0.39	0.40	0.40	0.83	0.85	0.86	0.86
	6	0.17	0.16	0.17	0.18	0.39	0.38	0.39	0.41	0.85	0.85	0.86	0.87
	8	0.17	0.18	0.18	0.18	0.39	0.40	0.40	0.40	0.85	0.86	0.87	0.87
	10	0.17	0.17	0.17	0.17	0.38	0.39	0.40	0.40	0.86	0.86	0.87	0.87

Table S2: Ranking at the individual level: power comparison after matching the type I error rates for the tests based on the RSS and SRS designs under different levels of the effect size δ. Here, the test sizes based on RSS were controlled at the nominal level 0.05 by using simulated critical values.

S7 HSLS09 data example

					δ=0.15		δ=0.25		δ=0.45		δ=0.8	
Design	H^c	m^c	H^{id}	m^{id}	SRS	RSS	SRS	RSS	SRS	RSS	SRS	RSS
i-1	3	5	1	6	0.10	0.12	0.19	0.24	0.49	0.61	0.93	0.98
i-2	5	3	1	6	0.10	0.13	0.19	0.26	0.50	0.66	0.93	0.99
ii-1	1	15	2	3	0.10	0.10	0.19	0.20	0.50	0.51	0.94	0.94
ii-2	1	15	3	2	0.10	0.10	0.19	0.21	0.49	0.54	0.93	0.95
ii-3	1	15	6	1	0.10	0.11	0.19	0.21	0.50	0.55	0.93	0.96
iii-1-1	3	5	2	3	0.10	0.12	0.19	0.26	0.49	0.65	0.93	0.99
iii-1-2	3	5	3	2	0.10	0.13	0.19	0.27	0.49	0.67	0.93	0.99
iii-1-3	3	5	6	1	0.10	0.13	0.19	0.28	0.50	0.70	0.93	0.99
iii-2-1	5	3	2	3	0.10	0.13	0.19	0.28	0.49	0.70	0.93	0.99
iii-2-2	5	3	3	2	0.10	0.13	0.19	0.29	0.49	0.72	0.93	0.99
iii-2-3	5	3	6	1	0.10	0.14	0.19	0.32	0.50	0.76	0.94	1.00

Table S3: HSLS09 example: power comparison using empirical critical values to match the type I error rates ($\alpha = 0.05$) for the tests based on the RSS and SRS designs. The power of the F-test (based on SRS) in each column should be constant, but subject to Monte Carlo errors.

References

Chen, Z., Bai, Z., and Sinha, B. K. (2006). *Ranked Set Sampling: Theory and Applications*. Springer.

Dell, T. and Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28(2):545–555.

Durrett, R. (2010). *Probability: Theory and Examples*. Cambridge University Press, New York, fourth edition.

Takahashi, K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of Institute of Statistical Mathematics*, 20:1–31.