Annals of Operations Research (2023) 325:939–949
https://doi.org/10.1007/s10479-022-05026-2

ORIGINAL RESEARCH

External validity of multi-criteria preference data obtained from non-random sampling: measuring cohesiveness within and between groups

Saeideh Babashahi[1] · **Paul Hansen**[2] · **Ronald Peeters**[2]

Accepted: 12 October 2022 / Published online: 2 November 2022

Abstract
An important component of multi-criteria decision analysis (MCDA) in the public sector is the elicitation and aggregation of preference data collected via surveys into the relative importance of the criteria for the decision at hand. These aggregated preference data, usually in the form of mean weights on the criteria, are intended to represent the preferences of the relevant population overall. However, random sampling is often not feasible for public-sector MCDA for logistical reasons, including the expense involved in identifying and recruiting participants. Instead, non-random sampling methods such as convenience, purposive or snowball sampling are widely used. Nonetheless, provided the preference data collected are sufficiently 'cohesive' in terms of the extent to which the weights of the individuals belonging to the various exogenously defined groups in the sample are similar, non-random sampling can still produce externally valid aggregate preference data. We explain a method for measuring cohesiveness using the Kemeny and Hellinger distance measures, which involve measuring the 'distance' of participants' weights (and the corresponding rankings of the criteria) from each other, *within* and *between* the groups respectively. As an illustration, these distance measures are applied to data from a MCDA to rank non-communicable diseases according to their overall burden to society. We conclude that the method is useful for evaluating the external validity of preference data obtained from non-random sampling.

Keywords Multi-criteria decision analysis (MCDA) · Non-random sampling · External validity · Distance measure · Preference cohesiveness · Cluster analysis

✉ Ronald Peeters
ronald.peeters@otago.ac.nz

Saeideh Babashahi
s.babashahi@bsms.ac.uk

Paul Hansen
paul.hansen@otago.ac.nz

[1] Department of Global Health and Infection, Brighton and Sussex Medical School, University of Sussex, Sussex, UK

[2] Department of Economics, University of Otago, 9054 Dunedin, New Zealand

1 Introduction

Multi-criteria decision analysis (MCDA) is increasingly used to support decision-making in the public sector, including in publicly-funded health systems (the focus of the illustrative application later in the article). For reviews of MCDA in the health sector, see Thokala et al. (2016), Marsh et al. (2017), Baltussen et al. (2019) and Hansen and Devlin (2019). An important component of public-sector MCDA is the elicitation and aggregation of preference data concerning the relative importance of the criteria for the decision at hand. These aggregated preference data, usually in the form of mean weights on the criteria, are intended to represent the preferences of the relevant population overall and are often elicited using surveys.

When recruiting survey participants, random sampling (where each individual in the population of interest has an equal chance of being selected) is generally considered to be the gold standard of sampling methods because of its unbiasedness and the external validity of its results (where, to be clear, by "external validity" we mean, as is conventional, that the preferences data for the sample are representative of, and hence generalizable for, the relevant population overall). However, random sampling is often not feasible for public-sector MCDA for logistical reasons, including the expense involved in identifying and recruiting participants. Instead, non-random sampling methods such as convenience, purposive or snowball sampling are widely used (Etikan et al., 2016; Goodman, 1961).

Nonetheless, provided the preference data collected are sufficiently 'cohesive' in terms of the extent to which the weights of the individuals belonging to the various exogenously defined groups in the sample are similar, non-random sampling can still produce externally valid aggregate preference data. Such exogenously defined groups could be distinguished in terms of individuals' observable socio-demographic characteristics or, depending on the application, their stakeholder type–e.g. in the context of the health sector, patients versus healthcare providers, researchers or policy-makers, etc.

With the objective of evaluating whether preference data obtained from non-random sampling is likely to be externally valid (or the opposite), this article addresses the following methodological question. When non-randomly sampled participants differ with respect to their membership of exogenously defined groups, how do we ex-post evaluate the cohesiveness of their elicited weights?

The fundamental concepts for thinking about this question are illustrated in Fig. 1 via two simple cases corresponding to the left and right plots. Each plot (case) represents the weights on three criteria for 75 individuals who each belong to one of three groups of 25, represented by the blue circles, red triangles and green squares.

In the left case, the within-group variation is similar across the three groups, but the groups themselves are clearly different. In contrast, in the right case, the within-group variation for each group is larger than in the left case, but the three groups are not obviously different. Thus, the weights of the three groups are more cohesive in the right case than in the left case. If a sample was being drawn from each population, the sampling method – random versus non-random – would be less of an issue for the right than for the left case: non-random sampling would be more likely to produce externally valid preference data for the right than for the left case.

In this article, we explain and illustrate a method for ex-post evaluating the external validity of preference data (in the form of weights and the corresponding rankings of the criteria) obtained from non-random sampling based on measuring the cohesiveness of the data from the various exogenously defined groups in the sample. As alluded to in the illustration above, fundamental to measuring the cohesiveness of participants' weights (or rankings) is their

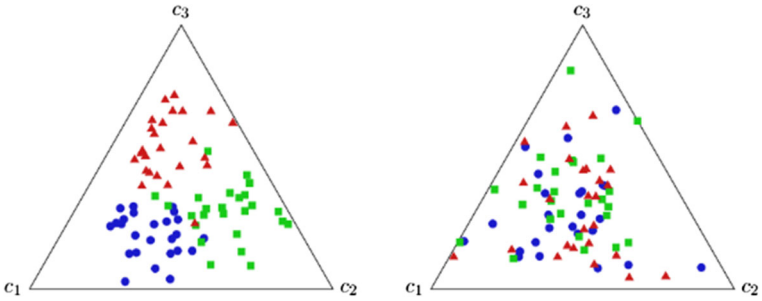

Fig. 1 Graphical illustration of two cases of weights (non-normalized) on three criteria (axes) for 25 people from three stakeholder groups

'distance' from each other, *within* and *between* the groups respectively. In the next section, we develop specific distance measures and suggest a bootstrapping method that allows for proper statistical testing.

In Sect. 3, as an illustration, we apply the developed techniques to existing preference data from a MCDA to rank non-communicable diseases according to their overall burden to society (Babashahi et al., 2021). The 476 participants were sampled in a non-random manner. Yet, there are three clearly distinct types of participants: patients and general public (group P), researchers and policy-makers (group R), and healthcare service providers (group S). In brief, our results may be summarized here as follows.

Although the three groups ranked the underlying criteria identically, using our method we can conclude that based on the weights, differences in preferences between members of group P and members of group S are not significantly different to the differences in preferences among members in group S. For all other pairs of groups, we find significant differences. Given that effect sizes are modest, we conclude that non-random sampling may not have undermined external validity. If we had had statistically significant differences and substantial effect sizes, we would have concluded the opposite regarding external validity and have recommended the weights be re-estimated used a randomly drawn sample.

The article closes with our short discussion and conclusion. We provide substantive arguments for some of the choices underpinning the measures that we develop in Sect. 2 and offer possible (theoretically founded) tweaks to them. We also acknowledge that we are not the first to research external validity issues caused by non-random sampling and compare our method with the rich literature of other methods available (including cluster analysis). Overall, we believe our method is complementary to those methods.

2 Methods

The method for quantifying the cohesiveness of participants' preference data within and between groups respectively is based on the Kemeny (1959) and Hellinger (1909) distance measures, which are explained below after some basic notation is introduced.

Let there be sets of individuals $N = \{1, \ldots, n\}$ and criteria $M = \{1, \ldots, m\}$, with each individual having preferences over the criteria. The preferences of each individual i are represented ordinally by a ranking vector $r^i = \left(r^i_k\right)_{k \in M}$, with $r^i_k \in M$ and lower numbers

indicating more preferred criteria, and cardinally by a weight vector $w^i = \left(w_k^i\right)_{k \in M}$, with $w_k^i \geq 0$, $\sum_{k \in M} w_k^i = 1$ and higher weights indicating more preferred criteria.

For each pair of individuals, i and j, we can use both their rankings and weights to quantify the extent of the dissimilarity (or, conversely, similarity) of the individuals' preferences. One metric based on rankings is the normalized Kemeny distance (Kemeny, 1959), defined as:

$$d^r(i, j) = \frac{1}{2(m-1)m} \sum_{k \in M} \sum_{l \in M} \left| sign\left(r_k^i - r_l^i\right) - sign\left(r_k^j - r_l^j\right) \right|$$

This normalized Kemeny distance is increasing in the minimum number of interchanges of two adjacent elements (ranks) required to transform one person's ranking into the other person's ranking (i.e. increase their similarity), and has a minimum of 0 if the two rankings are identical (perfectly similar) and a maximum of 1 if they are opposite (perfectly dissimilar). One metric based on weights, and that we will also use in this article, is the Hellinger distance (Hellinger, 1909), defined as:

$$d^w(i, j) = \sqrt{\frac{1}{2} \sum_{k \in M} \left(\sqrt{w_k^i} - \sqrt{w_k^j} \right)^2}$$

The Hellinger distance has been developed to quantify the dissimilarity between two probability distributions (in our context, two weight vectors) and has a minimum of 0 if the two distributions are identical, and a maximum of 1 if the supports of the two probability distributions are disjoint.

Let now the individuals in N be (exogenously) partitioned in different groups. For a given group G, we define the within-group dissimilarity as the mean distance between all possible pairs in this group:

$$d^z(G) = \frac{1}{(|G|-1)|G|} \sum_{i \in G} \sum_{j \in G \setminus \{i\}} d^z(i, j)$$

where $z \in \{r, w\}$, depending on whether the dissimilarity is based on rankings or weights.[1] Further, for a pair of groups, G and H, we define the between-group dissimilarity as the mean distance between all possible pairs of members from both groups:

$$d^z(G, H) = \frac{1}{|G||H|} \sum_{i \in G} \sum_{j \in H} d^z(i, j)$$

where again z refers either to rankings or weights. The within-group dissimilarity represents the coherence of the preferences of the individuals within the group, whereas the between-group dissimilarity represents the coherence of the individuals' preferences between the two groups.

For similar reasons as underly the infinite monkey theorem, the probability that any pair of individuals within a group has an identical ranking over criteria is increasing in the size of the group. As a result, both dissimilarity metrics are sensitive to group size – which, in particular, affects metrics for rankings, and the extent to which it affects metrics for weights depends on the resolution at which the weights can be specified. One way to correct for such

[1] Though there is no gold standard, smaller values – indicating smaller distances between individuals' preference data – indicate more cohesiveness. The within- and between-group cohesiveness reflects the extent of agreement between participants: a higher level of agreement corresponds to more similar preferences.

potential biases – and the route we take because it also opens the road to statistical testing – is to use bootstrapping methods.

For bootstrapping purposes, we fix a sample size s conveniently below the minimum size of all the groups: $s < min_G|G|$. By frequent (t times) sampling s individuals from the respective groups, we create distributions of our within- and between-group dissimilarities: $f^r(G)$, $f^w(G)$, $f^r(G,H)$ and $f^w(G,H)$. We can use these distributions in three ways: (1) to perform (in a fair manner) statistical inferences regarding differences in within-group dissimilarities across groups, (2) to compare between-group differences between pairs of pairs of groups, and (3) to compare within- and between-group dissimilarities (e.g. the within-group dissimilarity of group G and the between-group dissimilarity of groups G and H).

If within-group dissimilarities are not statistically significantly different from the between-group dissimilarities (like the right plot in Fig. 1), we may conclude that the weights of the groups are cohesive, and so preference data obtained from non-random sampling is likely to be externally valid; in other words, any bias due to non-representative sampling is likely to be modest. If, instead, between-group dissimilarities are statistically significantly larger than within-group distances (like the left plot in Fig. 1), the sampling method used is an issue and preference data obtained from non-random sampling may not be externally valid.

3 Application to a MCDA for ranking non-communicable diseases

As an illustration, we apply the method explained above to data from a MCDA performed in New Zealand for ranking 19 non-communicable diseases (NCDs, e.g. arthritis, cancer and depression) according to their overall burden to society, with the ranking available to support decision-making about health research funding; see Babashahi et al. (2021) for details.

In brief, an online survey to elicit weights on five criteria reflecting NCDs' burden to society was administered in late 2017. Participants ($N = 476$) were invited using convenience and purposive sampling via the researchers' personal and professional networks and relevant New Zealand health organizations, with 'snowballing' whereby participants were asked to forward the survey link to other eligible and interested people. The survey implemented the PAPRIKA method – an acronym for Potentially All Pairwise RanKings of all possible Alternatives (Hansen & Ombler, 2008) – to produce weights on the five criteria for each participant.

Table 1 lists the five criteria, with their mean weights over the 476 participants reported in the final column. These 476 people self-identified as belonging to one of three key stakeholder groups: (1) 149 were patients or members of the general public (referred to as group 'P' in the tables and figures), (2) 161 were health researchers or policy-makers (R), and (3) 166 were healthcare providers (e.g. doctors and nurses) (S). The mean weights for each of these three groups (denoted P, R and S) are also reported in Table 1.

Although the mean weights for the three stakeholder groups are similar, there may still be significant differences (or 'dissimilarities') in individuals' weights within each group, and these differences may differ significantly between the groups. We use the distance measures explained in the previous section to quantify these within- and between-group differences in order to evaluate the cohesiveness of participants' weights – and ultimately, the external validity of Babashahi et al.'s (2021) preference data.

For the criterion weights and their rankings respectively, we computed the within- and between-group dissimilarities for 1000 random samples of 75 individuals from the three stakeholder groups. Figure 2 shows the cumulative distributions representing the distribution

Table 1 Criteria for ranking NCD's burden to society and their weights, for three stakeholder groups and all participants

Criterion	Stakeholder groups			All participants
	P *(n=149)*	*R* *(n=161)*	*S* *(n=166)*	*N* *(n=476)*
Deaths across the population – i.e. reduced life expectancy	0.2807	0.2748	0.2807	0.2770
Loss of quality of life across the population – e.g. pain, disability	0.2350	0.2299	0.2280	0.2309
Cost of the disease to patients, families and community – e.g. unpaid family support	0.1829	0.1852	0.1892	0.1858
Cost of the disease to the health system – i.e. publicly-funded healthcare	0.1650	0.1822	0.1678	0.1720
Disproportionately affects vulnerable groups – e.g. Māori, children, poor people	0.1363	0.1279	0.1343	0.1343

of the (relevant) distances resulting from all samples, with Table 2 reporting key summary statistics. For example, as shown in Fig. 2, bottom left plot, the between-group dissimilarity based on rankings between groups P and R takes a value of 0.75 at 0.3842, indicating that 750 out of the 1000 samples (75%) resulted in a between-group distance of at most 0.3842.

With respect to the within-group dissimilarities–i.e. comparing the cumulative distribution functions (CDFs) within Fig. 2's top two plots–we find that only the CDFs of groups P and S based on weights are not statistically distinguishable (t-test: $\mathrm{p} = 0.1119$); all other differences are statistically significant ($p < 0.0001$).

Though a bit harder to interpret, for the between-group dissimilarities – i.e. comparing the CDFs within Fig. 2's bottom two plots – we find that all differences are statistically significant ($p = 0.0011$ for $d^w(P, S)$ and $d^w(R, S)$; $p < 0.0001$ for all others). This means, for example, that group P stakeholders are not equally different to group R stakeholders compared to group S stakeholders.[2]

Most interesting is the comparison of between-group dissimilarity of a pair of groups, $d^z(G, H)$, with the within-group dissimilarities of each of the groups in this pair, $d^z(G)$ and $d^z(H)$ – i.e. comparing CDFs in the bottom two plots with the corresponding CDFs in the top two plots. For example, the statistically significant difference between $d^z(G)$ with $d^z(G, H)$ tells us that group H stakeholders are statistically distinguishable from group G stakeholders.

The only comparison that is not statistically significant ($p = 0.2816$) is between $d^w(P, S)$ and $d^w(S)$; all other differences are significant ($p < 0.0034$). If we adopt a more conservative test (in the sense of making it more difficult to reject equality) by decreasing the sample size to $s = 30$, thereby increasing the variation, we can reject equality of distributions for a few more comparisons.

[2] More conservative testing is possible by reducing the sample sizes (s) or the number of samples (t). Reductions from $s = 75$ to $s = 50$ and $s = 30$ and from $t = 1000$ to $t = 500$ produced similar results to the ones reported here.

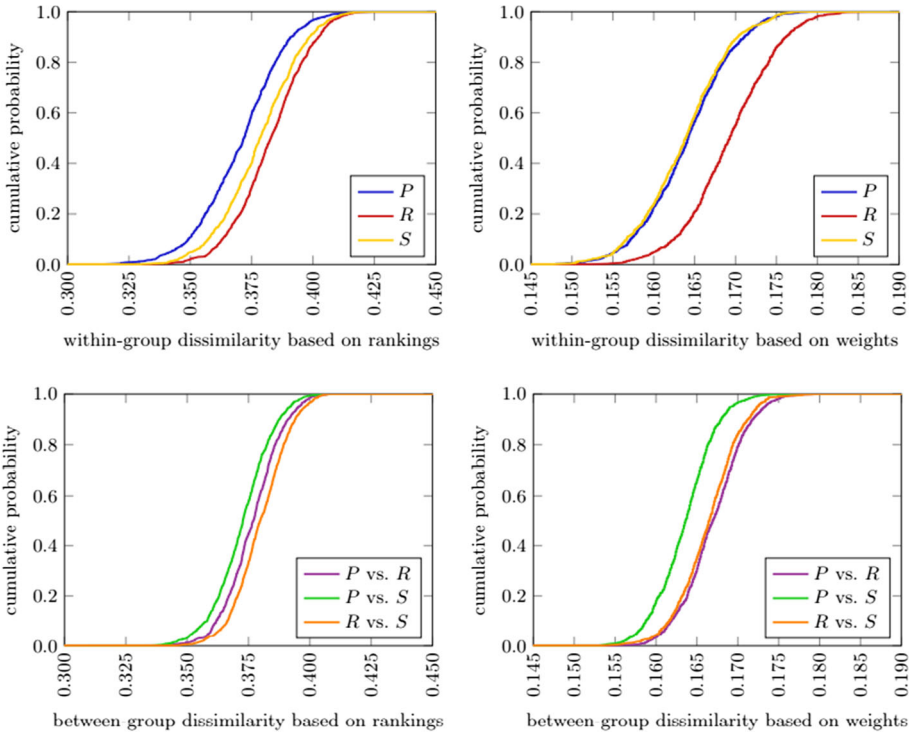

Fig. 2 Cumulative distributions of the within-group dissimilarities (top plots) and between-group dissimilarities (bottom plots) based on rankings (left plots) and weights (right plots)

Table 2 Mean within- and between-group dissimilarities (standard deviations in parentheses)

	Within-group			Between-group		
	P	R	S	Pvs.R	Pvs.S	Rvs.S
Rankings (d^r)	0.3707 (0.0169)	0.3826 (0.0149)	0.3783 (0.0160)	0.3766 (0.0113)	0.3726 (0.0118)	0.3799 (0.0110)
Weights (d^w)	0.1641 (0.0053)	0.1693 (0.0054)	0.1637 (0.0053)	0.1670 (0.0038)	0.1635 (0.0036)	0.1664 (0.0038)

However, statistical significance is not always informative regarding effect sizes. To evaluate the strength of the eventual differences, for each of the latter comparisons, we compute Cohen's d (Cohen, 1988), defined as the absolute difference between the two means divided by the pooled standard deviation; the resulting values are reported in Table 3.

According to Sawilowsky's (2009) rule of thumb, effect sizes are classified according to Cohen's d: 0.01 = very small, 0.20 = small, 0.50 = medium, 0.80 = large, 1.20 = very

Table 3 Effect sizes calculated using Cohen's d

	$d^r(P)$	$d^r(R)$	$d^r(S)$		$d^w(P)$	$d^w(R)$	$d^w(S)$
$d^r(P, R)$	0.4139	0.4529	–	$d^w(P, R)$	0.6129	0.4984	–
$d^r(P, S)$	0.1296	–	0.4070	$d^w(P, S)$	0.1315	–	0.0482
$d^r(R, S)$	–	0.2040	0.1215	$d^w(R, S)$	–	0.6198	0.5758

large and 2.0 = huge. Generally, the use of smaller sample sizes generates a larger (pooled) standard deviation and is likely to under-estimate the true effect sizes.[3]

Based on the numbers in Table 3, we can conclude that effect sizes are medium at most and hence participants' preference data (rankings and weights) are cohesive. Thus, we conclude that the non-random sampling methods employed in Babashahi et al., (2021) produced externally valid preference data.

4 Discussion and conclusion

Although random sampling is generally considered to be the gold standard of sampling methods, for logistical reasons non-random sampling is often used for public-sector MCDA surveys. With the objective of bolstering confidence in the legitimacy of MCDA's use in the health sector, ex-post sensitivity analysis of preference data across different stakeholders has been recommended (e.g. Angelis & Kanavas 2017), including in the reports of the ISPOR (International Society for Pharmacoeconomics and Outcomes Research) Emerging Good Practices Task Force (Thokala et al., 2016; Marsh et al., 2016).

For the purpose of ex-post evaluating the external validity of preference data obtained from non-random sampling, we developed and illustrated methods for measuring and assessing the cohesiveness of the weights and rankings within and between various exogenously defined groups included in the sample.

We used the Kemeny distance and the Hellinger distance as building blocks for evaluating the cohesiveness of preference data for rankings and weights. However, we could potentially have used other measures instead. Well-known possible and equally valid alternatives for the Kemeny distance include (the closely related) Kendall's W, Goodman-Kruskal's gamma, Pearson and Spearman correlation coefficients, some of which have been applied in related literature assessing cohesiveness in preference data (Sałabun & Urbaniak, 2020). Alternatives for the Hellinger distance include the Jensen-Shannon and Kullback-Leibler divergence measures.

One notable advantage of the Kemeny and Hellinger distances is that, unlike the alternative options, they do not violate the triangle inequality – i.e. for any three objects a, b and c the distance from object a to c through object b is at least as great as the distance from a to c directly (Bossert et al., 2016). Also, Kemeny distance allows the measurement of preference distances (i.e. disagreement) more accurately based on both strict and weak (or partial) rankings (Can & Storcken, 2018; Kemeny, 1959).

[3] It may be tempting to try to prevent under-estimating potential differences by choosing a larger sample size. However, a 'too large' sample eliminates all variation and results in even the smallest effect size being classified as huge.

We acknowledge that in the process of reducing the dimensionality of the data, by assigning a single number (representing the distance or dissimilarity) to a pair of informationally rich objects (i.e. herein rankings or weight vectors), there is the risk of possibly valuable information being lost. In this respect, the methods are capable of only partially identifying potential external validity problems. Nonetheless, we believe they are fit for purpose as a screening tool as they can discover serious issues (e.g. when the between-group distances are far above the within-group distances – like in the left situation in Fig. 1).

To allow for statistical testing, while correcting for biases caused by differences in observations across groups, we used bootstrapping techniques. This requires two choices: the sample size and the number of replications. In our illustrative application, we opted for a sample size of half the smallest number of observations for the groups and 1000 replications. Typically, statistical significance of differences is easier to obtain with larger sample sizes and is harder to obtain if sample sizes are too small. Optimal bootstrapping design is beyond the scope of this article and the optimal design may be context-dependent; it may, for example, depend on the number of criteria. In any case, the use of different sample sizes may reveal insights with respect to the robustness of observed differences.

As already acknowledged in the introduction, we are not the first to research external validity issues caused by non-random sampling. A common alternative approach for analyzing preference data is via cluster analysis (e.g. Kaltoft et al., 2015). First, groups ('clusters') of survey participants with similar patterns of weights or rankings are identified. By construction, these groups exhibit a high degree of within-group similarities and between-group dissimilarities. Second, the extent to which these clusters are related to participants' socio-demographic and background characteristics is investigated. During this second step, potential inaccuracies stemming from the first step are ignored, such that the errors produced in both steps compound, which may produce false positives as well as false negatives. An advantage of our distance-based method is that it only involves one layer of statistical testing, and hence errors cannot compound.

In short, with external validation performed via cluster analysis, groups (clusters) are first identified based on preference similarities and then similarities across these groups are assessed based on exogenous variation in individual characteristics. Instead, we form groups based on exogenous variation in individual characteristics and then assess similarities across preferences. Notwithstanding these fundamental differences, measures and methods developed in the clustering literature, including in pattern recognition, are of interest in the context of our method and are potentially worthwhile areas for future research.

Most of the multicriteria clustering literature is concerned with the categorization (ranking and sorting) of alternatives based on their scores for different criteria (Zopounidis & Doumpos, 2002; De Smet & Montano Guzmán, 2004; Meyer & Olteanu, 2013; Sarrazin et al., 2018). In the context of our method, these techniques are of most interest after individual preferences have been aggregated. Prior to aggregation, we advise using our method to validate elicited individual preferences if non-random sampling methods have been used in the elicitation process. Measures developed to assess (external) cluster validity are very similar to the measures we develop. For instance, the intra- and inter-heterogeneity measures developed in Rosenfeld et al., (2021) are fully in line with our within-group and between-group similarity measures. Given this methodological accordance, we believe our suggested bootstrapping techniques in support of statistical inference have the potential to contribute to this literature.

Acknowledgements The preference data used in this article are from the first author's PhD thesis, which was supervised by the two other authors and Trudy Sullivan.

Authors contributions Conceptualization: SB, RP; Methodology: SB, RP; Formal analysis and investigation: SB, RP; Writing - original draft preparation, and review and editing: SB, RP, PH; Funding acquisition: SB, PH; Software: PH; Supervision: RP, PH.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. The data used in this paper are from a research study that received financial support in the form of a University of Otago Doctoral Scholarship to the first author and two small grants from the Department of Economics and Department of Preventive and Social Medicine.

Data availability Data analyzed in this study are drawn from another study, which is accessible at https://doi.org/10.1016/j.healthpol.2020.12.003.

Code availability MATLAB codes are available via https://osf.io/9ae6w/.

Declarations

Conflicts of interest Authors have no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Angelis, A., & Kanavas, P. (2017). Multiple criteria decision analysis (MCDA) for evaluating new medicines in health technology assessment and beyond: The advance value framework. *Social Science & Medicine*, *188*, 137–156. https://doi.org/10.1016/j.socscimed.2017.06.024

Babashahi, S., Hansen, P., & Sullivan, T. (2021). Creating a priority list of non-communicable diseases to support health research funding decision-making. *Health Policy*, *125*(2), 221–228. https://doi.org/10.1016/j.healthpol.2020.12.003

Baltussen, R., Marsh, K., Thokala, P., Diaby, V., Castro, H., & Cleemput, I. (2019). Multicriteria decision analysis to support health technology assessment agencies: Benefits, limitations, and the way forward. *Value in Health, 22*(11), 1283–1288. https://doi.org/10.1016/j.jval.2019.06.014

Bossert, W., Can, B., & D'Ambrosio, C. (2016). Measuring rank mobility with variable population size. *Social Choice and Welfare*, *46*, 917–931. https://doi.org/10.1007/s00355-015-0942-z

Can, B., & Storcken, T. (2018). A re-characterization of the Kemeny distance. *Journal of Mathematical Economics*, *79*, 112–116. https://doi.org/10.1016/j.jmateco.2018.04.007

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Routledge. https://doi.org/10.4324/9780203771587

De Smet, Y., & Montano Guzmán, L. (2004). Towards multicriteria clustering: An extension of the k-means algorithm. *European Journal of Operational Research*, *158*, 390–398. https://doi.org/10.1016/j.ejor.2003.06.012

Etikan, I., Musa, S., & Alkassim, R. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, *5*(1), 1–4. https://doi.org/10.11648/j.ajtas.20160501.11

Goodman, L. (1961). Snowball sampling. *Annals of Mathematical Statistics*, *32*(1), 148–170. https://doi.org/10.1214/aoms/1177705148

Hansen, P., & Devlin, N. (2019). Multi-criteria decision analysis (MCDA) in health care decision making. *The Oxford Encyclopedia of Health Economics, Oxford University Press*. https://doi.org/10.1093/acrefore/9780190625979.013.98

Hansen, P., & Ombler, F. (2008). A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives. *Multi-Criteria Decision Analysis*, *15*(3–4), 87–107. https://doi.org/10.1002/mcda.428

Hellinger, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik*, *136*, 210–271. https://doi.org/10.1515/crll.1909.136.210

Jager, J., Putnick, D., & Bornstein, M. (2017). More than just convenient: The scientific merits of homogeneous samples. *Monographs of the Society for Research in Child Development*, *82*(2), 13–30. https://doi.org/10.1111/mono.12296

Kaltoft, M., Turner, R., Cunich, M., Salkeld, G., Nielsen, J., & Dowie, J. (2015). Addressing preference heterogeneity in public health policy by combining cluster analysis and multi-criteria decision analysis: proof of method. *Health Economics Review*, *5*, 10. https://doi.org/10.1186/s13561-015-0048-4

Kemeny, J. (1959). Mathematics without numbers. *Daedalus*, *88*(4), 577–591. https://www.jstor.org/stable/20026529

Marsh, K., Goetghebeur, M., Thokala, P., & Baltussen, R. (2017). *Multi-criteria decision analysis to support healthcare decisions*. Springer Cham. https://doi.org/10.1007/978-3-319-47540-0

Marsh, K., IJzerman, M., Thokala, P., Baltussen, R., Boysen, M., Kaló, Z., Lönngren, T., Mussen, F., Peacock, S., Watkins, J., & Devlin, N. (2016). Multiple criteria decision analysis for health care decision making—emerging good practices: Report 2 of the ISPOR MCDA emerging good practices task force. *Value in Health, 19*(2), 125–137. https://doi.org/10.1016/j.jval.2015.12.016

Meyer, P., & Olteanu, A. L. (2013). Formalizing and solving the problem of clustering in MCDA. *European Journal of Operational Research*, *277*(3), 494–502. https://doi.org/10.1016/j.ejor.2013.01.016

Rosenfeld, J., De Smet, Y., Debeir, O., & Decaestecker, C. (2021). Assessing partially ordered clustering in a multicriteria comparative context. *Pattern Recognition*, *114*, 107850. https://doi.org/10.1016/j.patcog.2021.107850

Sałabun, W., & Urbaniak, K., et al. (2020). A new coefficient of rankings similarity in decision-making problems. In V. Krzhizhanovskaya (Ed.), *Computational Science - ICCS 2020.* (Vol. 12138). Cham: Springer. https://doi.org/10.1007/978-3-030-50417-5_47

Sarrazin, R., De Smet, Y., & Rosenfeld, J. (2018). An extension of PROMETHEE to interval clustering. *Omega*, *80*, 12–21. https://doi.org/10.1016/j.omega.2017.09.001

Sawilowsky, S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 597–599. https://doi.org/10.22237/jmasm/1257035100

Thokala, P., Devlin, N., & Marsh, K. (2016). Multiple criteria decision analysis for health care decision making – an introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force. *Value in Health*, *19*, 1–13. https://doi.org/10.1016/j.jval.2015.12.003

Valerio, M., Rodriguez, N., Winkler, P., Lopez, J., Dennison, M., Liang, Y., & Turnr, B. (2016). Comparing two sampling methods to engage hard-to-reach communities in research priority setting. *BMC Medical Research Methodology, 16*, 146. https://doi.org/10.1186/s12874-016-0242-z

van Hoeven, L., Janssen, M., Roes, K., & Koffjberg, H. (2015). Aiming for a representative sample: Simulating random versus purposive strategies for hospital selection. *BMC Medical Research Methodology*, *15*, https://doi.org/10.1186/s12874-015-0089-8

Zopounidis, C., & Doumpos, M. (2002). Multicriteria classification and sorting methods: A literature review. *European Journal of Operational Research*, *138*(2), 229–246. https://doi.org/10.1016/S0377-2217(01)00243-0