NBER WORKING PAPER SERIES

AGGREGATING PARTIAL RANKINGS FROM NEIGHBORS: METHODOLOGY AND EMPIRICAL EVIDENCE

Pascaline Dupas
Marcel Fafchamps
Deivy Houeix

Working Paper 29911
http://www.nber.org/papers/w29911

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2022, Revised March 2026

We thank our editor Garance Genicot and three anonymous referees for their invaluable guidance. For excellent comments, we are also grateful to Christian Ghiglino, Francis Bloch, Gabrielle Demange, Marcin Dziubinski, Rakesh Vohra, Yann Bramoullé, Steven Glazerman, Andrew Crawford, Florian Grosset, Sneha Subramanian, Matthew Olckers, Josh Blumenstock, Gayatri Koolwal, Elliott Collins, Abdul Noury, Doulo Sow, Carly Trachtman, and the participants at the IPA 2021 Methods and Measurement Conference and at the NYU Abu Dhabi 2023 Conference on the Economics of Networks. We thank Francis Bloch for inspiring this research. We would like to acknowledge and thank Eva Lestant and Arsène Zongo for excellent field research assistance, and Innovations for Poverty Action (IPA) Côte d'Ivoire for collecting the Abidjan data. We are grateful to the Stanford King Center on Global Development and to the IPA Research Methods Initiative Grant IPA-M5 for funding this research. Human Subjects Research approval was obtained from the Stanford University IRB (Protocol 42884). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2022 by Pascaline Dupas, Marcel Fafchamps, and Deivy Houeix. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Aggregating Partial Rankings from Neighbors: Methodology and Empirical Evidence
Pascaline Dupas, Marcel Fafchamps, and Deivy Houeix
NBER Working Paper No. 29911
April 2022, Revised March 2026
JEL No. D31, O12

ABSTRACT

Many decisions require ordering alternatives: for example, the selection of top candidates for a competitive academic program or the selection of the poorest individuals for a cash transfer program. One common approach consists in aggregating orderings reported by different observers (e.g., committee or community members), but those orderings are typically partial: not all observers rank all applicants. We introduce a novel type of approach, based on pairwise rankings, to (i) aggregate partial orderings reported by multiple observers and (ii) construct confidence intervals for the resulting aggregate ordering. We identify, both theoretically and using simulations, the conditions under which a pairwise approach dominates rank averaging: when reporting error is low, reported orderings are partial, and observers rank alternatives that are close to each other in their true latent ordering. We introduce improvements to rank averaging and pairwise methods and illustrate them using several datasets. We find that, with partial reported orderings, Borda counts (i.e., simple rank averages) are dominated by the averaging of normalized ranks and should never be used in practice.

Pascaline Dupas
Princeton University
Department of Economics
and CEPR
and also NBER
pdupas@princeton.edu

Marcel Fafchamps
Stanford University
and NBER
fafchamp@stanford.edu

Deivy Houeix
Harvard University
deivy.houeix@gmail.com

1 Introduction

Numerous practical situations demand that we rank alternatives in order of priority. Yet too often individuals have information that allows them to produce only a *partial* ranking of alternatives. Examples include: farmers experimenting with new crops and techniques; workers observing co-workers; consumers trying new products; universities identifying the most qualified applicants for a selective program; or a government determining which individuals in a community face the greatest economic hardship and should receive financial assistance. In all these cases, individual economic agents have specific information that enables them to rank some—but typically not all—of the available alternatives according to a common latent ordering. One commonly used solution to this aggregation problem is to take the average of the ranks given by different observers. This approach, however, often fails to recover the common latent ordering whenever observers report partial (i.e., incomplete) orderings of the alternatives (e.g., Alvo and Yu, 2014; Marden, 1995). Mis-ranking arises due to size variation and imperfect overlap between sets of ranked alternatives, and it is particularly severe when some observers only report on alternatives that have a high rank in the common ordering, while others report only on lowly ranked ones. Averaging orderings normalized by the size of ranked sets does alleviate the problem somewhat—and should always be preferred to the Borda count/rank averaging method when observers rank different subsets of alternatives. But it does not eliminate mis-ranking.

We develop a novel methodology to overcome this problem. The methodology aggregates partial orderings reported by multiple observers by making use of the transitive closure of a graph. This method extends the theoretical work of Tangian (2000) on the use of complete pairwise rankings to recover aggregate orderings. In contrast with the commonly used Borda count method that averages reported ranks, our method first averages *pairwise* rankings and then computes their transitive closure by using graph theory to 'stitch' them together (Lidl and Pilz, 1997) and recover the aggregate ordering.[1] We show that this method outperforms the Borda count method when observers only rank some of the available alternatives (e.g., some of their neighbors) and particularly when ranked alternatives tend to be close in the common latent ordering (e.g., observers mostly compare similarly poor households or similarly rich households). We also develop a simple procedure to construct robust confidence intervals for estimated ranks from both partial and complete orderings reported by multiple observers. This method can be applied to estimated ranks obtained through rank averaging or through our pairwise method.

[1] Truthful reporting is necessary for our method to recover the common latent ordering, but this is also true of other approaches.

We apply the method to data collected to rank households by poverty level. Many developmental interventions aim to identify the poor. In some instances, such as when poverty is highly concentrated, community-based or spatial targeting may be sufficient (Diamond et al., 2016; Premand and Barry, 2022). However, identifying the poor usually requires within-community targeting mechanisms. In the absence of universal administrative records (e.g., income tax filings), programs often rely on self-reported information collected through surveys. Those can however be costly, time-consuming, and/or subject to manipulation.[2] A common compromise is to elicit relative rankings from community members. This approach, often cheaper than surveys and more transparent than relying on local elites, has been shown to be informative in rural contexts (Alatas et al., 2016, 2012; Trachtman et al., 2026) and to identify high-potential entrepreneurs (Hussam et al., 2022). But a key challenge is how much information people have or are willing to share about each other. Partial orderings arise when respondents are unable or unwilling to rank certain individuals.

To showcase the pitfalls associated with partial orderings and the value of our proposed pairwise aggregation method, we rely on the data gathered by Alatas et al. (2012) in 640 rural communities of Indonesia (case study 1) and original data we collected in 34 poor neighborhoods of urban Côte d'Ivoire (case study 2). The measurement of relative material welfare is a topical policy issue in both contexts. For example, in 2019, the Ivorian government started rolling out universal health care coverage (CMU), which targets the poorest using a combination of observables and community assessments. Our data show substantial within-neighborhood heterogeneity in welfare, making purely geographic targeting ineffective and motivating the question of whether peer-based orderings can improve targeting.

In both contexts, we compare our pairwise rank approach to the simple Borda count method used by Alatas et al. (2012), and to an improved rank averaging method that addresses heterogeneity in the number of alternatives ranked across observers. We assess performance by comparing aggregated orderings to welfare orderings constructed from household survey data. In both datasets/case studies, target households complete a survey

[2]More broadly, survey-based orderings can be distorted by measurement error and strategic misreporting (Banerjee et al., 2020), leading to mis-assignment (Cruces et al., 2013). Measurement error arises when respondents have imperfect knowledge, e.g., recall problems; and subjective well-being often correlates only weakly with material living standards across contexts (Blanchflower and Oswald, 2004; Fafchamps and Shilpi, 2008; Layard, 2009). Response bias occurs when respondents have incentives to misreport, e.g., if lower reported welfare increases eligibility. Even when such incentives do not change *relative* ranks, they can still produce misclassification of individuals as poor or non-poor (Ravallion, 2008). An alternative is to delegate targeting decisions to local actors—e.g., Malawi chiefs identify beneficiaries for an input subsidy (Basurto et al., 2020)—but this raises concerns about capture or favoritism (Alatas et al., 2019).

covering consumption, assets, and household characteristics. We construct survey-based material welfare proxies using this data, and compare the households' implied orderings to those obtained from peer rankings.

In the rural Indonesia dataset, where reported rankings are relatively complete and consistent across observers, our method yields complete orderings without ties in all villages. In contrast, in the urban context of Abidjan, Côte d'Ivoire, where observers were asked to rank 14 households in neighborhoods that often contain more than 200, we find that even our approach often fails to produce complete orderings without ties. This result highlights the limitations of using peer rankings in high-density neighborhoods: most observers simply do not know many of the households around them. Using a higher number of observers should in principle yield more precise and complete rankings, albeit at a higher cost. Nevertheless, we show that orderings obtained from the pairwise method systematically outperform those obtained by rank averaging, compared to survey-based measurements of material welfare.

This paper makes a contribution to the literature on rankings. In his careful theoretical analysis of ranking data, Marden (1995) investigates partial orderings (i.e., block designs) in Chapter 11, but only if they are randomly assigned to observers and are of equal length. He also mentions transitivity in passing, without examining its potential as a ranking strategy. We show that, under certain conditions that we identify, our pairwise implementation of transitive closure can handle situations in which partial orderings are of unequal length and are not randomly assigned—e.g., self-selected by observers. Alvo and Yu (2014) study block designs in more detail and discuss various statistical approaches to averaging ranks, including the use of matrix properties to determine if a complete ordering is identified. But they rely on random assignment and do not investigate the potential use of transitive closure. Building on earlier work by Zermelo (1929) and Bradley and Terry (1952), Newman (2022) proposes an algorithmic estimator to aggregate the type of partial ranking data that originates from sporting events, with and without ties. This approach makes a number of functional assumptions and does not rely on transitive closure. We have also found that, in the datasets examined here, it offers little or no improvement relative to the pairwise estimators without testing.[3]

The remainder of the paper is organized as follows. Section 2 presents the main methodological contribution of this paper. Section 3.1 applies our proposed method to rankings data from rural Indonesia and Section 3.2 applies it to rankings data we collected in urban Côte d'Ivoire. Section 4 concludes.

[3]More tangential to our work, Fok et al. (2012) investigate partial orderings with heterogeneity in ranking abilities in the context of mixed-logit estimation. Their objective is to estimate how characteristics of observers or alternatives affect rankings, not to recover an aggregate ordering.

2 Methodology

2.1 Intuition

The aggregation of individual orderings over alternatives has long been an object of study, starting with seminal contributions by the 18th and 19th century thinkers Condorcet, Borda, and Laplace (Black, 1958; Gehrlein, 1983; Tanguiane, 1991). Condorcet examined situations in which individuals sequentially vote on pairs of alternatives, and observed that combining vote results need not produce a transitive aggregate ordering—an occurrence often referred to as Condorcet cycles. Borda examined situations in which individuals rank all the alternatives and these ranks are averaged over all individuals. Although this method may generate ties, it does not produce cycles. Laplace focused on the case in which each of the n ranking individuals k has a well-defined cardinal preference function $u_k(y^i)$ over each i alternative, and reports the value of $u_k(y^i)$ for each of the $i = \{1, ..., m\}$ alternatives. The aggregate ordering is obtained by averaging the reported values. This method is more precise but more demanding for the reporting individuals.

Much attention has been devoted in the social choice literature on the theoretical properties of different elicitation mechanisms (e.g., Arrow, 1950; Harsanyi, 1955; Sen, 1970) and on their robustness to strategic manipulation (e.g., Gibbard, 1973; List, 2022; Satterthwaite, 1975). While we draw inspiration from this literature, our focus is elsewhere: it is on the relative merits of different methods to aggregate *partial* orderings over alternatives reported by individuals who share the *same* cardinal preference function $u(y^i)$. This includes many situations encountered in practice, such as: the ordering of applicants to a job, prize, or teaching program; the ordering of contestants; and the selection of papers for a conference. All these cases often require the aggregation of partial orderings reported by different observers—such as the combination of orderings of batches of applicants reported by different members of a selection committee; the reconciliation of orderings given by multiple judges over different subsets of contestants; or the harmonization of the rankings reported by different conference committee members over subsets of submitted papers.[4] Throughout this paper we will assume that observers report their preference ordering truthfully, an assumption that is reasonable for our purpose.

Since each observer k reports an *ordering* of alternatives, averaging Borda ranks across observers immediately springs to mind as a way of obtaining an aggregate ordering—and is

[4]These examples differ from situations in which reporting individuals have *different* preference orderings and the objective of aggregation is to obtain a correct weighting of population preferences from a sample of reports—e.g., as would be obtained by polling voters, for instance (Tangian, 2000). In these cases, the representativeness of the sample of observers is key. We ignore these cases here.

often the method used in practice. This approach works well when each observer ranks all the m alternatives. But, because of sampling error in reported orderings across observers, it may incorrectly rank some alternatives when observers ranks a different number of alternatives m_k. Normalizing reported ranks ${r_i}^k$ by $m_k + 1$ can mitigate some of the issues arising from variation in the number of ranked alternatives across k observers (e.g., Tangian, 2000; Tanguiane, 1991), but not all, as the two simple examples below illustrate.

Example 1: Consider two observers with common preferences over $m = 6$ alternatives $y_1 < y_2 < y_3 < y_4 < y_5 < y_6$. Imagine that one observer ranks them all while the other only ranks the last two. Averaging the Borda ranks incorrectly puts y_5 (ranked 5th and 1st, so average rank is 3) ahead of alternative y_4 (ranked only once, in 4th place), and it ties alternative y_5 with y_3 and alternative y_6 with y_4. Normalizing the ranks eliminates the ties but still ranks alternative y_4 ahead of y_5.[5]

Averaging ranks can also yield different (incorrect) results depending on the number of observers reporting on specific sets of alternatives.

Example 2: Consider two sets of m_1 and m_2 observers with common preferences over $m = 5$ alternatives $y_1 < y_2 < y_3 < y_4 < y_5$. There are m_1 observers who rank alternatives $\{y_1 < y_2 < y_3\}$ and another m_2 observers who rank alternatives $\{y_3 < y_4 < y_5\}$. The average of the normalized ranks over the two sets of observers is $\{1/4, 2/4, \frac{0.75m_1+0.25m_2}{m_1+m_2}, 2/4, 3/4\}$. This shows that not only are options y_2 and y_4 tied, option y_3 is ranked above y_4 or below y_2 depending on the relative number of observers in each set.

In both examples above, rank averaging fails to recover the common ordering. The main intuition behind our approach is that, in both cases, we can deduce the common ordering by focusing on pairs of ranked alternatives since, by the assumption of common preferences, they are ranked correctly by all observers. This means that the information needed to recover the common ordering *is* present in the data, but it is not recovered by averaging ranks. A different way of aggregating reported orderings is needed. Given how widely rank averaging is used in practice, this realization is of wide practical relevance.

In this paper we propose an alternative *pairwise* aggregation method for "stitching together" partial orderings. Under mild coverage conditions, it recovers the common latent ordering and therefore improves upon rank averaging in settings where observers rank different subsets of alternatives. We demonstrate that this advantage extends to the presence

[5]The average Borda ranks for the six alternatives are $\{1, 2, 3, 4, 3, 4\}$ while the average normalized ranks are $\{0.14, 0.28, 0.43, 0.57, 0.52, 0.76\}$.

of reporting errors, provided either that the variance of errors is sufficiently small or that the number of observers is sufficiently large. When the variance of observation errors is large, both rank averaging and our proposed pairwise aggregation yield aggregate orderings that are estimated imprecisely. To address this, we introduce corrective measures to the pairwise method that produce more reliable results, that is, results that are closer to the common ordering. We then illustrate the benefits of this methodology with different empirical datasets.

2.2 The pairwise method in a nutshell

Our pairwise method for 'stitching together' partial orderings is equivalent to taking the transitive closure of a Boolean matrix representation (Lidl and Pilz, 1997). In practice, we achieve this by building on a well-known result from graph theory, namely, that all the directed paths extending from a node i to other nodes can be calculated by taking successive powers of the $m \times m$ adjacency matrix R of that network: a directed path—i.e., chain of links or pairs—from node i to node j exists if and only if element ij in R^k becomes non-zero for some power $k \leq m$ (e.g., Jackson, 2010). If the matrix of all paths is upper-triangular, the ordering it represents is complete and transitive.

To apply this result to our setting, note that any reported ordering can be represented as a directed graph in which each of the m alternatives is a node and each pairwise rank relationship is a directed link (e.g., Alvo and Yu, 2014). For instance, each m_1 observer in our Example 2 above, ranks alternative y_2 higher than y_1. This can be thought of as a 'looking-up' link going from node i to node j. Similarly for the other pairwise rank relationships reported by m_1 observers. Hence the ordering of the 5 alternatives reported by m_1 observers can be represented by an adjacency matrix A_1 of the form:

$$A_1 = \begin{bmatrix} . & 1 & 1 & 0 & 0 \\ 0 & . & 1 & 0 & 0 \\ 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & . \end{bmatrix}$$

where element r_{ij} of A_1 is 1 if m_1 observers report a higher rank for alternative j than i, and 0 otherwise. Since A_1 is not upper-triangular, the ordering is incomplete. The equivalent

matrix A_2 for m_2 observers is:

$$A_2 = \begin{bmatrix} . & 0 & 0 & 0 & 0 \\ 0 & . & 0 & 0 & 0 \\ 0 & 0 & . & 1 & 1 \\ 0 & 0 & 0 & . & 1 \\ 0 & 0 & 0 & 0 & . \end{bmatrix}$$

To 'stitch together' the two orderings, we start by forming the union $A_{1\&2}$ of matrices A_1 and A_2:

$$A_{1\&2} = \begin{bmatrix} . & 1 & 1 & 0 & 0 \\ 0 & . & 1 & 0 & 0 \\ 0 & 0 & . & 1 & 1 \\ 0 & 0 & 0 & . & 1 \\ 0 & 0 & 0 & 0 & . \end{bmatrix}$$

whereby any element that is 1 in either A_1 or A_2 is 1 in $A_{1\&2}$, and 0 otherwise. Taking successive powers of $A_{1\&2}$ reveals directed paths linking $\{y_1, y_2\}$ to $\{y_4, y_5\}$ through y_3. The result is an upper-triangular matrix A of the form:

$$A = \begin{bmatrix} . & 1 & 1 & 1 & 1 \\ 0 & . & 1 & 1 & 1 \\ 0 & 0 & . & 1 & 1 \\ 0 & 0 & 0 & . & 1 \\ 0 & 0 & 0 & 0 & . \end{bmatrix}$$

To verify that this matrix defines a complete and transitive ordering from y_1 to y_5, note that the sum of row i gives the number of alternatives ranked higher than i while the sum of column j gives the number of alternatives ranked lower than alternative j. It follows that the rank of an alternative i is equal to 1 plus its column sum or, equivalently in the transitive case, as m minus its row sum.

In the next two subsections, we first identify what, in the *absence of observation errors*, is required for the pairwise method to recover the common latent ordering with probability 1. We then discuss how the performance of the two methods is affected by the presence of observation errors, and we introduce possible palliative measures that can be taken for rank averaging and the pairwise methods in that case.

2.3 Common orderings reported without error

We now formalize our approach. We continue to focus on the case where all observers share a common complete ordering O_l which we assume to contain no ties.[6] By definition, a complete ordering O of m alternatives associates each alternative with one and only one of the integers from 1 to m. For now, we assume that observers report their ordering without error. We also ignore reports that only contain a single alternative, since they do not contain any rank-relevant information.

2.3.1 Complete orderings

We start by demonstrating the equivalence between the rank averaging and the pairwise method when all observers rank *all* alternatives. Let there be n observers indexed by k. Let r_i^k denote the rank (or order), from 1 (lowest) to m (highest), assigned by observer k to alternative i (without ties). The average rank over observers—or Borda count (Tanguiane, 1991)—is $R_b = \{r_1^b, ..., r_m^b\}$ with:

$$r_i^b = \frac{1}{n} \sum_{k=1}^{n} r_i^k$$

Since, in this special case, reported ranks are identical across observers and equal to the common ordering, their averages $\{r_i^b\}$ form a proper ordering O_b of the alternatives ranging from 1 to m, and $O_b = O_l$.

We now formally introduce the pairwise method. It comes in two variants. The first, which we refer to as pairwise voting, is an immediate application of the model presented in Tangian (2000). It only relies on pairwise ordinal information. The second variant seeks to improve efficiency by making use of the cardinal information contained in within-observer pairwise rank differences.

For the first variant, we start by constructing a matrix P^k of all the ij pairwise comparisons reported by observer k. The elements of matrix P^k are $p_{ij}^k = 1$ if $r_i^k < r_j^k$ and 0 otherwise. Each $p_{ij}^k = 1$ is akin to a 'vote' by observer k in favor of j over i. Let P be the element-by-element average of these matrices over the n observers, with elements:

$$p_{ij} = \frac{1}{n} \sum_{k=1}^{n} p_{ij}^k \tag{1}$$

Each p_{ij} is the proportion of observers who 'vote' for j over i. Next, construct an $(m \times m)$ matrix P_a from P by setting each of its ij elements equal to 1 if $p_{ij} > 0$ and 0 otherwise.

[6] This assumption is natural if alternatives take continuous values, as in this paper.

Since this is akin to having a simple majority of observers 'voting' for alternative j over i, we refer to this approach as pairwise voting: it is purely ordinal (e.g., Tangian, 2000). Because matrix P_a is filled with either 1 or 0, it can be seen as representing a directed network or graph. Diagonal elements are set to 0 by construction, as is done for all network adjacency matrices. In the special case of complete and identical orderings, matrix P_a can be rearranged by suitable permutation into an upper-triangular matrix similar to matrix A above. Hence a complete ordering $O_d \equiv \{r_i^d\}$ can be recovered using either of the two methods illustrated in subsection 2.2: either as 1+ the sum of column i of P_a or as $m-$ sum of row i.

The same dataset can be used to construct an alternative set of $(m \times m)$ matrices D^k in which each element $d_{ij}^k = r_j^k - r_i^k$ is the (signed) rank difference between j and i reported by observer k. Unlike in the pairwise P^k, each $d_{ij}^k \in \{-m+1, ... -1, 1, ...m-1\}$. Now let D be the element-by-element average of these matrices over the n observers, with elements:

$$d_{ij} = \frac{1}{n} \sum_{k=1}^{n} d_{ij}^k \tag{2}$$

An alternative network adjacency matrix D_a is then constructed by setting each of its ij elements equal to 1 if $d_{ij} > 0$ and 0 otherwise. Diagonal elements are similarly set to 0. Unlike P_a, matrix D_a makes use of cardinal information contained in the average rank difference between alternatives and, for this reason, is expected to be more informative when we later introduce reporting errors. When, however, all observers report the same thing, as is assumed in this subsection, matrix $D_a = P_a$ and thus it can be similarly rearranged into an upper-triangular matrix and a complete ordering $O_p \equiv \{r_i^p\}$ can be recovered using either as 1+ the sum of column i of D_a or as $m-$ sum of row i.

Proposition 1: When observers have a common ordering that they report in full without error, then: (a) $d_{ij} = r_j - r_i$ for all ij pairs; (b) $D_a = P_a$; (c) orderings $\{r_i^b\} = \{r_i^d\} = \{r_i^p\} = O_l$.

Proof. (a) This follows by the definition of dij and the fact that all r_i^k are identical. (b) Since all observers rank all alternatives in the same way, it follows that all p_{ij}^k are identical. Hence their average value p_{ij} is either 1 or 0, depending on whether $r_j >$ or $< r_i$ in O_l. (c) Since rank ordering $\{r_i^d\}$ is derived from D_a without any matrix multiplication, it has to be identical to $\{r_i^a\}$ since, in that case, D_a is nothing but the complete latent ordering presented in matrix form. □

2.3.2 Partial orderings

Proposition 1 demonstrates that when observers report identical complete orderings, there is no reason to use pairwise methods D or P when the simpler rank averaging method delivers the common ordering. Things are different when each observer k only ranks a different subset m_k of alternatives. We refer to this as partial (or incomplete) orderings. It is in this case that the pairwise approaches work better than rank averaging. To show this, we continue to assume that all observers have a *common* latent ordering O_l on the whole set of m alternatives, but each observer only ranks a subset of alternatives of size $m_k < m$.

As we illustrated in Example 2 above, when observers report partial (or incomplete) orderings of alternatives, rank averaging suffers from *sampling error*. One possible sufficient condition for averaging normalized ranks to recover the common latent ordering is what Marden (1995) in Chapter 11 calls a 'balanced incomplete block design' whereby each observer is assigned to rank a subset (i.e., 'incomplete set') of alternatives such that all equal size combinations of alternatives are covered by at least one observer (i.e., 'block design'), and the number of observers ranking any subset of alternatives is the same (i.e., 'balanced'). This requires the researcher to control which subset of alternatives observers rank, and a large enough number of observers to form a balanced block design.

An alternative sufficient condition is that each observer be assigned (by the researcher or by nature) a *randomly selected* subset of $m_s < m$ alternatives (Marden, 1995), in which case the average of the observers' normalized orderings converge, as the number of observers increases, to a sequence of numbers that defines an ordering equal to the common latent ordering. For instance, let $m_s = 3$ in Example 2 and let the number of observers tend to ∞. In this case, the averaged normalized ranks converge to a sequence $Q = \{1/4, 1.5/4, 2/4, 2.5/4, 3/4\}$ as the number of observers gets large enough. While Q itself is not an ordering, a proper ordering (e.g, a Borda count) can be constructed from it by ordering the alternatives according to Q. Since, in most applications, the number of observers is limited and the assignment of alternatives to observers is either outside the control of the researcher or not guaranteed to be random, these conditions are seldom satisfied in practice.

When neither of Marden's conditions are satisfied, averaging Borda counts from 1 to m_k results in a sampling bias caused by the fact that individuals ranked by observers with a large observed set m_k are unduly advantaged since alternatives ranked by observers with a small m_k cannot receive the same high Borda count. As illustrated in Example 1, this bias can be reduced by relying instead on normalized ranks $z_i^k \equiv \frac{r_i^k}{m_k+1}$ for all $i \in m_k$ and 0

otherwise. Averaging over observers yields:

$$r_i^b = \frac{1}{n_i} \sum_{k \in N_i}^{n} z_i^k$$

where N_i is the set of observers who rank alternative i and n_i is their number. As before, $R_b \equiv \{r_1, ..., r_m\}$, but since each z_i^k takes values between 0 and 1 by construction, R_b is not a proper ordering, that is, a sequence of integers from 1 to m. It can nonetheless be turned into an ordering O_b by sorting alternatives by their r_i^b value. As Examples 1 and 2 above have demonstrated, O_b need not be equal to O_l, that is, the rank averaging method is not guaranteed to recover the common latent ordering. It is in such situations that pairwise methods come in handy because they can recover the common ordering even with a very small number of observers, as we now show.

To demonstrate that pairwise methods D or P can potentially recover the common ordering with few observers, we start by calculating pairwise p_{ij}^k and d_{ij}^k for each observer as before, and add them to construct matrices D and P. Some observers do not rank alternative i relative to j, in which case p_{ij}^k and d_{ij}^k are missing. Let N_{ij} be the set of observers ranking i relative to j and n_{ij} be their number. The corrected formulas are:

$$p_{ij} = \frac{1}{n_{ij}} \sum_{k \in N_{ij}} p_{ij}^k \tag{3}$$

$$d_{ij} = \frac{1}{n_{ij}} \sum_{k \in N_{ij}} d_{ij}^k \tag{4}$$

For those ij pairs that are not ranked by any observers, p_{ij} and d_{ij} are set to 0.[7] The desired matrices are $P \equiv [p_{ij}]$ and $D \equiv [d_{ij}]$. Adjacency matrix D_a is then formed by setting each of its elements equal to 1 if $d_{ij} > 0$ and the rest equal to 0; and adjacency matrix P_a is formed by setting its elements equal to 1 if $p_{ij} > 0.5$—which is equivalent to majority voting on pairs of alternatives (Tangian, 2000). Diagonal elements are set to 0 by construction, as before.

We then iteratively take $m-1$ powers of adjacency matrix P_a to obtain a ranking matrix $E_{up}^p \equiv [\eta_{ij}^p]$ made of only 0's and 1's. Any ij element that turns positive in any of the $m-1$ powers of matrix P_a is set to 1 in E_{up}^p; others are set to 0. This procedure identifies all the directed paths leading 'up' or *from* node i to other nodes, that is, all the alternatives ranked above i. To identify all the alternatives ranked lower than i, that is, leading 'down' from i, we apply the same procedure to the transpose P_a'. This results in a matrix $E_{down}^p \equiv [\eta_{ji}^p]$ that

[7] And similarly for p_{ji} and d_{ji}.

identifies all the directed paths leading *to* i from other nodes, that is, all the alternatives below i. We then combine the results into an index z_i^p defined as:

$$z_i^p \equiv \frac{1}{2}\frac{P_i^p + m + 1}{m+1} \tag{5}$$

with:

$$P_i^p \equiv \sum_j \eta_{ji}^p - \sum_j \eta_{ij}^p$$

where $\sum_j \eta_{ji}^p$ is the number of alternatives ranked below i and $\sum_j \eta_{ij}^p$ is the number of alternatives ranked above i. The same procedure can be applied to adjacency matrix D_a to construct the ranking matrix E_{up}^d and E_{down}^d, and compute z_i^p. As discussed before, these indices can then be turned into an ordering by ranking alternatives according to their z_i^p or z_i^d value, from the lowest to the highest.

When the above procedure identifies the common ordering O_l, then $m - \sum_j \eta_{ij}^d = 1 + \sum_j \eta_{ji}^d$, i.e., the rank r_i of alternative i can be calculated using either of these two sums, and z_i^d boils down to a normalized rank $\frac{r_i}{m+1}$. Identifying the common ordering requires that all *consecutive* pairs in common ordering O_l be ranked by at least one observer k.[8] While this condition cannot be verified ex ante, it is verifiable ex post: if the pairwise algorithm produces a complete transitive ordering of the alternatives, then the condition must be satisfied. This is because when the condition is not satisfied, there will be no directed link between two consecutively ranked nodes in O_l. For instance, if observers only report that $y_1 < y_2$ and $y_1 < y_3$, but not that $y_2 < y_3$, we cannot tell the relative ranks of alternatives 2 and 3. This follows from the properties of directed graphs/networks: for a graph to contain a directed path leading from node 1 to node 3 via node 2, there has to be a directed link from 2 to 3 (Lidl and Pilz, 1997).

When the identification condition is not satisfied, the recovered ordering typically has multiple branches and isolated leaves leading into or from the rest of the network. There can also be multiple nodes in the middle of the ordering that are not ranked relative to each other: there are ties which show up as non-zero elements in the lower-triangle of adjacency matrices E_{up}^d and E_{up}^p. That's when formula (5) comes into its own: it summarizes, in a parsimonious manner, all the recoverable information about the ordering.

Proposition 2: When observers have a common ordering that they report partially but without error, then: (a) a complete ordering O_b of the m alternatives is produced if each

[8] For instance, if $O_l = \{y_1 < y_2 < y_3 < y_4\}$, this means that pairs $\{1,2\}$, $\{2,3\}$ and $\{3,4\}$ are each ranked by at least one observer. Other pairs $\{1,3\}$, $\{1,4\}$, and $\{2,4\}$ need not be ranked by observers because they will be recovered through the matrix multiplication algorithm described earlier.

alternative in $\{1, ..., m\}$ appears in at least one observer's ordering $\{r_i^k\}$, but O_b need not be equal to O_l; (b) O_b converges asymptotically to O_l as $n \to \infty$ when alternatives are assigned independently and randomly to observers; (c) irrespective of the value of n (the number of observers) and even if alternatives are not randomly assigned to observers, O_d and O_p recover the common ordering when all *consecutively ranked* pairs in common ordering O_l are ranked by at least one observer k.

Proof. (a) The first part of the statement follows mechanically from the rank averaging formula: for an alternative to appear among the average ranks, it has to be ranked by at least one observer. To demonstrate the second part, we only need to prove it can happen. This was done in Examples 1 and 2. (b) See Appendix A. (c) This results from the multiplicative properties of adjacency matrices (Jackson, 2010; Lidl and Pilz, 1997; Vega-Redondo, 2007). □

Proposition 2 implies that when we obtain a complete ordering of the alternatives O_p starting from matrix P, or O_d starting from matrix D, then $O_p = O_d = O_l$. This property hinges heavily on the assumption that partial orderings are reported without error. Proposition 2 also implies that recovering the common ordering O_l by a pairwise method does *not* depend on asymptotics: it can be achieved even with a small sample of observers, provided they report the common ordering truthfully and accurately. It also does not rely on random assignment of alternatives to observers: it can be achieved even with purposeful assignment or when observers self-select the alternatives they choose to rank. This stands in contrast with average ranking methods that crucially rely on both. This means that pairwise methods *can* recover the true ordering O_l in situations where rank averaging cannot.

Proposition 2 also brings to light the respective weaknesses of the rank averaging and pairwise methods. Part (b) indicates that the asymptotic consistency of the averaging of normalized ranks relies critically not just on the random assignment of rankable sets to observers, but also on the random assignment of *individual* alternatives to observers. This rules out correlated rankable sets, whereby observers order subsets of alternatives in which alternatives have proximate ranks in the latent ordering O_l.[9] This could arise by design (e.g., stratified assignment of observers) or by self-selection (e.g., observers select the alternatives they rank, and which alternatives each observer selects varies systematically with their own characteristics). Pairwise methods are not vulnerable to these issues.

Part (c) of Proposition 2 brings to light the main weakness of the pairwise methods: since the true ordering O_l is, by construction, unknown to the researcher, recovering O_l

[9] E.g., if observers are asked to rank households by their material welfare and some observers ranked mostly high welfare households while others rank mostly low welfare households.

with certainty *ex ante* requires having all pairs of alternatives ranked by observers. This, by extension, implies that the expected accuracy of the pairwise methods depends on the proportion of *pairs* of alternatives that are ranked: the larger the fraction of possible pairs that are not ranked relative to each other, the least likely will condition (c) in Proposition 2 be satisfied—in which case the pairwise methods are not guaranteed to recover the true ordering O_l.

To illustrate, let us assume that alternatives are independently assigned to observers, which allows us to compare rank averaging with pairwise methods. Let m_b be the number of alternatives ranked by each of n observers. The probability that any alternative is ranked by a specific observer is $\frac{m_b}{m}$ and the probability that an alternative is ranked by at least one observer is thus $1-\left(1-\frac{m_b}{m}\right)^n$. The probability that a specific ij pair is ranked by a specific observer is the number of pairs ranked by each observer, which is $\frac{m_b(m_b-1)}{2}$, divided by the total number of pairs $\frac{m(m-1)}{2}$—that is, $\frac{m_b(m_b-1)}{m(m-1)}$. It follows that the probability that any pair is ranked by at least one observer is $1-\left(1-\frac{m_b(m_b-1)}{m(m-1)}\right)^n$. It immediately follows that, for a given n and m_b, the proportion of unranked pairs $\left(1-\frac{m_b(m_b-1)}{m(m-1)}\right)^n$ falls at a faster rate with m than the proportion of unranked alternatives $\left(1-\frac{m_b}{m}\right)^n$. In other words, other things being equal, the accuracy of the pairwise methods falls rapidly with m, the number of alternatives.[10]

This comparison of the relative weaknesses of the rank averaging and pairwise methods can therefore be summarized by observing that, while the pairwise methods have a strong advantage in samples with few alternatives, this advantage falls rapidly as the number of alternatives increases.

We also note that since the rank averaging method always mathematically produces an ordering O_b (possibly with some ties), and since O_b is not guaranteed to be equal to O_l by Proposition 2, this means that rank averaging methods fail to disclose the uncertainty of what they report. The pairwise methods do not: whatever partial information can be recovered from the data is presented accurately in E^m_{up} and E^m_{down} for $m=\{d,p\}$. Orderings that deviate from the complete ordering often are non-transitive, implying that their network representation does not take the form of a series of links from the lowest rank node to the highest ranked node. First, some alternatives may not be ranked at all. This can only arise when no observer has ranked it relative to others. Such alternatives cannot, therefore, be included in either O_b, O_d or O_p. Second, orderings may be disjoint, that is, there may exist one or more subsets of alternatives that are not ranked relative to any of the remaining

[10]Similarly, keep m_b/m at a constant $\kappa = 10\%$. The number of observers needed to ensure that each alternative is ranked at least once with a 99% probability is around 44. This number rises to around 460 for pairs of alternatives.

alternative in O_l. The network representation of the ordering then has multiple components. When this arises, it is often the case that one component is much larger than the others (e.g., Jackson, 2010).[11] Third, within a component, there can be side branches with one or more links attached to a line running through multiple nodes. In these cases, the nodes in the side branches are not ranked relative to *all* the nodes in the main ordering. For instance, we may have evidence that $y_1 < y_2 < y_4$ and that $y_3 < y_4$. But we do not know where to place y_3 relative to y_2 and y_1. There may also be splits, whereby two or more nodes are reported to be ranked above a node i and below a node j, but are not ranked relative to each other. In such cases, it makes sense to regard them as ties. Our proposed indices z_i^d and z_i^p do away with all this complexity by treating alternatives that are not ranked relative to each other as ties. This allows comparing these non-transitive orderings with orderings coming from a rank averaging method. Orderings based on z_i^d and z_i^p typically have more ties but, in the cases considered in this subsection, these ties present a more honest picture of the amount of ranking information that can be recovered from the data.

2.4 Common orderings reported with error

We now introduce random noise in the orderings reported by individual observers, while maintaining the assumption of a common latent ordering. Formally, let $y = \{y_i\}$ be a vector that represents the common latent value that each observer attaches to each of the n alternatives. These latent values produce the common ordering O_l. Latent values are observed with noise by each observer, i.e., $\tilde{y}_i^k = y_i + \epsilon_i^k$ where ϵ_i^k is the realization of a disturbance term with mean zero and fixed variance σ^2. We assume throughout that the distribution of ϵ_i^k is i.i.d. across observers and alternatives. The orderings reported by each observer is obtained by ranking alternatives by their observed values $\tilde{y}_i^k$, not the true latent values y_i.

2.4.1 Complete orderings with reporting errors

We first discuss what happens with complete orderings (i.e., all n observers rank all m alternatives). In this case, the application of the central limit theorem guarantees that, with a large enough sample, the application of the rank averaging and pairwise methods converge to the latent ordering—see Tangian (2000) for a formal proof covering both methods.

[11] In empirical analysis, it is possible to keep track of this phenomenon by recording whether an alternative belongs to the largest component—and is therefore ranked relative to a larger number of alternatives.

Proposition 3: Assume complete orderings. (a) Given a number of alternatives m and a number of observers n, there exist a variance σ^2 small enough that Proposition 1 holds. (b) Given a variance σ^2 and a number of alternatives m, pairwise ranks converge to latent pairwise ranks for a finite sample size n. (c) When pairwise convergence is achieved, the pairwise methods based on P or D recover the complete latent ordering O_l.

Proof. The variance of the noise in $\tilde{y}_i^k$ must be large to change the ordering reported by observers. When the variance of the noise is small enough, reported orderings remain unaffected and observers report the common latent ordering. This proves Part (a). When σ^2 is large enough to distort individual orderings, the orderings reported by individual observers, and thus the reported ranks, vary across observers. A larger sample of observers is then needed to average out the errors in reported orderings through formulas (1) and (2). Tangian (2000) uses the central limit theorem to demonstrate that pairwise rank differences averaged across observers converge to latent pairwise rank differences for a large enough sample of observers. He also calculates this minimal sample size for a special case, but the intuition holds more generally. When convergence is achieved, matrices D_a and P_a are purged of observation error ϵ_i^k, and Part (c) then follows by application of Proposition 2. □

Proposition 3 mirrors Proposition 1 in the presence of noise: with complete orderings, the presence of reporting noise introduces a need for a large enough sample to eliminate the effect of the noise. In this case, we expect convergence to be faster with the D method than the P method since the former implicitly uses cardinal information on the distance between values of y_i for different alternatives that the P method ignores. Put differently, the distortion to estimated aggregate orderings may vary between methods before convergence to common average orderings is achieved. But since both the accuracy of the rank averaging method and that of the pairwise methods rest on the same condition, namely that averaged ranks—and thus averaged differences in ranks—be consistent with the common ordering, they converge at the same time in the case of complete orderings. This means that, in this case, there is no anticipated advantage from using the pairwise methods.

2.4.2 Partial orderings with reporting errors

Partial orderings is where pairwise methods can outperform rank averaging. If sample sizes are sufficiently large for the reported partial orderings to converge to the common ordering (e.g., Tangian, 2000; Tanguiane, 1991), Proposition 2 applies. If convergence is achieved, the respective strengths and weaknesses of the two methods are those discussed in subsection 2.3.2: even when reporting error is eliminated by averaging, sampling error continues to distort averaged ranks, in which case pairwise methods generates a better ordering as long

as all *consecutively ranked* pairs in the common ordering are ranked by at least one observer. We already pointed out in subsection 2.3.2 that, other things being equal, the latter condition is less likely to be satisfied when m, the number of alternatives, is large.

The same forces affect the absolute and relative performance of rank averaging and pairwise methods when convergence to the common ordering is not achieved: sampling error, compounded by reporting error, biases the results of rank averaging; and reporting error affects the pairwise averaging formulas (3) and (4). But a new complication arises for pairwise methods: errors in (3) and (4) result in incorrect pairwise rankings when constructing matrices D_a and P_a. The contamination of these matrices then creates back loops when taking powers of D_a and P_a.[12] To illustrate this contamination process, consider the following example:

Example 3: Suppose the common ordering is $y_1 < y_2 < y_3 < y_4 < y_5$. There are two reported orderings: $y_1 < y_2 < y_3$ and $y_3 < y_4 < y_2 < y_5$. The second ordering contains a misranking that will propagate to the reconstructed 'looking up' and 'looking down' matrices E^w_{up} and E^w_{down} for $w = d, p$, since there is a looking-up path looping back from alternative 4 to alternative 2. As a result, the number of nodes above and below each of the five alternatives become: $\{4, 3, 3, 3, 0\}$ and $\{0, 3, 3, 3, 4\}$, respectively. Hence the values of P^w_i are $\{-4, 0, 0, 0, 4\}$, and the corresponding values of z^w_i are $\{1/6, 3/6, 3/6, 3/6, 5/6\}$, resulting in ordering $\{y_1 < y_2 = y_3 = y_4 < y_5\}$. By creating a loop, the reporting error has led to alternatives 2, 3 and 4 being tied: the algorithm is unable to rank them. The misreporting by observer 2 also affects averaged normalized ranks,[13] which are $\{5/20, 11/20, 8.5/20, 8/20, 16/20\}$ for the five alternatives, resulting in incorrect ordering $\{y_1 < y_4 < y_3 < y_2 < y_5\}$.

Example 3 is fairly representative of the different ways that reporting errors affect the two methods. In the pairwise methods, incorrect reported orderings create looping back, resulting in ties. In averaged normalized ranks, incorrect reporting messes up the inferred ordering but need not result in ties. In other words, the average ranks method dissimulates the existence of inconsistent information in observer reports, while the pairwise method exposes it. Some may see this as a disadvantage of the pairwise methods: unambiguous orderings are precious in many social contexts (e.g., hiring, student admission, marking, prize selection). We tend to view the ties generated by backward loops as a useful diagnostic

[12] While back loops are reminiscent of Condorcet cycles, they are not due to differences in latent preference orderings between observers, they arise due to reporting error.

[13] The normalized ranks for the first observer are $\{1/4, 2/4, 3/4\}$ for $\{y_1, y_2, y_3\}$ and those for the second observer are $\{1/5, 2/5, 3/5, 4/5$ for $\{y_3, y_4, y_2, y_5\}$.

on the confidence one should have in orderings obtained from the average ranks method when there are reporting errors. In the empirical part of the paper, we compare the extent to which different methods discriminate between alternatives.

2.4.3 Adding a significance adjustment to reduce back loops

While reporting an abundance of ties may appear more conservative, it may also result from an inefficient use of the available information. Back loops are created by reporting errors which should disappear with a larger sample size. Therefore, we may reduce the incidence of back loops by discarding pairwise rankings that are either too small in terms of rank difference, or are based on a sample of observations that is too small to be reliable.

For this reason, we introduce two additional pairwise methods: the first seeks to improve on formula (3) by rejecting pairwise 'majority voting' results that are too tight to rise above a given Bayesian odds ratio; the second seeks to improve formula (4) by rejecting pairwise rank differences that fail a t-test of significance. The details of both correction methods are given in Appendix B.

By rejecting pairwise links that are deemed less reliable, we hope to eliminate back loops. Of course, we may also eliminate correct pairwise rankings and, by reducing the density of matrices P_a and D_a, reduce the ability of the pairwise method to reconstruct the common ordering O_l. To protect against this possibility, the researcher may want to estimate pairwise orderings using different test rejection thresholds/significance levels, and keep the threshold that produces the most discriminating ordering (i.e., with the lowest proportion of ties).

2.5 Monte-Carlo simulations

To illustrate the relative trade-offs between the different estimation approaches, we simulate their performance when observers' rankable sets only partially overlap with each other. We maintain the assumption of a latent common ordering O_l. All our simulations focus on the relative roles of partial orderings and reporting error.

We compare the performance of the rank averaging method, the pairwise majority voting method, and a "scoring" method based on cardinal reports, in recovering a common latent ordering over 30 alternatives organized in 100 sets—3000 observations in total. The scoring method estimates the common ordering by averaging the values $\tilde{y}_i^k$ of alternatives reported, with error, by observers (i.e., it uses much richer information—reports of quantitative income data—not just ranks). For rank averaging, we use the averaging of normalized ranks instead of the Borda counts in order to reduce bias. The majority voting method P (without significance adjustment correction) is selected here to provide a lower bound on the

performance of pairwise methods and to facilitate comparison with the findings of Tangian (2000), which are based on pairwise voting.

We calculate the Mean Square Error (MSE) of each estimator $e \in$ {s=scoring, b=rank averaging, p=pairwise} as:

$$MSE^e = \frac{1}{3000}\sum_{i=1}^{3000}[\omega_i^e - z_i]^2$$

where ω_i^e is the renormalized ordering of alternatives calculated from estimator e and z_i is the true normalized rank of alternative i in its set.[14] The lower bound of MSE^e is 0 when the estimated aggregate ordering perfectly matches the true ordering. Given our choice of simulation parameters, MSE^e takes value 0.0832 when all alternatives are tied at the mean rank and it is equal to 0.333 when the estimated aggregate ordering is the reverse of the true ordering. This means that any MSE above 0.0832 is worse than having no ranking information. In all simulations we expect the MSE obtained by the scoring method to perform the best because it relies on richer information provided by observers. But it too can diverge from the true ordering due to observation error. Hence the MSE value for the scoring method should be seen as the lowest achievable MSE of the pairwise and rank averaging methods relying on the same observers.

We report in Tables 1 and 2 the results from two sets of Monte Carlo simulations. All simulations use the same vector of value realizations so as to eliminate any additional noise across simulations. Each table reports the three MSE^e for 20 different simulations based on the amount of observation error V from 0 to 0.9 (to recall, the standard deviation of log(value) is 1) and the size of the rankable sets S from 10% to 70% of the set of 30 alternatives in the observer's set. Each table has two panels. In the right-hand panel, we only report the MSE estimates for the alternatives that end up being ranked and the total number of ranked alternatives. Given that methods and simulations vary in the fraction of alternatives that they manage to rank, we also report in the left-hand panel MSE estimates calculated over the whole sample of 3000 alternatives, so as to facilitate comparison. In this panel, unranked alternatives are assigned the median normalized rank of 0.5—meaning 'no ranking information'.

In Table 1, rankable sets are uncorrelated by construction, which means that each observer has an independent probability (S%) of observing each of the 30 alternatives in their set. We know that, in this case, the rank averaging method is expected to do reasonably well, and we want to see how the pairwise method performs relative to it. Results show that,

[14] More precisely, for each estimator e, we rank alternatives by their value of e (allowing for ties), and we then normalize for sample size by dividing these ranks by $1 + m_e$ where m_e is the number of alternatives ranked by e.

for all three methods, estimated aggregate orderings deteriorate with observation error: for the worst case scenario ($V = 0.9$ and $S = 0.1$), orderings estimated by the pairwise and rank averaging methods are both worse than no information, i.e., their MSE exceeds 0.0832. All three methods improve with observer coverage S, as could be expected. For instance, with a coverage of 70% and low observation error ($V <= 0.1$), all methods do quite well. We also see that all three methods fail to rank a sizable proportion of alternatives when coverage is low (e.g., $S = 0.1$ or 0.2). Finally, and more importantly, we note that, in half of the simulations with less observation error, the pairwise method does better than the rank averaging method. This is particularly noticeable in the right-hand panel where we ignore unranked alternatives (which add noise to the MSE's when we include them).

Earlier in this section, we have argued that the pairwise method is expected to outperform rank averaging when rankable sets are correlated. To confirm this prediction, we repeat in Table 2 the same set of 20 simulations with maximally correlated rankable sets, in the sense that the set of alternatives observed by each observer k is contiguous in y_i.[15] As predicted, we find that the pairwise method does much better than rank averaging, outperforming it in 18 of the 20 simulations. In some cases, the difference is qualitatively large. For instance, when $S = 0.2$ and $V = 0$, the pairwise method yields an MSE of 0.061 (0.043 if only using ranked alternatives) while the MSE of the rank averaging method well exceeds 0.0832, meaning that it does worse than having no information. The superiority of the pairwise method is maintained even with a high level of observation error, although both methods perform poorly when coverage is very low ($S = 0.1$). In that case, only the scoring method manages to recover a meaningful ordering estimate of 0.074, which is less than 0.0832 (and 0.063 when ignoring unranked alternatives).

From this exercise we conclude that the pairwise method has a useful role to play in the estimation of orderings when observers rank different subsets of the available alternatives. Furthermore, the pairwise method tends to outperform the commonly used rank averaging method when observers rank subsets of alternatives that are proximate in the common latent ordering (i.e., either all highly ranked or all lowly ranked alternatives in the common ordering)—as is likely to be the case in many empirical applications when observers select which alternatives to rank.

[15] This is achieved by randomly picking an integer U between 1 and $30(1 - S)$ and letting observer k see alternatives U to $\text{int}(U + 30S)$ in the true ordering. To illustrate, if $S = 0.1$ and $U = 5$ then the set of observed alternatives is those with true ranks $\{r_5, r_6, r_7, r_8\}$.

2.6 Precision and confidence intervals

Pairwise methods can yield more ties than rank averaging, and our significance adjustment correction methods are designed to reduce ties by pruning unreliable pairwise links. To compare the *apparent* precision of estimated aggregate orderings across methods—i.e., their ability to discriminate between alternatives—we calculate the Herfindahl-Hirschman Concentration Index (HHI) over estimated ranks. If all alternatives are given the same rank, the index takes value 1, indicating maximum concentration onto a single normalized rank value. If all alternatives receive distinct ranks (no ties), indicating maximum discriminatory power, the index takes the lowest possible value allowed by the number of alternatives.

Discriminatory power, however, is not the same as statistical precision. The simulations presented in the sub-section above have indeed brought to light that orderings obtained by rank averaging or pairwise methods are subject to considerable imprecision—something that is seldom recognized in the estimation of orderings. To assess the accuracy of ordering estimates obtained by any method, we construct bootstrapped confidence intervals for estimated aggregate ranks. Since ranks reported by individual observers are necessarily correlated with each other, we use an observer-level cluster bootstrap that simulates counterfactual samples from the orderings of observers randomly selected with replacement within alternative sets (Cameron and Miller, 2015; Cameron et al., 2008).[16]

The approach is based on the two maintained assumptions behind our methodology, namely that: (1) rankable sets are assigned to observers independently from their preference ordering over all alternatives; and (2) observers have common latent orderings over the full set of alternatives.[17] Under these two assumptions, each reported partial ordering r^k can be seen as an i.i.d. realization of a data generating process that randomly samples from the true orderings. This data generating process can therefore be mimicked by drawing with replacement from the set of observers, so as to produce counterfactual samples of orderings. By calculating rank estimates for each counterfactual sample, we can approximate their distribution in the data generating process. The end result is a confidence interval for each estimated rank, i.e., the range of values in which $(1-\alpha)\%$ of bootstrapped ranks reside, where α is the desired level of significance.

Finally, to summarize precision at the level of the *entire* ordering, we aggregate rank-specific intervals into a single index: the normalized area covered by confidence intervals

[16] Mogstad et al. (2024) offers an alternative method of constructing confidence intervals for ranks. Because they do not have multiple observers, they cannot use our method and must rely on a more convoluted approach that requires more assumptions than the cluster bootstrap.

[17] More generally, the procedure continues to apply when latent orderings are heterogeneous, as long as these orderings are distributed independently from their rankable sets. See Appendix C for discussion of this case.

across alternatives. A value of 0 means that all ranks are perfectly estimated: any sample of observers produces the same ordering. A value of 1 means that the confidence interval for each estimated rank encompasses the full range of feasible ranks, i.e., the estimated ranks are entirely uninformative. Any value in between gives a sense of how accurately the common ordering has been estimated. The performance of an estimator depends on *both* measures: a perfectly uninformative estimator (all ties), for instance, typically will have a zero confidence interval (all alternatives are tied in all samples).

3 Empirical analysis

We apply the methods discussed above to two datasets. In both cases, we compare two different flavors of rank averaging (e.g., Borda count and averaging normalized ranks) to four versions of the pairwise method: the rank difference method D based on formula (4), without and with significance adjustment correction to weed out weak links; and the majority voting method P based on formula (3), similarly without and with significance adjustment correction. Based on the methodology section, we expect the Borda count to perform less well than normalized rank averaging, and the rank difference method to perform better than the majority voting method. We wish to ascertain the magnitude of the advantage conferred by normalization (for rank averaging) and by incorporating the cardinal information on common ordering contained in rank differences (for pairwise methods).

We first apply our various estimators to a secondary dataset from rural Indonesia. This dataset is close to complete orderings of a small number of alternatives with moderate reporting errors, repeated over relatively large number of alternative sets. In this type of data, we expect rank averaging methods to work well, and pairwise methods to match them closely even without the need for significance adjustment correction. We also expect pairwise methods to be nearly as discriminating as rank averaging (as measured by the HHI), to provide a comparable level of precision (as measured by the size of the confidence intervals), and to correlate equally well a survey-based proxy for household material welfare.

The second dataset is a novel dataset from urban Côte d'Ivoire where, due to the high-density setting, individuals may know fewer people by name, but potentially observe more about those in close proximity. This data was collected for the purpose of investigating the efficacy of the pairwise method in a context for which it is potentially well suited: a small number of alternatives partially ranked with error by a relatively large number of observers, repeated over a number of alternative sets. In this setting, we expect rank averaging to produce mis-ranking due to the sampling error induced by partial reporting. We wish to ascertain whether pairwise methods can improve on this situation in spite of considerable

reporting error.

In Appendix E, we further apply our methods to three settings in which partial subsets of a large number of alternatives are ranked by a moderately large number of observers, with considerable reporting variation among them: the Ethiopia data from Ayalew et al. (2024) in which 85 judges rank 916 applicants; the Indonesia data from Trachtman et al. (2026) in which inhabitants of ten villages rank up to 30 of their neighbors; and the data from golf tournaments studied in Guryan et al. (2009), in which we treat 433 international golfers as alternatives and rankings in 81 tournaments as observers. With partial ranking, pairwise methods offer a possible edge. But as we argued in section 2, this edge may be defeated in situations with a large number of alternatives, a situation in which pairwise methods may produce a lot of ties (high HHI) due to occasional misrankings contaminating the matrix multiplication process. We examine whether this lack of discriminatory power is ameliorated by significance adjustment correction. Furthermore, because of the extent of reporting variation, we expect rank averaging to produce an estimated aggregate ordering that is discriminating (low HHI) but has large confidence intervals. We examine the extent to which this is the case. Detailed results from one dataset are presented in Appendix E, while the other two are briefly summarized.[18]

3.1 Material welfare rankings in rural Indonesia

In a study in rural Indonesia, Alatas et al. (2012) collected orderings of material welfare, from richest to poorest, of randomly selected rural households (i.e., alternatives) in multiple villages (sets of alternative). There are 5,741 ranked households in total, divided into 640 villages. The average number of ranked households per village is 8.9 (9 in 601 villages, 8 in 32, 7 in 4 and 6 in 3). These households are ranked by 5,711 observers, who belong to the sample of ranked households and only rank other households in the same village. On average, each observer ranks 7.4 households, with a minimum of 1 and a maximum of 8. Three quarter of the observers rank exactly 8 households; only 7.9% rank fewer than 6. On average, each alternative is ranked by 7.5 observers, with a minimum of 1 and a maximum of 9. Only 12% of alternatives are ranked by fewer than 7 observers. These conditions are good for rank averaging, but also ideal for the pairwise methods.

Apart from the fact that observers do not always include themselves in their reported ranking, there is near universal overlap in the set of households on which each observer reports an income ordering in a given village. This feature allows Alatas et al. (2012) to construct a unique ranking index $\tilde{r}_i$ for each target household by averaging reported

[18] Estimation results have been shared with the authors of the two recent papers.

rankings within villages. Since the number of reported ranks on each household is large, this reduces reporting error. In addition, there is sizable concordance in reported rankings across observers: the standard deviation of the ranks reported by different observers around the average normalized rank of a given household is 0.137, compared to a standard deviation of ranks across the entire sample of 0.252.

We apply our pairwise method to the same data in order to obtain four sets of estimates of $\widetilde{z}_i$ for each sampled household in each village: pairwise rank differences, without and with the significance adjustment correction; and majority voting, without and with the significance adjustment correction.[19] From these values, we construct four sets of estimated rank $\hat{r}_i$ by sorting observed households according to $\widetilde{z}_i$ in each village—and taking proper care of ties. We then compare these estimates to the ranks obtained by sorting households by their average Borda count, as done by Alatas et al. (2012),[20], and by their averaged normalized ranks. We find high levels of correlation between all six orderings—the two rank averaging measures and the four pairwise measures. This correlation is 0.99 between orderings based on Borda counts and averaged normalized ranks, and above 0.96 for all other ordering pairs. This does not, however, imply that they are identical: for instance, only 61% of estimated aggregate ranks are identical between the averaged normalized ranks and pairwise rank difference methods—but 98% are within +1 or -1 of each other.

Table 3 presents a comparison of the six estimators. The first column shows the Herfindahl- Hirschman Index of concentration: the higher this value is, the less discriminating the estimated aggregate ordering is. Results show that, as already suggested in the methodology section, rank averaging methods produce fewer ties—and thus less concentrated orderings—than pairwise methods. The averaged normalized ranks method reaches close to the maximum discrimination power of 11.1% that is achievable with 9 alternatives. In this particular dataset, the pairwise rank difference method without testing is the most discriminating.

The reported discriminating power can be misleading if the estimated aggregate ordering is not consistently estimated, as may happen for rank averaging methods with incomplete reported orderings. A highly discriminating estimated aggregate ordering may also hide a great deal of imprecision (see Figure 1). Table 3 shows that all estimators have confidence intervals that, on average, contain between 36% (averaged normalized differences) and 49% (averaged Borda counts) of the range of possible rank values. This means that, had

[19] Because of the detailed data coverage, we chose a generous significance level of 25% (odds of 4 to 1) for the significance adjustment correction. Tighter significance levels yield worse estimates.

[20] In the documentation coming with the dataset available online, Alatas et al. (2012) report using a variable RANK in their analysis. Since this variable is the Borda count divided by 9, it produces the same aggregate ordering.

another set of observers been chosen, the aggregate ordering estimates would have been quite different. As anticipated from the methodology section, pairwise estimators based on rank differences perform better in terms of precision than those based on pairwise majority voting. We also see that pairwise estimators with testing have tighter variance: the proportion of simulated rank estimates that falls outside a $[-1, +1]$ interval is markedly smaller. From this evidence, we conclude that, in this particularly rich dataset with high overlap between rankable sets, the averaged normalized ranks performs best in terms of both discriminating power and reported precision. But all pairwise methods beat the Borda count method on both, meaning what which the choice of rank averaging method matters a lot.

Alatas et al. (2012) also compare average rankings to material welfare measured by consumption expenditure data collected on 5,352 of the ranked households. They find that average rankings are only poorly correlated with reported consumption. We repeat this exercise with all six ordering estimates. Since village-level orderings contain no information on the average level of household consumption in each village, we compare them instead to the within-village *orderings* of households by consumption expenditure. This should give a better chance for estimated aggregate orderings to predict reported consumption. Results are presented in Table 4.[21] We see that the correlation between the two sets of orderings is significant and large in magnitude. We also note that it is of comparable magnitude for all six estimators. Orderings obtained by averaging normalized ranks give the best fit, but three of the pairwise estimates have the same estimated coefficient, and thus the same correlation with consumption orderings, as averaged normalized ranks. The lowest correlation and fit are found for averaging unnormalized Borda ranks, which is the method most commonly used in the literature. To summarize, pairwise methods compare well with rank averaging in this dataset in terms of reported precision and confidence intervals. From Section 2, we also expect that they may also offer benefits in terms of consistency, justifying their additional computational cost.

3.2 Material welfare rankings in urban Côte d'Ivoire

We now perform the same analysis for an urban sample from Côte d'Ivoire. A purposefully recruited sample of observers were asked to rank nearby residents in terms of their material

[21] Chetverikov and Wilhelm (2023) argue that when OLS is used to estimate rank-on-rank correlations, the OLS-reported (robust or not) variance is not consistent in the presence of pointmasses because they induce many ties. The authors introduce a bootstrap correction method that requires knowing the variable from which both ranks are derived. In the case of Table 4, we know the variable behind the consumption ranks, but not that behind the reported ranks, so we cannot apply their method. However, pointmasses are not an issue with the consumption data since it is continuous and the HHI index of all estimators is low in the Indonesian data (see Table 3), indicating few ties. We also report the Spearman p-values at the bottom of the Table, because they are consistent in this case (Chetverikov and Wilhelm, 2023).

welfare (see details in Appendix F). There are 457 ranked households located in 34 separate neighborhoods. They were ranked by 440 individuals located in the same neighborhoods. Of those, 146 observers ranked only one household, therefore providing no usable information.[22] This leaves 294 observers ranking on average 3.8 households, with a minimum of 2 and a maximum of 9. 72% of them rank between 3 and 6 households. On average, a ranked household is ranked 2.7 times, with nearly half of them (46%) ranked only once. These numbers are much lower than we anticipated based on other similar studies, such as those of Alatas et al. (2012) and Trachtman et al. (2026), and they imply a low level of averaging to correct for reporting error in pairwise ranks. This also penalizes rank averaging because of the limited overlap between rankable sets across observers—a feature that increases sampling bias in orderings estimated using rank averaging methods. Which method performs best in such a difficult environment thus remains a valid empirical question.

Comparison of the different estimators A comparison of all six estimated aggregate orderings is presented in Table 5. There is moderate correlation between orderings across methods. The correlation between Borda counts and averaged normalized ranks is only 0.78. On the other hand, the correlation between the pairwise majority voting and pairwise rank difference methods is 0.98 without significance adjustment correction—a figure that betrays the small size of the rankable sets, which mechanically reduces rank differences. With testing, this correlation drops to 0.59, primarily due to the adverse effect that testing has on the pairwise majority voting method in this dataset. We do, however, note a high correlation between the normalized ranks method, the majority voting method without testing, and the pairwise rank difference method with or without testing: all correlations are 0.9 or above.

In terms of the discriminatory power of the reported orderings, none of the methods approaches the minimum value of the HHI concentration index, indicating that maximum discriminating power is not achieved in this dataset and that there are many ties. The most discriminating method is, unsurprisingly perhaps, the averaged normalized rank method, which reaches a 16.7% concentration index. The pairwise rank difference method is a close second, with an HHI of 20.9%. As could be expected, the two pairwise methods without testing perform equally well, given that they are highly correlated.

As noted before for Indonesia, apparent discriminating power hides a lot of imprecision (Figure 2). Table 5 shows that the two best performing methods have bootstrapped confidence intervals that, on average, span from 36% (rank averaging) to around 40% (pairwise rank difference with or without testing) of the possible range. We also find

[22] They included other households in their reported ordering, but these households could not subsequently be identified unambiguously and had to be dropped from the analysis.

that a large fraction of bootstrapped estimates fall outside a $[-1, +1]$ interval around the estimated rank. While these confidence intervals are large, they are nonetheless similar to those obtained in the Indonesia data of Alatas et al. (2012), in spite of the Côte d'Ivoire data having a lower neighborhood coverage in general.

Next, we examine the correlation between estimated aggregate orderings and our main survey measures of welfare collected in the survey: a proxy means test (PMT) index reproducing the method used by the Ivorian government to identify the poor. Results are summarized in Table 6 for our six estimators.[23] The coefficients of the two pairwise estimates with testing are around twice that of the two averaged rank estimate, suggesting that, in very noisy data such as the Abidjan sample, the pairwise method offers a slight improvement. But none of the coefficients is statistically significant, consistent with the sparsity of the ranking information collected among Abidjan respondents.[24] We also investigate whether reported rankings correlate more with the conspicuous consumption expenditures of the target households. They do not. Those results are omitted here to save space.

In Table 7, we repeat the exercise using as dependent variable a dummy equal to 1 if the ordering estimator correctly identifies households below the median PMT index of the sample within each EA.[25] We see that the pairwise difference estimators outperform both averaged ranks estimators, with a large difference in magnitude between them. But none of them is statistically significant.

4 Conclusion

This paper introduced a new method for producing an ordering of alternatives by aggregating partial orderings reported by multiple observers who share a common latent ordering, possibly observed with error. The method relies on the same data as the commonly used Borda count method, but it averages pairwise ranks instead of reported ranks. We offer different flavors of the pairwise method to suit different data configurations. We show, theoretically and through simulations, that pairwise methods can outperform the rank averaging methods when observers only rank a subset of alternatives—especially so when the

[23] As for Indonesia, we report Spearman p-values at the bottom of the Table, as additional check. Detailed estimation results for other measures of household welfare are reported in Table D1 and Table D2.

[24] From Chetverikov and Wilhelm (2023), a bias in OLS-reported standard errors is created by the presence of pointmasses in the regressor. From Table 5, we see that five of the six estimators have a relatively low HHI index, indicating few ties and thus a low potential. Only the pairwise voting estimator with testing has a high HHI, casting doubt on its reported standard error. However, since none of the estimated coefficient is statistically significant, this does not change our conclusion. We nonetheless report the Spearman p-values at the bottom of the Table, as they should not be affected by pointmasses bias.

[25] Detailed estimation results for other measures of household welfare are reported in Table D3 and Table D4.

subset of alternatives that each observer ranks are close to each other in the latent ordering, a phenomenon we call correlated rankable sets.

We compare pairwise and rank averaging methods in two distinct contexts. We first demonstrate that both types of methods fare equally well in data collected by Alatas et al. (2012) in which villagers rank neighbors in terms of perceived material welfare. In this dataset, most observers all rank the same set of neighbors, a feature that minimizes the theoretical advantage of pairwise methods. We show that, in this case, pairwise methods nonetheless perform at par with rank averaging methods in terms of discriminating power and in the ability to predict material welfare measured in a household survey.

Second, we collect our own, similar kind of data in urban Côte d'Ivoire, West Africa, a setting in which observers provide sparse and incomplete rankings and where the potential for correlated rankable sets is high due to geographical segregation. We find that, as a result of incompleteness in reported rankings, both the pairwise and rank averaging methods fail to rank a large proportion of target households. For those they are able to rank, the two approaches produce correlated aggregate—but imprecisely estimated—orderings. A more accurate picture would require increasing the density of reporting. We also find that orderings produced by pairwise methods offer a slight advantage in terms of predicting survey-collected material welfare measures, but this advantage is not statistically significant in the data.

These results suggest that, relative to rural settings where neighbors have lived side by side for decades, urban and peri-urban areas may experience too much spatial mobility to allow neighbors to accurately guess each other's relative economic standing. Alternatively, the challenge may come from social considerations. On the one hand, social arrangements forcing households to share resources with those around them may incentivize relatively well-off individuals to hide their income (Baland et al., 2011)—to appear 'average' to avoid attracting requests for assistance aimed at those who appear too rich. Consistent with this, we find that self-ranking observers tend to rank themselves richer than how others rank them. On the other hand, as we show in a companion paper using data from the same Abidjan setting, relatively poor households may be keen to manipulate their consumption/behavior to appear 'average' and avoid being stigmatized as too poor (Dupas et al., 2025). The combination of these two forces creates a 'race to (appear in) the middle' that is a possible explanation for why it is difficult for observers in Abidjan to infer the material welfare of their neighbors.

What important lessons should practitioners take from this paper? First, the estimation of an aggregate ordering from rankings reported by observers is an inherently challenging process. This is best illustrated by the calculations reported in Tangian (2000) who, at the end of his paper, calculates that 6000 is the number of poll respondents needed to recover,

with a 99% probability, a population's transitive ordering over 5 candidates. Many villages or urban neighborhoods do not even have 6000 potential respondents, let alone respondents with the necessary information. Given this, it should be good practice for practitioners to report confidence intervals around their estimated aggregate orderings, e.g., using cluster bootstrapping. Second, try different methods: the four pairwise ranking methods proposed here require the same data as rank averaging. Third, if you are using rank averaging, average normalized ranks instead of Borda counts, and turn rank averages into proper orderings before comparing them to (proper) orderings constructed from survey data or other sources. Fourth, if it is not possible to require all observers to rank all alternatives, assign alternatives to observers using a balanced block design. If this too is not possible, ensure or verify that alternatives are randomly assigned to observers. If this fails, rank averaging is not reliable: prioritize pairwise methods to recover the aggregate ordering. Fifth, minimize reporting error: (1) focus the attention of observers on an identifiable and observable aspect of the rankable alternatives; (2) facilitate the ranking process, e.g., through props or visual aids; and (3) incentivize observers to do a good job. Researchers may also consider first training observers to accurately rank alternatives that are known to the researchers (e.g., provided by the researcher to the observer), before asking them to rank alternatives unknown to the researcher.

References

Alatas, Vivi, Abhijit Banerjee, Arun G. Chandrasekhar, Rema Hanna, and Benjamin A. Olken, "Network Structure and the Aggregation of Information: Theory and Evidence from Indonesia.," *American Economic Review*, 2016, *106* (7), 1663–1704.

_ , _ , Rema Hanna, Benjamin A. Olken, and Julia Tobias, "Targeting the Poor: Evidence from a Field Experiment in Indonesia.," *American Economic Review*, 2012, *102* (4), 1206–4–.

_ , _ , _ , Benjamin Olken, Ririn Purnamasari, and Matthew WaiPoi, "Does Elite Capture Matter? Local Elites and Targeted Welfare Programs in Indonesia," *AEA Papers Proceedings*, 2019, p. 109: 334–339.

Alvo, Mayer and Philip LH Yu, *Statistical methods for ranking data*, Springer, 2014.

Arrow, Kenneth J, "A Difficulty in the Concept of Social Welfare," *Journal of Political Economy*, 1950, *58* (4), 328–346.

Ayalew, Shiberu, Shanti Manian, and Ketki Sheth, "Discrimination and Access to Capital: A Field Experiment in Ethiopia," Technical Report, Washington State University 2024.

Baland, Jean-Marie, Catherine Guirkinger, and Charlotte Mali, "Pretending to be poor: Borrowing to escape forced solidarity in Cameroon," *Economic development and cultural change*, 2011, *60* (1), 1–16.

Banerjee, Abhijit, Rema Hanna, Benjamin Olken, and Sudarno Sumarto, "The (lack of) Distortionary Effects of Proxy-Means Tests: Results from a Nationwide Experiment in Indonesia," *Journal of Public Economics Plus 1*, 2020.

Basurto, Maria Pia, Pascaline Dupas, and Jonathan Robinson, "Decentralization and efficiency of subsidy targeting: Evidence from chiefs in rural Malawi," *Journal of public economics*, 2020, *185*, 104047.

Black, D., *The Theory of Committees and Elections*, Cambridge University Press, 1958.

Blanchflower, David G. and Andrew J. Oswald, "Well-being over time in Britain and the USA," *Journal of Public Economics*, 2004, pp. 88(7–8), 1359–1386.

Bradley, R. A. and M. E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, 1952, *39*, 324–345.

Cameron, A. Colin and Douglas L. Miller, "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 2015, *50* (2), 317–372.

_ , Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

Chetverikov, Denis and Daniel Wilhelm, "Inference for Rank-Rank Regressions," *The Review of Economic Studies*, 2023, *90* (5), 2296–2330.

Cruces, Guillermo, Ricardo Perez-Truglia, and Martin Tetaz, "Biased Perceptions of Income Distribution and Preferences for Redistribution: Evidence from a Survey Experiment," *Journal of Public Economics*, 2013, *98*, 100–112.

Diamond, Alexis, Michael Gill, Miguel Rebolledo Dellapiane, Emmanuel Skoufias, Katja Vinha, and Yiqing Xu, "Estimating Poverty Rates in Target Populations: An Assessment of the Simple Poverty Scorecard and Alternative Approaches," Technical Report, Poverty Equity Global Practice Working Paper 080, The World Bank 2016.

Dupas, Pascaline, Marcel Fafchamps, and Laura Hernandez-Nunez, "Keeping Up Appearances: An Experimental Investigation of Socio-Economic Signaling to Avoid Discrimination," Technical Report, National Bureau of Economic Research 2025.

Fafchamps, Marcel and Forhad Shilpi, "Subjective welfare, isolation, and relative consumption," *Journal of Development Economics*, 2008, *86* (1), 43–60.

Fok, Dennis, Richard Paap, and Bram Van Dijk, "A rank-ordered logit model with unobserved heterogeneity in ranking capabilities," *Journal of Applied Econometrics*, 2012, *27* (5), 831–846.

Gehrlein, W. V., "Condorcet's paradox," *Theory and Decision*, 1983, *15*, 161–197.

Gibbard, Allan, "Manipulation of Voting Schemes: A General Result," *Econometrica*, 1973, *41* (4), 587–601.

Guryan, Jonathan, Kory Kroft, and Matthew J. Notowidigdo, "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments," *American Economic Journal: Applied Economics*, October 2009, *1* (4), 34–68.

Harsanyi, John C, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy*, 1955, *63* (4), 309–321.

Hussam, Reshmaan, Natalia Rigol, and Benjamin N Roth, "Targeting high ability entrepreneurs using community information: Mechanism design in the field," *American Economic Review*, 2022, *112* (3), 861–98.

Jackson, Matthew, "Social and Economic Networks," *Princeton University Press*, 2010, *Princeton.*

Layard, Richard, "Well-Being Measurement and Public Policy," *NBER Chapters, in: Measuring the Subjective Well-Being of Nations: National Accounts of Time Use and Well-Being*, 2009, pp. 145–154.

Lidl, R. and G. Pilz, *Applied abstract algebra, Undergraduate Texts in Mathematics (2nd ed.)*, Springer, 1997.

List, C., "Social Choice Theory," in "The Stanford Encyclopedia of Philosophy," E. N. Zalta U. Nodelman (Eds.), Metaphysics Research Lab, Stanford University, 2022.

Marden, John I, *Analyzing and modeling rank data* Monographs on Statistics and Applied Probability, Chapman and Hall/CRC, 1995.

Mogstad, Magne, Joseph P Romano, Azeem Shaikh, and Daniel Wilhelm, "Inference for Ranks with Applications to Mobility across Neighbourhoods and Academic Achievement across Countries," *The Review of Economic Studies*, 2024, *91* (1), 476–518.

Newman, M.E.J., "Fast computation of rankings from pairwise comparisons," *arXiv*, 2022, *2207*, 00076.

Premand, Patrick and Oumar Barry, "Behavioral change promotion, cash transfers and early childhood development: Experimental evidence from a government program in a low-income setting," *Journal of Development Economics*, 2022, *158*, 102921.

Ravallion, Martin, "Miss-targeted or Miss-measured?," *Economics Letters*, 2008, pp. 100 (1): 9–12.

Satterthwaite, Mark Allen, "Strategy-proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *Journal of Economic Theory*, 1975, *10* (2), 187–217.

Sen, Amartya K, *Collective Choice and Social Welfare*, San Francisco: Holden-Day, 1970.

Tangian, A. S., "Unlikelihood of Condorcet's paradox in a large society," *Social Choice and Welfare*, 2000, *17*, 337–365.

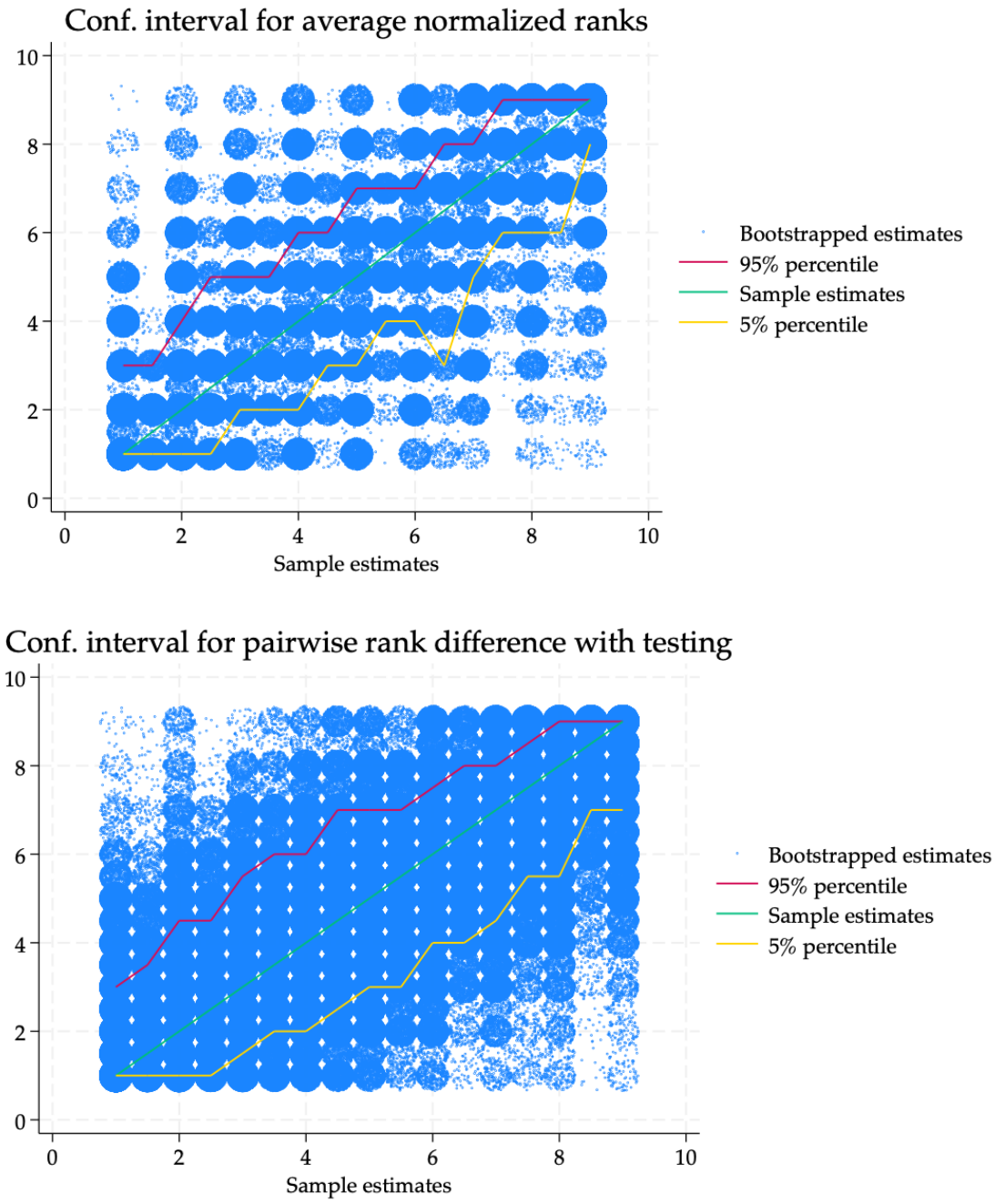
Tanguiane, Andranick S., *Aggregation and Representation of Preferences*, Springer-Verlag, 1991.

Trachtman, Carly, Yudistira Hendra Permana, and Gumilang Aryo Sahadewo, "How much do our neighbors really know? The limits of community-based targeting," *Journal of Development Economics*, 2026, *178*, 103555.

Vega-Redondo, Fernando, *Complex Social Networks*, Cambridge University Press, 2007.

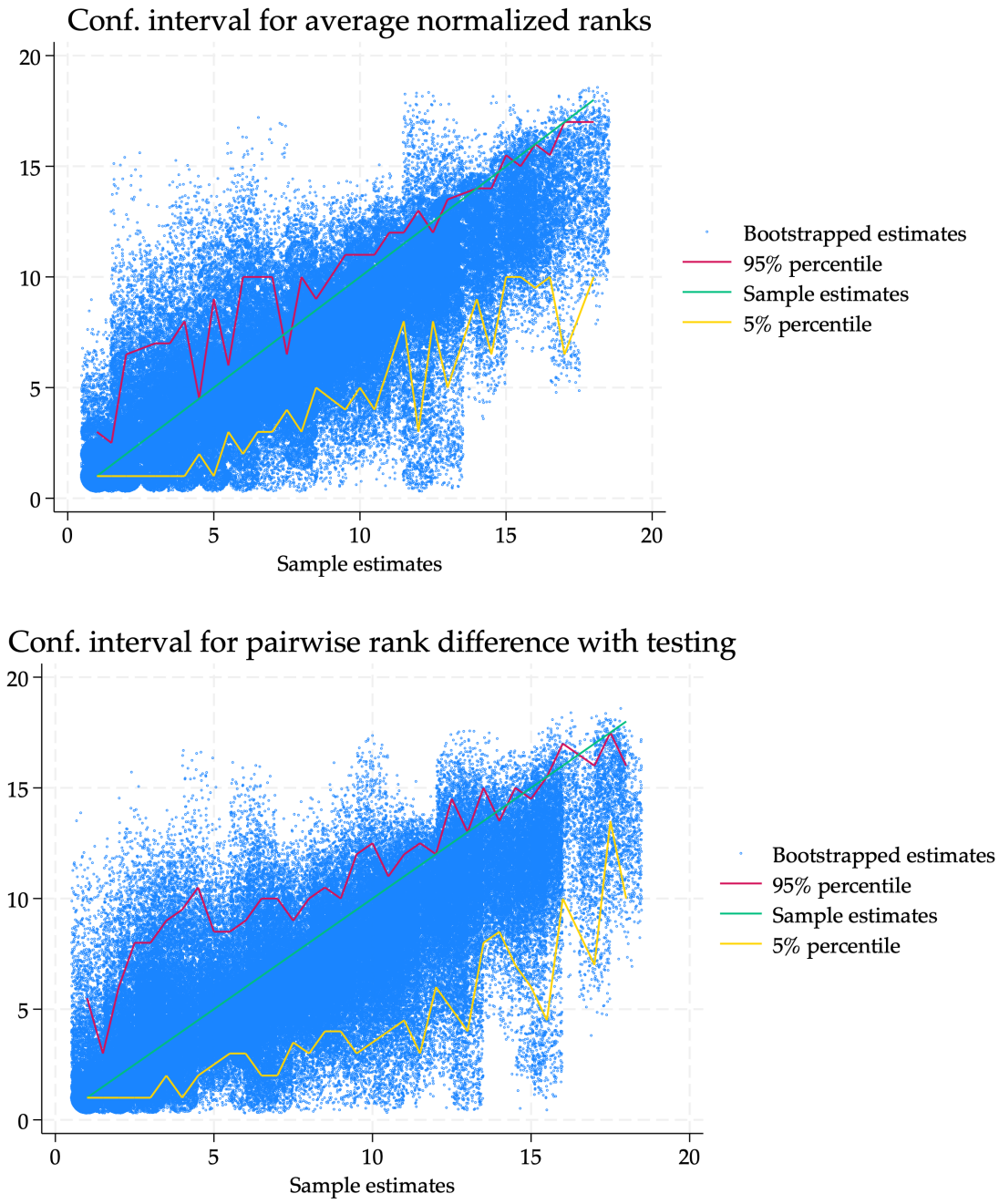
Zermelo, E., "Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, 1929, *29*, 436–460.

Figure 1: Confidence Intervals for selected estimators in the Indonesia data



Notes: Obtained by randomizing observers with replacement.

Figure 2: Confidence Intervals for selected estimators in the Côte d'Ivoire data



Notes: Obtained by randomizing observers with replacement.

Table 1: Mean Square Error - uncorrelated rankable sets

		Using all alternatives				Using only ranked alternatives							
		Size of rankable set S =				Size of rankable set S =							
Estimator:	V =	0.1	0.2	0.4	0.7	0.1	N	0.2	N	0.4	N	0.7	N
Pairwise	0	0.059	0.020	0.001	0.00009	0.042	1,801	0.009	2,579	0.000	2,969	0.00001	2,999
Normalized Ranks		0.060	0.026	0.004	0.00021	0.046	1,849	0.016	2,580	0.002	2,969	0.00013	2,999
Scoring		0.041	0.015	0.001	0.00006	0.015	1,849	0.003	2,580	0.000	2,969	0.00000	2,999
Pairwise	0.1	0.059	0.018	0.003	0.00087	0.043	1,847	0.010	2,623	0.002	2,973	0.00087	3,000
Normalized Ranks		0.061	0.024	0.004	0.00093	0.047	1,905	0.016	2,623	0.004	2,973	0.00093	3,000
Scoring		0.040	0.012	0.002	0.00060	0.014	1,905	0.003	2,623	0.001	2,973	0.00060	3,000
Pairwise	0.3	0.063	0.026	0.013	0.00541	0.049	1,847	0.019	2,623	0.012	2,973	0.00541	3,000
Normalized Ranks		0.064	0.029	0.008	0.00316	0.052	1,905	0.022	2,623	0.008	2,973	0.00316	3,000
Scoring		0.044	0.017	0.005	0.00263	0.022	1,905	0.009	2,623	0.005	2,973	0.00263	3,000
Pairwise	0.5	0.072	0.040	0.038	0.01663	0.064	1,847	0.034	2,623	0.037	2,973	0.01663	3,000
Normalized Ranks		0.071	0.037	0.014	0.00660	0.063	1,905	0.031	2,623	0.014	2,973	0.00660	3,000
Scoring		0.052	0.025	0.011	0.00610	0.034	1,905	0.018	2,623	0.010	2,973	0.00610	3,000
Pairwise	0.9	0.090	0.072	0.066	0.04783	0.094	1,847	0.066	2,623	0.065	2,973	0.04783	3,000
Normalized Ranks		0.085	0.056	0.031	0.01574	0.086	1,905	0.053	2,623	0.030	2,973	0.01574	3,000
Scoring		0.070	0.045	0.028	0.01731	0.062	1,905	0.040	2,623	0.027	2,973	0.01731	3,000

Notes: This Table reports the Mean Square Error of the pairwise majority voting, Normalized Ranks, and scoring methods applied to the same simulated data. Simulations are based on 100 sets of 30 alternatives ranked by 9 observers per set—3000 alternatives in total. V is the standard deviation of the noise added to the 'true' value of a log(income) variable with mean zero and unit variance. Rankable sets are uncorrelated by construction: each observer sees each of the 30 alternatives with equal probability S. It follows that S is the average share of the 30 alternatives ranked by each observer. N is the number of ranked alternatives. We fix the random seed across simulations to ensure that the realizations of income are identical across all parameter values. The Mean Square error is calculated as the square of [(estimated rank minus the true rank) divided by 30]—where estimated rank and true rank are both a number from 1 to 30 and the division by 30 is used to normalize the MSE estimates. With 30 alternatives, the MSE of no information (all ranks tied at 15.5) is 0.0832 and the MSE of the reverse ranking (the worst possible outcome) is 0.333. MSE estimates increase as we move down (more observation noise) and left (less coverage). The pairwise method and the Normalized Ranks methods rely on the identical reported rank data. We show in yellow those simulations in which the pairwise method performs better than the Normalized Ranks, and in green those in which the Normalized Ranks performs better than the pairwise method. The scoring method relies on income reported, possibly with error, by each observer. Reported incomes are averaged across observers within each set to compute ranks in a set. The scoring method performs better than the pairwise and Normalized Ranks methods, but it requires observers to report quantitative income data, not just ranks. In the left-hand panel, MSE calculations includes all alternatives. Unranked alternatives receive a median rank. This serves to compare methods that generate differences in the number of ranked alternatives. In the right-hand panel, MSE's are calculated using ranked alternatives only. There is no difference between the two panels when the number of ranked alternatives is 3000 (the maximum).

Table 2: Mean Square Error - correlated rankable sets

		Using all alternatives				Using only ranked alternatives							
		Size of rankable set S =				Size of rankable set S =							
Estimator:	V =	0.1	0.2	0.4	0.7	0.1	N	0.2	N	0.4	N	0.7	N
Pairwise	0	0.127	0.061	0.022	0.00713	0.136	2,111	0.043	2,563	0.004	2,737	0.00076	2,914
Normalized Ranks		0.136	0.102	0.034	0.00752	0.149	2,111	0.091	2,563	0.017	2,737	0.00120	2,914
Scoring		0.036	0.028	0.020	0.00700	0.012	2,111	0.006	2,563	0.004	2,737	0.00075	2,914
Pairwise	0.1	0.135	0.070	0.023	0.00825	0.148	2,079	0.053	2,537	0.005	2,734	0.00150	2,908
Normalized Ranks		0.143	0.109	0.038	0.00903	0.159	2,084	0.099	2,537	0.021	2,734	0.00221	2,908
Scoring		0.040	0.029	0.020	0.00750	0.015	2,084	0.008	2,537	0.005	2,734	0.00131	2,908
Pairwise	0.3	0.145	0.088	0.031	0.01058	0.163	2,061	0.072	2,535	0.013	2,734	0.00397	2,908
Normalized Ranks		0.152	0.125	0.052	0.01191	0.171	2,084	0.119	2,537	0.037	2,734	0.00505	2,908
Scoring		0.045	0.033	0.022	0.00926	0.023	2,084	0.014	2,537	0.007	2,734	0.00303	2,908
Pairwise	0.5	0.150	0.101	0.048	0.01875	0.171	2,057	0.088	2,533	0.029	2,734	0.01140	2,908
Normalized Ranks		0.155	0.135	0.070	0.01585	0.176	2,084	0.131	2,537	0.057	2,734	0.00898	2,908
Scoring		0.053	0.041	0.028	0.01219	0.034	2,084	0.022	2,537	0.013	2,734	0.00611	2,908
Pairwise	0.9	0.154	0.120	0.076	0.05172	0.176	2,050	0.109	2,529	0.055	2,734	0.04311	2,908
Normalized Ranks		0.157	0.145	0.099	0.02805	0.179	2,084	0.143	2,537	0.088	2,734	0.02159	2,908
Scoring		0.074	0.061	0.045	0.02395	0.063	2,084	0.046	2,537	0.031	2,734	0.01776	2,908

Notes: This Table reports the Mean Square Error of the pairwise, Normalized Ranks, and scoring methods applied to the same simulated data. Simulations are based on 100 sets of 30 alternatives ranked by 9 observers per set—3000 alternatives in total. V is the standard deviation of the noise added to the 'true' value of a log(income) variable with mean zero and unit variance. Rankable sets are correlated by construction: each observer sees S*30 consecutive alternatives, with alternatives ranked by their true income value. A set has 30*(1-S) different sequences of consecutive alternatives. Each of the 9 observers is randomly assigned, with equal probability, to one of these sequences. It follows that S is the share of the 30 alternatives that is ranked by each observer. In the Table, N is the number of ranked alternatives. We fix the random seed across simulations to ensure that the realizations of income are identical across all parameter values. The Mean Square error is calculated as the square of [(estimated rank minus the true rank) divided by 30]—where estimated rank and true rank are both a number from 1 to 30 and the division by 30 is used to normalize the MSE estimates. With 30 alternatives, the MSE of no information (all ranks tied at 15.5) is 0.0832 and the MSE of the reverse ranking (the worst possible outcome) is 0.333. MSE estimates increase as we move down (more observation noise) and left (less coverage). The pairwise method and the Normalized Ranks methods rely on the identical reported rank data. We show in yellow those simulations in which the pairwise method performs better than the Borda method, and in green those in which the Borda method performs better than the pairwise method. The scoring method relies on income reported, possibly with error, by each observer. Reported incomes are averaged across observers within each set to compute ranks in a set. The scoring method performs better than the pairwise and Normalized Ranks methods, but it requires observers to report quantitative income data, not just ranks. In the left-hand panel, MSE calculations includes all alternatives. Unranked alternatives receive a median rank. This serves to compare methods that generate differences in the number of ranked alternatives. In the right-hand panel, MSE's are calculated using ranked alternatives only. There is no difference between the two panels when the number of ranked alternatives is 3000 (the maximum).

Table 3: Comparison of Estimators in the Indonesia Dataset

	Lack of: Discriminating power	Precision	
	HHI		
	(1)	(2)	(3)
Averaging ranks (Borda count)	12.3%	48.8%	47.6%
Averaging normalized ranks	11.8%	36.3%	36.5%
Pairwise majority voting, no test	16.0%	46.2%	42.1%
Pairwise majority voting, with test	15.5%	46.6%	38.6%
Pairwise rank differences, no test	14.2%	40.9%	43.6%
Pairwise rank differences, with test	15.8%	41.6%	33.9%

Notes: The Herfindahl–Hirschman Index (HHI) of concentration is calculated separately for each village and averaged over all villages. With 9 alternatives, the minimum value it can take is 11.1%. Precision measure (1) is the percentage of the maximum inter-rank distance covered by the 95% confidence interval, averaged over all estimated ranks. Precision measure (2) is the percentage of simulated ranks that fall outside the interval [–1,+1] from the estimated rank, averaged over all estimated ranks.

Table 4: Correlation between estimated orderings and material welfare: Rural Indonesia

	(1)	(2)	(3)	(4)	(5)	(6)
	Borda	Norm. Borda	Votes NT	Votes T	Diff NT	Diff T
Rank	0.344***	0.368***	0.369***	0.357***	0.371***	0.367***
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Constant	0.328***	0.316***	0.315***	0.321***	0.315***	0.317***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
R2	0.118	0.136	0.127	0.123	0.132	0.131
Observations	5343	5343	5343	5343	5343	5343
Spearman p-val	0.000	0.000	0.000	0.000	0.000	0.000

Notes: For each column, we use the effective sample size to re-rank *both* variables within EA using only that sample, and re-normalize by (1 + EA sample size). OLS is then run on the re-normalized variables. SEs clustered at the EA level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Columns: *Borda* = Averaging ranks; *Norm. Borda* = Averaging normalized ranks; *Votes NT/T* = Pairwise majority voting w/o and with test; *Diff NT/T* = Pairwise rank differences w/o and with test.

Table 5: Comparison of Estimators in the Côte d'Ivoire Dataset

	Lack of: Discriminating power	Precision	
	HHI		
	(1)	(2)	(3)
Averaging ranks (Borda count)	25.2%	44.3%	80.5%
Averaging normalized ranks	16.7%	36.1%	74.7%
Pairwise majority voting, no test	24.0%	40.5%	70.6%
Pairwise majority voting, with test	51.2%	56.5%	79.4%
Pairwise rank differences, no test	24.7%	39.8%	70.3%
Pairwise rank differences, with test	20.9%	40.6%	71.5%

Notes: The Herfindahl–Hirschman Index (HHI) of concentration is calculated separately for each EA and averaged over all EAs. With 13.4 alternatives on average, the minimum value it can take is 7.4%. Precision measure (1) is the percentage of the maximum inter-rank distance covered by the 95% confidence interval, averaged over all estimated ranks. Precision measure (2) is the percentage of simulated ranks that fall outside the interval [–1,+1] from the estimated rank, averaged over all estimated ranks.

Table 6: Correlation between estimated orderings and the PMT index of material welfare

	(1) Borda	(2) Norm. Borda	(3) Votes NT	(4) Votes T	(5) Diff NT	(6) Diff T
Rank	0.056	0.058	0.054	0.117	0.052	0.099
	(0.081)	(0.084)	(0.103)	(0.105)	(0.103)	(0.094)
Constant	0.472***	0.471***	0.473***	0.442***	0.474***	0.451***
	(0.040)	(0.042)	(0.051)	(0.053)	(0.051)	(0.047)
R2	0.003	0.003	0.003	0.011	0.002	0.009
Observations	220	220	220	220	220	220
Spearman p-val	0.425	0.411	0.457	0.120	0.471	0.175

Notes: For each regression, we first mark the sample with missing handling. We then re-compute *both* the dependent variable and the estimator ranks *within EA and within the regression sample*, and re-normalize by (1+EA sample size). OLS is then re-run on these re-normalized variables. Standard errors clustered at EA level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Columns: *Borda* = Averaging ranks; *Norm. Borda* = Averaging normalized ranks; *Votes NT/T* = Pairwise majority voting w/o and with test; *Diff NT/T* = Pairwise rank differences w/o and with test.

Table 7: Can estimated orderings identify those below the median PMT Index?

	(1) Borda	(2) Norm. Ranks	(3) Votes NT	(4) Votes T	(5) Diff NT	(6) Diff T
Household	-0.002	0.012	0.089	0.078	0.099	0.090
Ranked Below Median	(0.069)	(0.069)	(0.084)	(0.084)	(0.084)	(0.082)
Constant	0.455***	0.450***	0.425***	0.429***	0.422***	0.423***
	(0.032)	(0.030)	(0.029)	(0.029)	(0.029)	(0.030)
R2	0.000	0.000	0.007	0.005	0.009	0.008
Observations	220	220	220	220	220	220

Notes: For each column, we first mark the regression sample by running an auxiliary OLS of the survey outcome on the constructed rank. Within that sample only, we compute within-EA medians for both the survey outcome and the constructed rank, and define below median indicators (strictly below). We then regress the outcome indicator on the constructed-rank indicator; SEs clustered at the EA level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Online Appendix

Table of Contents

A Sketch of the proof of Proposition 2, part (b) **1**

B Significance adjustment for pairwise ranking methods **2**

C Aggregation of heterogeneous latent orderings **4**

D Additional Tables for Côte d'Ivoire **5**

E Other empirical applications **7**

F Côte d'Ivoire Application: Data collection protocol **10**

F.1 Data Collection and Sampling Frame 10

F.2 Ranking Survey Protocol 11

F.3 Description of the sample 12

F.4 Material Welfare Benchmarks 13

F.5 Perceptions of Material Welfare 17

A Sketch of the proof of Proposition 2, part (b)

Proof. We start with a special case when $m = 4$ and $m_b = 2$ for all k. Normalized ranks are $\frac{1}{m_b+1} = 1/3$ and $\frac{2}{m_b+1} = 2/3$. Without loss of generality, let the common ordering O_l be $1, 2, 3, 4$. With random assignment of alternatives, each alternative has a 50% chance of being in the rankable set of any observer k. For alternative 1, the assigned normalized rank is always $1/3$ since any alternative it is paired with has a higher rank in O_l. Similarly, alternative 4 always has rank $2/3$ since it can only be paired with a lower ranked alternative in O_l. For alternative 2, it has a $1/3$ chance of being paired with alternative 1, in which case it gets rank $2/3$, and $2/3$ chance of being paired with a higher ranked alternative, in which case it gets rank $1/3$. With random assignment of alternatives to observers, the sample frequencies converge to these probabilities as the sample of observers gets large. The reverse holds for alternative 3. Hence the expected normalized ranks for the four alternatives are $3/9, 4/9, 5/9, 6/9$. Sample frequencies converge to these values as $n \to \infty$. Since frequencies increase linearly in true ranks r_i, sorting alternatives by their sample frequencies recovers the common ordering O_l when n is large enough. A more formal treatment of this case can be found in Marden (1995), Chapter 11.

This reasoning can be extended to any values of m and m_b: alternatives ranked 1 and m always get a rank $\frac{1}{/}m_b + 1$ and $\frac{m_b}{m_b+1}$ while alternatives 2 to $m_b - 1$ get an expected rank equal to $\frac{1}{m_b+1} + \frac{r_i-1}{m-1}\frac{m_b}{m_b+1}$ where r_i is the rank of alternative i in O_l. Sample frequencies converge to these values as $n \to \infty$. Since expected ranks increase in true ranks, sorting alternatives by their sample frequencies recovers the common ordering O_l when n is large enough. As Tangian (2000) illustrates for a special case, this occurs for a finite value of n.

Since this is true for any rankable set size m_b, adding large enough samples of normalized ranks with different m_k values recovers the true ordering since, for each m_k, the true ordering is preserved as long as all m alternatives have an equal chance of being in each of the rankable sets within each of the samples of rankable size m_k. If, on the other hand, some alternatives have a higher chance of being in rankable sets of a given size—as in example 1 above—then convergence to the true ordering O_l is not guaranteed.

□

B Significance adjustment for pairwise ranking methods

This section presents the details of the significance adjustment correction methods discussed in the paper.

Pairwise comparisons method P When information on pairwise rankings is collected from pairwise voting—or any pairwise ranking method (e.g., a sports championship or tournament)—pairwise method P is the relevant approach. Formally, we assume that, for a number of pairs p_{ij}, the researcher has a number m_{ij} of observations, s_{ij} of which are 1 when $i \prec j$ and the rest are 0 when $i \succ j$.[26] This means that, for each pair ij, variable p_{ij} has a Bernoulli distribution with unknown success probability π_{ij}. In our applications, the sample of observations on each pair ij is small, e.g., $m_{ij} = \{1, 2, 3, 4\}$.

The researcher starts with an uninformative prior over π_{ij} and observes m_{ij} realizations from its Bernoulli distribution. The researcher forms a posterior belief about $\pi_{ij} > 0.5$ (the probability that $y_j > y_i$) after observing s_{ij} success realizations (i.e., 1's) in sample of size m_{ij}. If this belief is large enough, the researcher posits that $\pi_{ij} > 0.5$. More precisely, the researcher decides to believe that $\pi_{ij} > 0.5$ if:

$$Prob[\pi_{ij} > 0.5 | s_{ij}, m_{ij}] > c$$

where c is the chosen threshold for accepting the link. For instance, an odds ratio of 9 to 1 in favor of $p_{ij} > 0.5$ is equivalent to having 90% of the posterior distribution above the value of 0.5. This posterior distribution takes the following form:[27]

$$(\pi_{ij} | s_{ij}, m_{ij}) \sim Beta(b + s_{ij}, b + m_{ij} - s_{ij})$$

where $\pi_{ij} \sim Beta(b, b)$ is the prior distribution. Assuming an uninformed prior, i.e., uniform on the interval $[0, 1]$, means setting $b = 1$. Putting it all together, we have:

$$Prob[\pi_{ij} > 0.5 | s_{ij}, m_{ij}] = \int_{0.5}^{1} Beta(1 + s_{ij}, 1 + m_{ij} - s_{ij}) d\pi_{ij}$$

The researcher then constructs the adjacency matrix P_a by setting element $a_{ij} = 1$ iff $Prob[\pi_{ij} > 0.5 | s_{ij}, m_{ij}] > c$, and 0 otherwise. The threshold value c is then chosen so as to reduce instances of non-transitive rankings (i.e., back loops) created by taking successive powers of P_a.

Partial ordering method D In this method, the construction of adjacency matrix D_a starts from r_i^k and r_j^k observations that each take values between 1 and n. The larger the

[26] We assume here that observers are not allowed to report a tie. The derivation of the posterior distribution below can be adapted to accommodate ties, but this case is ignored here because it does not arise in any of the datasets we examine in this paper.

[27] See the excellent lecture notes:
https://hedibert.org/wp-content/uploads/2014/01/workedexample-bernoullitrials1.pdf.

difference between r_j^k and r_i^k, the more likely it is that $i \prec j$. How confident the researcher is in this comparison, however, depends on how correlated individual rankings r_{ij}^k are across observers: if they are highly correlated, this means that the variance of the noise ϵ_i^k is low, and vice-versa. We can derive an estimate of the variance of ϵ_i^k in a dataset by calculating, across all alternatives, the variance of reported ranks r_i^k from their sample mean r_i. This gives an estimate of the variance $\hat{\sigma}_\epsilon^2$ of both r_i^k and r_j^k. We then use this variance estimate in the standard formula for the test statistic. Since by construction $n_j = n_i \equiv n_{ij}$ for any pairwise comparison, this gives: $\mathrm{t}_{ij} = \frac{r_j - r_i}{\sqrt{\frac{\sigma_{r_j}^2}{n_j} + \frac{\sigma_{r_i}^2}{n_i}}}$ $= \frac{r_j - r_i}{\sigma_\epsilon \sqrt{\frac{2}{n_{ij}}}}$ in which we replace σ_ϵ by its sample estimate $\hat{\sigma}_\epsilon$. The value t_{ij} of the t-statistic is then compared to the t-distribution with $2n_{ij} - 1$ degrees of freedom, and the ij element of matrix D_a is set equal to 1 if the t-test is significant at the $\alpha\%$ level, with α chosen to reduce instances of non-transitive rankings obtained by taking successive powers of D_a.

C Aggregation of heterogeneous latent orderings

In this Appendix, we discuss the applicability of the methodologies we have discussed in the paper to the situation where observers have heterogeneous latent orderings. In this case, the objective of the researcher is often to recover the average preference ordering of the **population**. This case is discussed in detail by Tangian (2000) who shows that, when all observers rank all alternatives, there exist a sample size large enough such that the pairwise ('majority voting') approach P recover the population average. The author shows that this property holds for the Borda count as well.

To extend this property to partial orderings, some conditions are required on the distribution of ranked alternatives across observers. To illustrate, let us assume that each alternative i has an equal probability ρ of being ranked by an observer, and that the realizations of these probabilities are i.i.d. across observers and alternatives. These assumptions guarantee that, in a large enough sample of observers n, each pair of alternatives will be ranked by a large enough number of observers for p_{ij} to approach the population average of individuals favoring j over i. Hence population preference for each pairwise comparison will be recovered. If the population voting preferences satisfy transitivity, the required sample size is finite, as demonstrated by Tangian (2000); Tanguiane (1991).

Population preferences elicited through pairwise majority voting need not, however, satisfy transitivity—a problem known as Condorcet cycles. In that case, the recovered aggregate preferences will not yield a full ordering of the alternatives: subsets of alternatives that fall within a Condorcet cycle are tied. The question then is whether an average Borda count O_b can do better. We first note that, by construction, averaged Borda counts always yield a transitive ordering—possibly with ties. This, however, may be misleading since, as we have shown in the case of partial orderings, rank averaging can produce incorrect orderings. In such cases, pairwise approaches P or D are more accurate: they will document the existence of Condorcet cycles in the E_{down} and E_{up} matrices, and represent them as ties in index z_i.

When Condorcet cycles arise, one obvious solution is to collect cardinal data on preferences, as in the Laplace social welfare approach. In this approach, observers are asked to report the value or score of each alternative to them. These values are then aggregated into social welfare values that are then ranked to obtain the population ordering over alternatives. While intellectually satisfying, this approach is more demanding for observers, and it is vulnerable to manipulation by observers and to inconsistency in the units used by different observers. The rank averaging approach may seem like a compromise between accuracy and feasibility. But we have shown that, when observers report partial orderings, rank averaging suffers from biases due to sampling error. In such cases, pairwise methods tend to dominate rank averaging provided that the variance of reporting errors is small.

D Additional Tables for Côte d'Ivoire

Table D1: Correlation between estimated orderings and the PPI index of material welfare

	(1) Borda	(2) Norm. Borda	(3) Votes NT	(4) Votes T	(5) Diff NT	(6) Diff T
Rank	-0.046	0.011	-0.031	-0.004	-0.023	0.003
	(0.058)	(0.055)	(0.063)	(0.070)	(0.063)	(0.055)
Constant	0.523***	0.494***	0.516***	0.502***	0.511***	0.498***
	(0.029)	(0.028)	(0.031)	(0.035)	(0.031)	(0.028)
R2	0.002	0.000	0.001	0.000	0.000	0.000
Observations	365	365	365	365	365	365
Spearman p-val	0.429	0.813	0.371	0.799	0.473	0.911

Notes: For each regression, we first mark the sample with missing handling. We then re-compute *both* the dependent variable and the estimator ranks *within EA and within the regression sample*, and re-normalize by (1+EA sample size). OLS is then re-run on these re-normalized variables. Standard errors clustered at EA level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Columns: *Borda* = Averaging ranks; *Norm. Borda* = Averaging normalized ranks; *Votes NT/T* = Pairwise majority voting w/o and with test; *Diff NT/T* = Pairwise rank differences w/o and with test.

Table D2: Correlation between estimated orderings and the Food exp per ca of material welfare

	(1) Borda	(2) Norm. Borda	(3) Votes NT	(4) Votes T	(5) Diff NT	(6) Diff T
Rank	0.017	-0.018	0.000	0.078	-0.000	0.004
	(0.052)	(0.048)	(0.050)	(0.049)	(0.047)	(0.047)
Constant	0.492***	0.509***	0.500***	0.461***	0.500***	0.498***
	(0.026)	(0.024)	(0.025)	(0.025)	(0.023)	(0.024)
R2	0.000	0.000	0.000	0.005	0.000	0.000
Observations	348	348	348	348	348	348
Spearman p-val	0.789	0.782	0.971	0.294	0.945	0.946

Notes: For each regression, we first mark the sample with missing handling. We then re-compute *both* the dependent variable and the estimator ranks *within EA and within the regression sample*, and re-normalize by (1+EA sample size). OLS is then re-run on these re-normalized variables. Standard errors clustered at EA level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Columns: *Borda* = Averaging ranks; *Norm. Borda* = Averaging normalized ranks; *Votes NT/T* = Pairwise majority voting w/o and with test; *Diff NT/T* = Pairwise rank differences w/o and with test.

Table D3: Can estimated orderings identify those below the median PPI Index?

	(1) Borda	(2) Norm. Ranks	(3) Votes NT	(4) Votes T	(5) Diff NT	(6) Diff T
Household	-0.011	0.057	0.005	-0.008	0.034	0.026
Ranked Below Median	(0.053)	(0.046)	(0.049)	(0.056)	(0.050)	(0.042)
Constant	0.454***	0.429***	0.448***	0.451***	0.439***	0.441***
	(0.029)	(0.025)	(0.024)	(0.021)	(0.023)	(0.022)
R2	0.000	0.003	0.000	0.000	0.001	0.001
Observations	365	365	365	365	365	365

Notes: For each column, we first mark the regression sample by running an auxiliary OLS of the survey outcome on the constructed rank. Within that sample only, we compute within-EA medians for both the survey outcome and the constructed rank, and define below median indicators (strictly below). We then regress the outcome indicator on the constructed-rank indicator; SEs clustered at the EA level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table D4: Can estimated orderings identify those below the median Weekly Consumption per ca?

	(1) Borda	(2) Norm. Ranks	(3) Votes NT	(4) Votes T	(5) Diff NT	(6) Diff T
Household	0.009	-0.100	-0.064	-0.064	-0.032	-0.043
Ranked Below Median	(0.060)	(0.060)	(0.050)	(0.055)	(0.048)	(0.053)
Constant	0.473***	0.514***	0.498***	0.492***	0.487***	0.491***
	(0.024)	(0.023)	(0.016)	(0.014)	(0.015)	(0.017)
R2	0.000	0.009	0.004	0.003	0.001	0.002
Observations	348	348	348	348	348	348

Notes: For each column, we first mark the regression sample by running an auxiliary OLS of the survey outcome on the constructed rank. Within that sample only, we compute within-EA medians for both the survey outcome and the constructed rank, and define below median indicators (strictly below). We then regress the outcome indicator on the constructed-rank indicator; SEs clustered at the EA level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

E Other empirical applications

In this appendix, we illustrate in detail how different methods perform in a dataset with a large number of ranked alternatives and large variation in reported orderings across observers. Based on the methodological section, we expect pairwise methods without significance adjustment correction to produce estimated aggregate orderings that are less discriminating than rank averaging. We wish to investigate the extent to which adding a testing step to our two pairwise methods is able to increase discrimination. We also expect all methods to produce imprecisely estimated aggregate orderings, based on confidence intervals estimated by randomization inference.

We rely on data from a business plan competition conducted by Ayalew et al. (2024) among small businesses in Ethiopia. The 916 submissions the authors received were evaluated by a panel of 100 loan officers from a variety of local financial institutions that includes both formal banks and micro-finance institutions. Each loan officer was randomly assigned between 50 and 100 randomly business plans for them to score (from 1 to 10) on four criteria. We use the scores given to the first of these criteria, 'overall evaluation', to construct reported orderings for each observer.

There are 916 applicants (alternatives) in total, ranked by 85 observers. On average, each observer ranks 49.9 applicants, with a minimum of 5 and a maximum of 79. Two-third of observers rank between 40 and 60 applicants. On average, an applicant is ranked by 4.6 observers, with a minimum of 2 and a maximum of 8. Three quarter of applicants are ranked by between 4 and 6 observers. As a result, the proportion of ranked pairs is low, increasing the possible impact of reporting errors on observed pairwise ranks (less averaging) as well as increasing the likelihood that misranked pairs contaminate the reconstruction of missing pairs through the matrix multiplication process. These are conditions under which pairwise methods are unlikely to be able to compete with rank averaging.

We now examine the estimated aggregate orderings obtained using the four pairwise estimators, the Borda counts, the averaged normalized ranks, as well as the averaged scores themselves, which are expected to produce more precise orderings by virtue of using more cardinal information. We find a high correlation of 0.93 between the orderings estimated using the two rank averaging methods, and a slightly lower correlation with the ordering derived using averaged scores (0.88 for normalized ranks and 0.81 for Borda counts). Pairwise methods without testing perform particularly poorly in this dataset—especially the pairwise majority voting method, which ties all alternatives and thus has zero correlation with the ordering obtained from averaging scores. Significance adjustment correction improves matters, with pairwise rank differences with testing achieving a 0.71 correlation with scores and correlations of 0.76 and 0.80 with Borda and normalize ranks, respectively. Pairwise majority voting with testing falls just short of 0.5 in all cases.

Table E1 compares the performance of all six estimators, plus average scoring. We notice immediately that the two pairwise methods without testing end up bunching most if not all alternatives into a single rank. This lack of discriminating power is also very precise: it is extremely robust across all bootstrapped simulations. In contrast, the two rank averaging methods perform well in terms of discriminating power. The ordering obtained using averaged normalized rank approaches the minimum feasible discrimination level of 0.1% attainable with 916 alternatives. Interestingly, this is better than using the ordering

obtained using averaged scores. This is probably because of bunching of scores on a few values: a score of 8 may have a different discriminating value if other alternatives are given similarly high scores, or if they stand out—a piece of information that normalized ranks capture better. Of the pairwise methods, pairwise rank differences with testing performs the best, reaching a respectable discriminating power of 1.8%.

Table E1: Comparison of Estimators in the Ethiopia Dataset

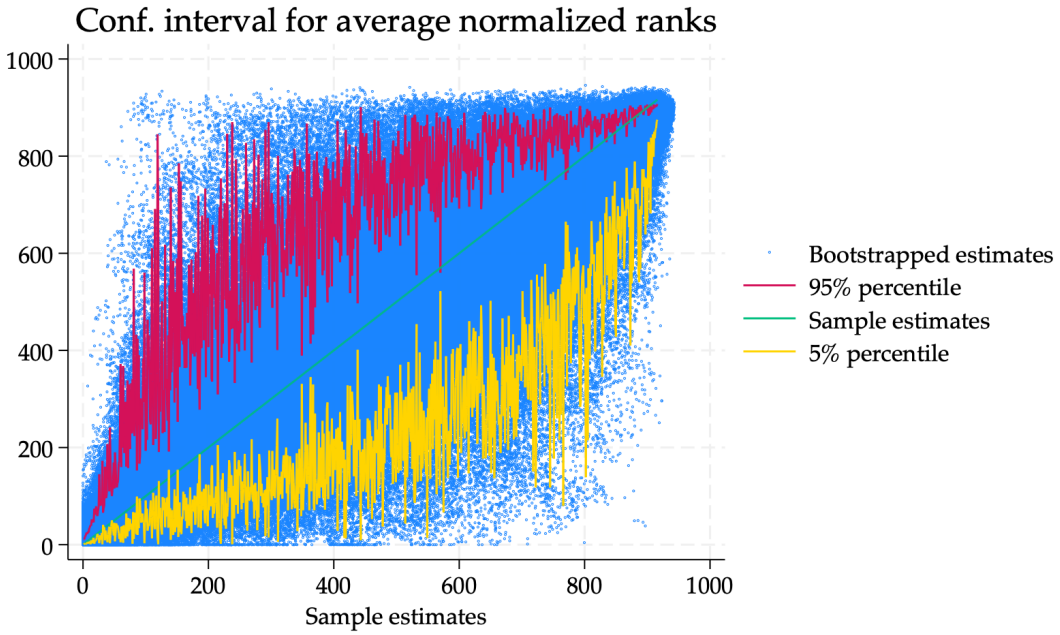
	Lack of:		
	Discriminating power	Precision	
	HHI		
	(1)	(2)	(3)
Averaging ranks (Borda count)	0.2%	65.1%	99.5%
Averaging normalized ranks	0.1%	44.7%	99.5%
Pairwise majority voting, no test	100.0%	1.6%	100.0%
Pairwise majority voting, with test	37.9%	62.6%	99.9%
Pairwise rank differences, no test	99.1%	1.6%	95.6%
Pairwise rank differences, with test	1.8%	67.5%	99.7%

Notes: The Herfindahl–Hirschman Index (HHI) of concentration is calculated separately. With 916 alternatives, the minimum value it can take is 0.1%. Precision measure (1) is the percentage of the maximum inter-rank distance covered by the 95% confidence interval, averaged over all estimated ranks. Precision measure (2) is the percentage of simulated ranks that fall outside the interval [–1,+1] from the estimated rank, averaged over all estimated ranks.

High discriminating power does not imply precision, however: even for the best ranking method, averaged normalized ranks, Monte Carlo simulated ranks cover 44.7% of the feasible range of ranks which, to recall, span from 1 to 916 in this case (see Figure E1). In the last column, we show the proportion of bootstrapped ranks that fall outside a $[-50, +50]$ range on either side of the estimated rank. The best method still have a 63.9% of simulated ranks falling outside of that large interval. This means that, even using the best available method for this case, the estimated aggregate ordering would be noticeably different if another sample of observers had been used. This is important because it implies that, if observers are seen as interchangeable because they are assumed to have common preferences, the estimated aggregate ordering should be regarded as having been estimated with considerable measurement error, something that would affect its predictive power in explanatory regressions, for instance. On the other hand, if observers have different latent orderings, the purpose of averaging their rankings is to obtain a representative ordering—in which case our method for assessing the precision of estimated aggregate orderings is incorrect: observers are not interchangeable.

Very similar results are obtained using the golf tournament data of Guryan et al. (2009): large-scale contamination of the matrix multiplication process in the pairwise approach, resulting in a large number of ties; more rank discrimination using rank averaging, but large confidence intervals for the estimated aggregate ordering.

Figure E1: Confidence Intervals for selected estimators in the Ethiopia data



Notes: Obtained by randomizing observers with replacement.

Application of the pairwise method to rural household data of Trachtman et al. (2026) yields better results. But even here the larger number of 30 alternatives causes an increase in the number of estimated ties when using the pairwise method. Other findings confirm earlier results: averaging normalized ranks dominates Borda counts; the pairwise rank difference method is more discriminating than pairwise majority voting; and pairwise methods with testing produce much better results.

F Côte d'Ivoire Application: Data collection protocol

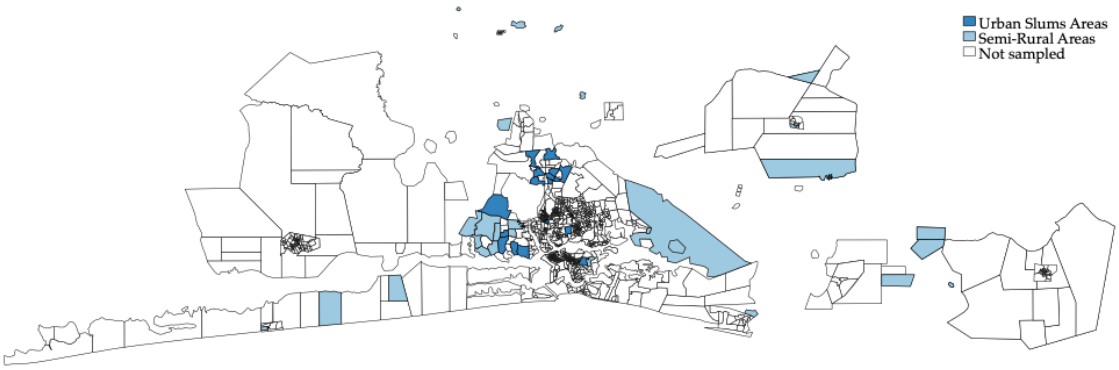
This appendix provides details on the collection of reported rankings by survey respondents living in various enumeration areas of the Côte d'Ivoire city of Abidjan and its periphery.

F.1 Data Collection and Sampling Frame

Summary: The ranking exercise was nested within the African Urban Development Research Initiative (AUDRI), which aimed to build representative data on urban and peri-urban populations in Greater Abidjan. For this study, we selected 34 enumeration areas (EAs)—20 urban "slum" EAs and 14 peri-urban areas—based on the national census definitions. In each EA, enumerators listed 14 consecutive households to define a "neighborhood," then administered three survey instruments: a short survey (N=207), a detailed individual survey (N=119), and a ranking survey (N=507).

This study was conducted under the African Urban Development Research Initiative (AUDRI), a collaboration between Stanford University and local partners in Côte d'Ivoire. The AUDRI sampling frame is based on the 2014 national enumeration areas (EAs) defined by Côte d'Ivoire's National Statistical Institute (INS). Urban EAs each include roughly 200 households; peri-urban EAs correspond to entire villages of varying size. AUDRI oversampled peri-urban areas on the urban fringe, resulting in a final frame of 706 EAs, of which 84 were classified as peri-urban and 622 as urban. Figure F1 shows the sampled areas on a district map of Greater Abidjan.

Figure F1: Areas sampled by the AUDRI study



Notes: Enumeration areas selected for the ranking study are indicated in blue.

For the peer ranking component of the study, we randomly selected 20 urban slum EAs and 20 peri-urban EAs from the AUDRI sample. Due to local refusals, we ultimately conducted the ranking survey in 34 areas (20 urban and 14 peri-urban).[28]

In each EA, enumerators followed a "right-hand rule" to list 14 consecutive dwellings beginning at the EA centroid. These listings were done in July–August 2019 and formed

[28]Six villages could not be reached by the team of surveyors because the village chief did not allow our enumerators access to the village.

the basis of our neighborhood definition. For logistical reasons, fewer than 14 households were listed in 8 EAs. The listing survey collected information from a randomly selected adult about household composition and asset ownership. Of the 476 listed dwellings, 207 households completed the listing survey. Of these, N=119 were randomly selected for a more detailed Individual Survey conducted between December 2019 and March 2020. This survey collected four hours of data per household, including modules on labor, transportation, health, and public services. To ensure completeness, enumerators made repeated visits to dwellings where no adult was home on the first attempt. In four EAs, dwellings that were closed on the first visit were mistakenly skipped due to field miscommunication. sampled households may be slightly more dispersed geographically.

The final analytical sample includes 207 households with complete listing data, of whom 119 also completed the Individual Survey. These households form the target population for peer rankings and survey-based welfare comparisons.

F.2 Ranking Survey Protocol

Summary: In March 2020, observers were asked to rank 5 to 14 neighboring households in terms of material well-being. Half were randomly assigned to also rank their own household. Enumerators used pre-coded household lists to ensure observer-household match quality. Four types of observers participated: (i) detailed-survey households (N=119), (ii) listed households (N=88), (iii) additional nearby residents (N=230), and (iv) key informants (N=70), primarily local shopkeepers.

The ranking survey was administered in March 2020 to 507 observers across the 34 ranking areas. Each observer was presented with a list of 5 to 14 nearby households (pre-coded from the listing exercise) and asked to rank them in order of material well-being. Respondents were free to include any immediate neighbor they knew by name and enumerators were instructed to identify—and confirm with the respondent—the names of the target households so that they could be matched with the data we collected on them. Enumerators received detailed training to ensure accurate matching of household names. Timestamps data from the SurveyCTO module indicate a median duration of 13 minutes for the ranking task. We did attempt to include the names of households other than the target households, in the hope of improving the quality of reconstructed rankings. But there were very few of them and we cannot be sure these names identify the same households. To avoid identification errors, we drop from the ranking analysis those households that could not be matched by the enumerator with a target household.

Observers were not informed that rankings would be used to allocate benefits. In practice, they were not. Half of observers were randomly assigned to include their own household in the ranking.

The precise wording of the question was:

"Maintenant que vous avez identifié les ménages que vous reconnaissez [dans notre liste], pourriez-vous me donner un classement de ces ménages, du plus pauvre au plus riche selon vous."

(*Now that you have identified the households you recognize [from our list], please rank these households from the poorest to the richest in your opinion.*)

Observer Types We classify observers into four groups (see Table F1):

- **Type A1: Individual Survey observers** (N=119) – selected from the listing sample, completed the full Individual Survey, and then the ranking survey.
- **Type A2: Listed-only observers** (N=88) – completed the short survey but not the full Individual Survey.
- **Type B: Additional nearby residents** (N=230) – initially absent during listing, but identified and interviewed during the ranking phase.
- **Type C: Key informants** (N=70) – identified as traders or shopkeepers in the vicinity, surveyed for their external perspective on neighborhood's material welfare.

To minimize recall or identification errors, only rankings involving matched households are used in the analysis. Households from Type A1-A2 and B reside in consecutive dwellings and are close neighbors by design.

Table F1: Breakdown of the Sample of Observers by Origin

Sample origin	**#**	**%**	**% of Household Head**	**% of Women**
Individual survey (A1)	119	23.47 %	47.06%	56.30%
Listing survey only (A2)	88	17.36%	%	69.32%
Selected on the spot (B)	230	45.36%	46.52%	46.96%
Informants (C)	70	13.81%	25.71%	71.43%
Total	**507**	**100%**	**42.41%**	**56.41%**

Notes: In EAs selected for the ranking exercise, we sought to interview all respondents to the Individual survey (December 2019 to March 2020) and to the Listing survey (July to August 2019). The Table shows those who could be found and surveyed for the ranking exercise. A number of additional households were recruited on the spot as observers to increase sample size. Informants were also recruited on the spot among traders and shopkeepers operating in the area.

F.3 Description of the sample

We define a sample of up to 14 target households per neighborhood—476 in total—for whom we collected survey data. Observers received a list of nearby households but could rank any households in their EA. Across the 34 EAs, this resulted in 3094 possible household pairs. We also surveyed 1–3 key informants per EA, for a total of 507 observers (14.9 per EA on average).

Each observer could rank up to 14 households (15 for those who self-ranked). If all observers had provided complete rankings, we would have 46,137 pairwise comparisons—up to 15 reports per pair. Such overlap would allow precise aggregation via average ranks (as in Alatas et al. (2012)) or our pairwise method. However, urban anonymity limited coverage: 17% of observers had lived in the area for under a year, and only 6% were born there. We compensated by doubling the observer-to-target ratio relative to Alatas et al. (2012), and we selected households in close proximity to improve familiarity. Still, in practice, many observers declined to rank even their immediate neighbors.

Table F2: Summary Statistics Across EAs

	(1) This Paper
# of intended targets	14.00
	(0.00)
# of observers	8.91
	(4.04)
Coverage overlap between observers >= 2	0.31
	(0.23)
Coverage overlap between observers = 1	0.15
	(0.16)
Coverage overlap between observers = 0	0.54
	(0.23)
Full Agreement on the Pairs i-j	0.24
	(0.44)

Notes: The extent of overlap in coverage between observers states that the share of target households that were ranked by at least 2 observers, 1, or none. Full agreement across observers' report, among all reported pairwise ranks by individual observers, the share where the proportion of responses that state that i < j is equal to 0 or 1.

Ultimately, we collected only 1,820 distinct pairwise rankings—just 3.9% of the theoretical maximum (46,137). These covered 837 household pairs (27% of all possible pairs) and 364 of the 476 target households—leaving 24% entirely unranked. For 52.8% of ranked pairs, we received only one report, making it infeasible to average out noise. Observers ranked an average of 3.6 household pairs (2.5 excluding self-ranks), with a maximum of 36. This is much lower than in Alatas et al. (2012), where nearly all households are ranked five times or more. Among multi-ranked pairs, full agreement occurs in just 24% of cases (see Table F2).

F.4 Material Welfare Benchmarks

Summary: The goal of the ranking exercise is to compare peer-reported rankings to survey-based proxies for material well-being. Since there is no universally agreed-upon measure of poverty, and each available proxy has limitations, we collected multiple indicators to approximate household material welfare. These include measures of recent consumption as well as indices designed to reflect longer-run poverty status. We emphasize that none of these proxies is a perfect or complete measure of poverty; rather, they capture different dimensions of material welfare with varying degrees of measurement error. For all measures, lower values indicate *greater* poverty.

Proxy Means Test (PMT) The PMT index is designed to estimate economic status from observable household characteristics. We follow the methodology used by the Government

of Côte d'Ivoire, which developed PMT weights in 2015 using LSMS survey data. These weights come from a regression of *log food consumption per capita* on roughly 25 asset and housing variables. We apply the same weights to our survey data to compute PMT scores. Since PMT coefficients differ by urban/rural classification, we treat peri-urban areas in our sample as rural, in line with government practice. In Table F3, we replicate the original PMT regression using our sample and confirm similar predictive power and coefficient magnitudes. Since the information required to compute the PMT was collected in the short survey, it is available for households initially listed, irrespective of whether they were interviewed in the full survey or not.

Table F3: Estimated Fit of the PMT and PPI indices w.r.t. log(consumption per capita)

	PMT Index					PPI Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Rural - Gov	Urban - Gov	Tot - Gov	Tot - Ranking	Tot - AUDRI	Tot - Ranking	Tot - AUDRI
R^2	0.497	0.612	0.568	0.563	0.491	0.484	0.446
Observations	7,076	5,748	12,773	193	2,871	493	2,666

Notes: The table reports the R^2 and the number of observations from the regressions run by the government of Côte d'Ivoire to build their PMT score (columns 1, 2, 3). The numbers were shared to us by the CNAM in Côte d'Ivoire. The government regressed log(food expenditure per capita) on the variables used to build the PMT score. Column 4 reports the R^2 from the same regression run on the households involved in the ranking exercise while Column (5) includes the full AUDRI sample. Column (6) reports the fit from the PPI regression, i.e., regressing log(food consumption per capita) on the variables used to build the PPI index. Column (7) reports the latter PPI regression on the full AUDRI sample. Note that the sample size is not exactly the same between columns (5) and (7) due to differential missing patterns between variables used in the PMT vs. the PPI score.

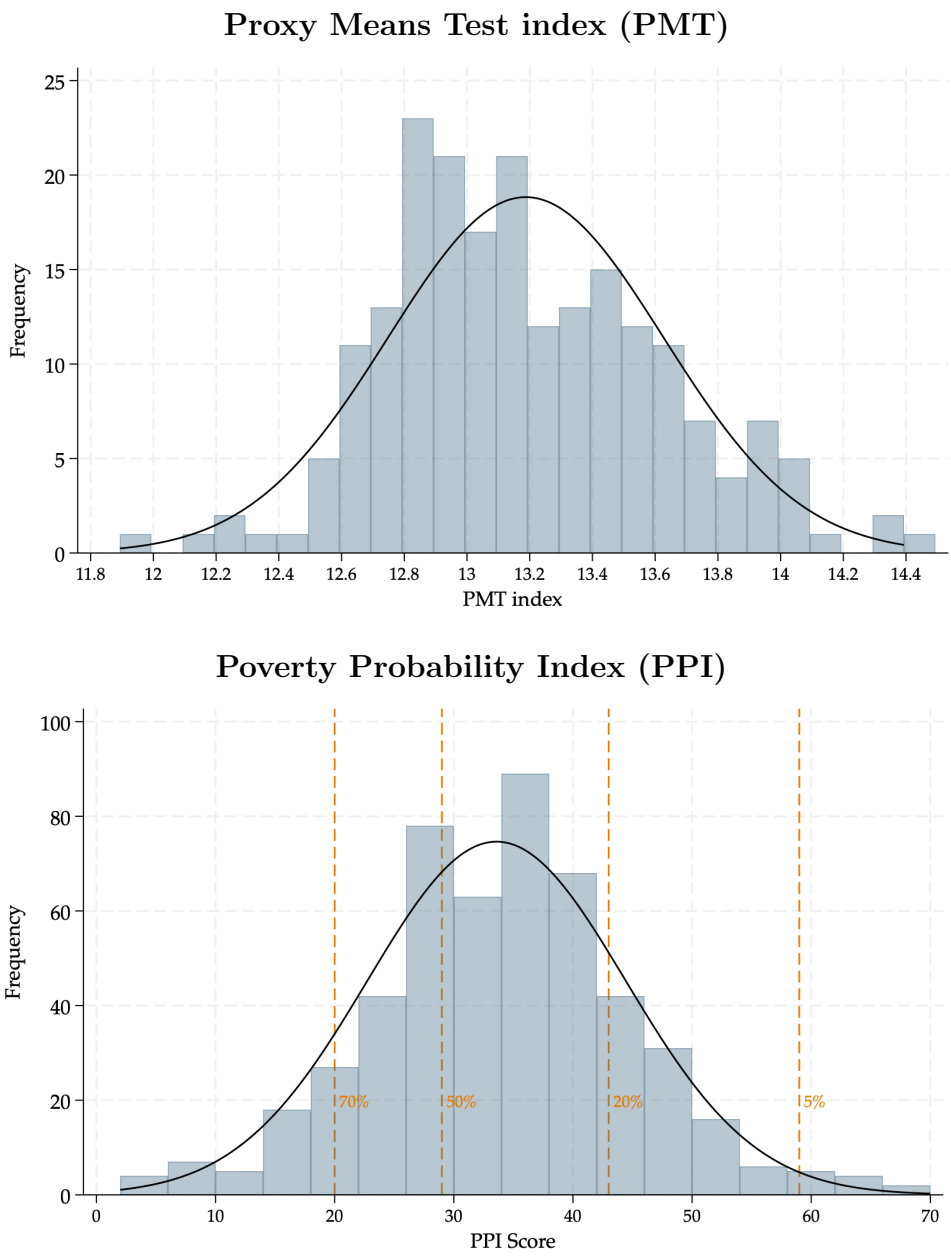
Poverty Probability Index (PPI) The PPI is an alternative index introduced by Innovations for Poverty Action (IPA) in 2018, also using Côte d'Ivoire's 2015 LSMS data. It estimates the probability that a household falls below the national poverty line, based on responses to ten standardized questions on location, housing, and household composition (Table F4). Unlike consumption data, PPI is designed to reflect longer-run welfare and be less sensitive to short-term shocks. We construct PPI scores for all households with sufficient data. Figure F2 shows the distributions of PPI and PMT across our sample, which reflect considerable variation in predicted poverty status—even in urban neighborhoods. As expected, households in and around the capital city have slightly higher PPI scores than the national average, reflecting both higher cost of living and infrastructure coverage.

Table F4: PPI Scorecard for the Côte d'Ivoire 2015 National Poverty Line

Indicators	**Responses**	**Points**
1. In which district does this household reside?	A. Abidjan	7
	B. Yamoussoukro	5
	C. Bas-Sassandra	9
	D. Comoé	4
	E. Denguélé	0
	F. Gôh-Djiboua	3
	G. Lacs	3
	H. Lagunes	2
	I. Montagnes	5
	J. Marahoué	0
	K. Savanes	2
	L. Vallée du Bandama	2
	M. Woroba	4
	N. Zanzan	4
2. How many members does the household have?	A. Three or less	17
	B. Four or more	0
3. What is the highest educational level that the household head has completed?	A. None	0
	B. Primary	4
	C. Secondary	5
	D. Higher	12
4. Did all children aged 6 to 16 attend school this school year?	A. There are no children aged 6 to 16	11
	B. All children aged 6 to 16 attended school this year	7
	C. At least one child aged 6 to 16 did not attend school this year	0
5. What is the mode of water supply?	A. Tap water in the dwelling	10
	B. Tap water in the yard	4
	C. Tap water outside of the property	4
	D. Well in the yard	1
	E. Public well	2
	F. Village pump	2
	G. Surface water (creek, river, etc.) or other	0
6. What type of toilet do you use?	A. W-C inside	7
	B. W-C outside	6
	C. Latrines in the yard	5
	D. Latrines out of the yard	5
	E. In nature (no toilet) or other	0
7. Where do you take your shower?	A. Outside	0
	B. Rudimentary shower	3
	C. Bathroom	9
	D. Other	1
8. Did the household own a moped, car or van in good working order in the last 3 months?	A. The household owns a car or van	15
	B. The household owns a moped and does not own a car or van	9
	C. None	0
9. Did the household own a fan in good working order in the last 3 months?	A. Yes	6
	B. No	0
10. Did the household own a bed in good working order in the last 3 months?	A. Yes	4
	B. No	0
PPI Index		Sum of points

Notes: The points provided here are those for the National Poverty Line.

Figure F2: Sample distribution of the PMT and PPI indices in the AUDRI Abidjan sample



Notes: We plot the distribution of the PPI (poverty probability index) and PMT (proxy means test index). The top Figure shows the PMT index developed by the Ivorian Government. In the bottom Figure, we show the PPI in our sample, using weights estimated by Innovations for Poverty Action (IPA) in April 2018 on the basis of Côte d'Ivoire's 2015 Enquête sur le Niveau de Vie des Ménages (Household Living Standard Measurement Survey). The "Poverty Likelihood", i.e., the probability to be below the National Poverty Line, is indicated in orange. The two indices are described in more details in Section F.4.

Consumption expenditures We construct three consumption-related measures from household survey data:

1. *Food consumption in the last week*: We administer a standard recall module to collect information on household consumption of cereals, pulses, spices, dairy, meat, bread/pasta, vegetables, fruits, beverages, alcohol, and other items. While the food consumption module was administered about a month earlier for observers in the individual survey, we pool all available data for consistency.

2. *Conspicuous or social consumption in the last month*: This includes non-food spending on communication, beauty products, entertainment (e.g., concerts, bars), and charitable contributions.

3. *Durable spending over the last 12 months*: We capture major expenses on clothing, footwear, school fees, and furniture.

Missing responses (typically 1–3% for any given item) are imputed using the enumeration area (EA) mean to preserve sample size. We use per capita versions of these consumption aggregates as our baseline, but acknowledge that this ignores household composition and economies of scale. In robustness checks, we construct equivalized consumption measures using a standard equivalence scale.[29]

Table F5 reports summary statistics for all measures, separately for urban and peri-urban areas. While all measures are positively correlated—including PPI, PMT, and the consumption aggregates—their correlations are moderate, suggesting that each captures distinct aspects of material welfare. In line with Trachtman et al. (2026), we examine all three as plausible benchmarks when evaluating the informational content of peer rankings.

F.5 Perceptions of Material Welfare

To complement the peer rankings, we asked observers to reflect on material welfare in their own terms. The following questions were included in the ranking survey: (1) Do you consider your own household poor? (2) Do you consider your household poorer than most other households in the neighborhood? (3) Do you think others in the neighborhood view your household as poor?

Summary statistics from these questions are presented in Table F6. Responses reveal systematic divergences between self-perceived and socially perceived poverty. While 29% of observers describe themselves as poor, only 21% believe they are poorer than their neighbors, and just 20% think others see them as poor. Among those who self-identify as poor, 53% believe others would disagree.

We also asked observers to describe the criteria they used to rank households. The most frequently mentioned indicators were "Food insecurity or hunger" (80%); "Unmet health needs" (43%), "Occupation of household head" (49%); "Known financial struggles" (49%)

Finally, we collected data on the nature of social interactions between observers and their neighbors. More than half reported regular visits, and around half had sought health

[29]Specifically, we use the OECD-modified equivalence scale: 1.0 for the first adult, 0.5 for each additional adult, and 0.3 per child. Log consumption per adult equivalent is then computed and used in parallel regressions. Results are qualitatively similar, and do not meaningfully affect the low predictive power of peer rankings.

Table F5: Summary Statistics - Measures of Material Welfare

	(1) Urban	(2) Rural
Consumption Expenditures		
Value of food expenditure in the last week	15.43	13.17
	(8.12)	(7.20)
Value of conspicuous expenditures in the last month	2.47	2.59
	(3.29)	(2.74)
- Communication expenditures	1.09	1.00
	(1.25)	(1.39)
- Entertainment (concert, bar, cinema, games) expenditures	0.22	0.29
	(1.55)	(0.91)
- Beauty products/hairdresser expenditures	0.48	0.75
	(0.79)	(1.40)
- Charitable expenditures	0.67	0.55
	(2.38)	(1.18)
Spending on durables in the last 12 months	2.34	2.10
	(3.00)	(2.38)
- Clothes/shoes HH expenditures	1.10	1.34
	(1.09)	(1.48)
- Furniture HH expenditures	0.39	0.27
	(1.16)	(0.84)
- School fees HH expenditures	0.89	0.52
	(2.80)	(1.35)
Value of food expenditure in the last week per capita	3.76	3.74
	(2.89)	(3.39)
Spending on durables in the last 12 months per capita	0.53	0.61
	(0.67)	(1.46)
Indexes		
PPI Index	37.13	28.70
	(9.55)	(10.64)
Score PMT	13.23	13.13
	(0.43)	(0.44)
Other variables		
HH's head unemployed or inactive	0.20	0.15
	(0.40)	(0.36)
# of mobile phones per capita	0.81	0.82
	(0.47)	(0.48)
Observations	294	213

Notes: Consumption expenditures are in 1,000 FCFA.

Table F6: How do Observers Think About Poverty? Summary Statistics from Survey Data

	(1) Share of observers
Uncertain about their ranking	0.09
	(0.29)
Perceptions of own poverty	
Consider their household to be poor	0.29
	(0.45)
Consider their household to be poorer than neighbors	0.21
	(0.41)
Think that other households consider their household to be poor	0.21
	(0.41)
Criteria used to classify	
Household expressed their financial problems	0.49
	(0.50)
Household members' health	0.14
	(0.35)
Household head's occupation	0.49
	(0.50)
Households' daily number of meals	0.19
	(0.39)
Household children's school enrollment	0.07
	(0.26)
Observers' own definition of poverty	
Food deprivations	0.80
	(0.40)
No decent housing	0.31
	(0.46)
Unresolved health problems	0.43
	(0.49)
No proper toilet/bathroom	0.16
	(0.37)
Knowledge about neighbors	
# of neighbors listed in total	5.75
	(1.47)
% of neighbors they regularly visit	0.56
	(0.38)
% of neighbors receive health/money advice from	0.44
	(0.39)
% of neighbors they'd ask money from	0.38
	(0.38)
Observations	507

Notes: Survey data collected early March 2020. Definition of poverty manually entered by the enumerators and re-classified by the research team.

or financial advice from the neighbors they ranked—supporting the idea that peer knowledge exists but may be partial or filtered by social proximity.