Trustworthy performance evaluations: The Performance Outcome Scoring Template (POS-T) for transparent assessments in real-world programs

Benjamin P. Raysmith (ben.raysmith@athletics.org.au)
Linköping University
Toomas Timpka
Linköping University
Jenny Jacobsson
Linköping University
Michael K. Drew
Australian Institute of Sport
Örjan Dahlström
Linköping University

Research Article

Keywords: POS-T, outputs and outcomes, inputs and activities, self-comparison, singular allocation

Posted Date: September 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-899494/v1

Title

Trustworthy performance evaluations: The Performance Outcome Scoring Template (POS-T) for transparent assessments in real-world programs.

Short Title

Trustworthy performance evaluation

Authors

Benjamin P. Raysmith[1, 2, 3]*, Toomas Timpka[1], Jenny Jacobsson[1, 4], Michael K. Drew[5, 6, 7], Örjan Dahlström[1,8]

Affiliations

[1] Athletics Research Centre, Linköping University, Linköping, Sweden

[2] Western Australian Institute of Sport, Perth, Australia

[3] Athletics Australia, Melbourne, Australia

[4] Swedish Athletics Association, Stockholm, Sweden

[5] Athlete Availability Program, Australian Institute of Sport, Bruce, Australian Capital Territory, Australia

[6] Australian Collaboration for Research into Injury in Sport and its Prevention (ACRISP), Australia

[7] University of Canberra Research Institute for Sport and Exercise (UCRISE), Canberra, Australian Capital Territory, Australia

[8] Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden

*Corresponding Author: ben.raysmith@athletics.org.au

Abstract

In applied program settings, such as in natural environment control and education, performance evaluation is usually conducted by evaluators considering both self-comparison and comparison with peers. We have developed the Performance Outcome Scoring Template (POS-T) for assessments with high face-validity in these settings. POS-T puts achievements of individuals or groups in context, i.e. the resulting performance outcome score (POS) reflects a meaningful measure of performance magnitude with regards to internal and external comparisons. Development of a POS is performed in four steps supported by a statistical framework. Software is supplied for creation of scoring applications in different performance evaluation settings. We demonstrate the POS-T by evaluation of CO_2 emissions reduction amongst 36 OECD member countries.

MAIN TEXT

Introduction

Performance evaluation seeks to examine the achievement of predetermined objectives or goals by individuals or groups through the broad assessment of processes (inputs and activities) and results (outputs and outcomes).[1-3] These evaluations are used to assess program efficiency and effectiveness and provide accountability to resource allocation, strategy and policy direction.[3,4] Performance evaluations are usually case-specific and defined by the stakeholders with the authority and responsibility to do so.[5,6] Furthermore, they are expected to provide a contextual judgement of performance at a moment in time and require the measurement of credible ongoing outputs (performance measures) that relate specifically to the needs of the evaluators.[3,7-9] The assessment of goal achievement lies with the process of performance evaluation and considers a broad array of factors that include addressing the "How" and "Why" questions of achieving pre-determined objectives. The complexity associated with performance evaluations has led to the development of performance measurement systems to collect ongoing data and to monitor and report progress towards pre-determined goals.[6,7]

Performance evaluations addressed through stakeholder questions require that the measures used to report performance results are relevant and trustworthy.[9] A performance 'outcome' is defined as the resultant effect of a system towards a pre-determined objective, whereas a performance 'output' is the data generated by a single unique metric impacting the outcome.[3,10-12] A balanced performance evaluation includes both internal measurement of the individual units in the program and of the external environment.[13,14] These

measurements permit a reflection on achievement to date as well as what *could* be possible to achieve within an equal context.[15] For instance, internal measures of output include 'exam scores' in an academic setting, 'race finish time' in a sport setting, 'greenhouse gas emissions per capita' in an environment setting. Each of these metrics can be used as a performance output measure that represents 'self-comparison' when collected over time. However, internal measures alone may be insufficient in performance evaluation as the appraisal of a performance output or trend can vary when assessed relatively against peer performance under comparable circumstances. External environment measures of output include 'league tables', 'event final rankings' or 'comparisons with industry benchmarks' and have become a popular way to compare peers within industries.[16-19] As with internal measures, external measures of relative performance in isolation equally may lack face-validity or attribution (singular allocation) in performance evaluation. This is seen in circumstances where the appraisal of a ranking against peers can vary when interpreted in the context of individual achievement or progress.[20-22]

Examples of bespoke performance measurement systems that comprise singular and multiple output measures exist across industry and sectors: academic,[23] health,[24,25] profit and non-profit organisations,[26,27] sport,[28,29] natural environnent,[30] government and private sectors,[31-33] and finance.[34,35] The use of a single output measure to reflect performance outcome may provide a too narrow perspective on achievement in a complex program and therefore risk evaluation face-validity in any of these areas.[9] The selection of performance output measures that enhance attribution and enable meaningful and trustworthy evaluations therefore benefits from appropriate consideration.[25,36,37] Different measurements of performance can be combined into composite scores to increase the face-

validity of a program evaluation appraisal without adding difficulty to its interpretation.[38,39] Measurements on different data scales need here to be transformed into a common scale before combining, even though the introduction of re-scaling may reduce reliability.[40] An alignment of scales that reflect a meaningful magnitude change within and between variables can improve face-validity in regards to program outcomes and thereby make a composite score preferred for decision-making.[38,41] It is therefore essential that the measurements are transparent and comprehensible for all stakeholders involved to avoid misleading program decisions using such composite scores.[9] A robust metric must be determined to assess objectively what has occurred and how this may influence future outcomes relative to the investment in any program or individual and any third-party interest.

The aim with this study was to develop a performance scoring template that combines internal and external measures by alignment of scales that reflect meaningful magnitudes of change in stakeholder defined contexts. The purpose is to provide evaluators of applied programs a means to report performance outcomes with convincing face-validity. The scoring template is exemplified by application to evaluation of CO_2 emissions reduction amongst OECD member countries.

Results

Application of the Performance Outcome Scoring Template (POS-T) commences with selection of data sources and concludes with a composite score that is adjusted for optimal face-validity (POS) (Fig. 1). Subject to the evaluation purpose and selected time-point the template can be utilised in comparing a result with a predetermined objective, benchmark

115 standard or appraise change over time. Statistical software (in the R language) is supplied to
116 support the development of a POS (Data S1). Stepwise instructions provided in the software
117 detail file data set-up for application in the code.
118
119 Two generic data sources, achievement (continuous scale) and rank (ordinal scale) are
120 handled in four cardinal and one optional (weighting) index development steps:
121
122 - **Components and Parameters** – Quantifiable domains deemed to have primacy with
123 respect to the face-validity of the performance outcome. Stakeholder selected
124 parameters that frame the evaluation.
125 - **Component Outputs** – Performance metrics collected from each component.
126 Representative of the data collection fields that comprise the performance output
127 measure and a comparative reference output within a distribution.
128 - **Component Indicators** – Component outputs transformed to normalised measures.
129 The performance outputs measured against a reference standard given assumptions
130 of both achievement and rank deviations.
131 - **Composite score** – A final viable measure is secured through aggregating and
132 optionally weighting normalised component indicators that meet the desired face-
133 validity.
134
135 Data handling through the POS-T is described in four detailed steps. Following stakeholder
136 selections made in step 1 the outputs from steps 2-4 are produced automatically when
137 applying the performance data to the statistical R-code software provided (Data S1).
138

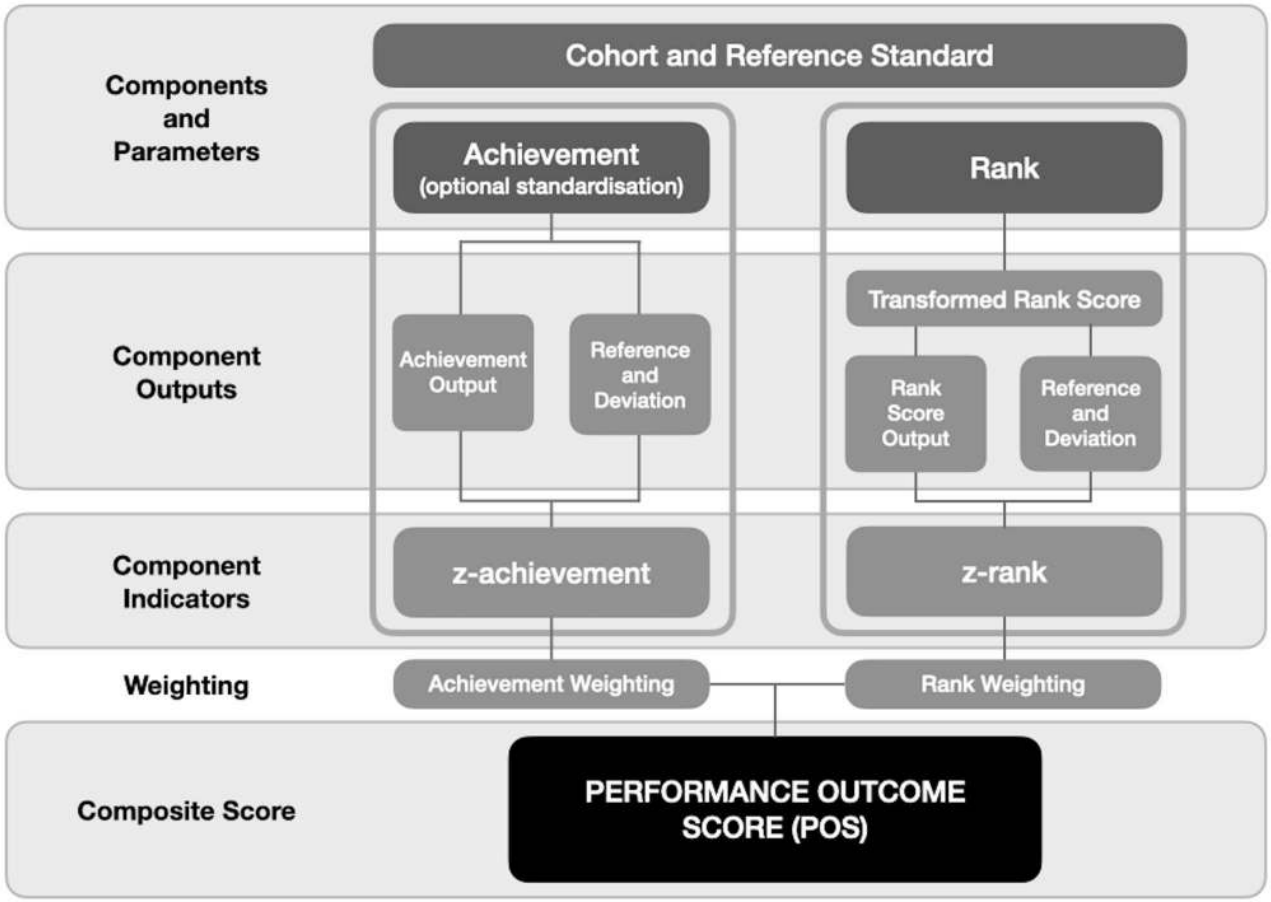


Fig. 1. POS-T signifying four cardinal and one optional (weighting) data handling levels.

Step 1 - Components (Quantifiable domains) and Parameters

Actions: Define evaluation entity (individual, group or population participating in a specified program), sample of entities to evaluate, comparison cohort, and comparator (frame of reference for evaluation). Select quantifiable domains representing internal and external measures that provide face-validity for the performance outcome of an entity (individual or group). Optional selection of a standardising parameter.

Outcome: Defined entity(s), cohort, comparator, and quantifiable domains with arguments for their selection. Optional standardising parameter.

Procedure: The entities for evaluation are selected within a cohort of interest framed by the context of the comparator. Examples of evaluation comparators include referencing a previous time point to evaluate performance over time, referencing a population mean to evaluate performance of the entity against a population standard, or referencing a single measure like a season average or predetermined objective to evaluate the entity against expectation. Next is selecting quantifiable domains that represent the **components** of 'achievement' (measured on a continuous scale) and 'rank' (measured on an ordinal scale). For the entity being evaluated 'achievement' is characterised as the component denoting self-comparison (internal measure), and 'rank' characterises the component denoting comparison with others (external measure). The optional selection of a standardising parameter is applied to the continuous data component. Standardising parameters are measures of exposure applied to the continuous data component and examples include: 'per capita' calculations, standardisation by funding, access to resources, or other parameters of exposure. Completion of step 1 is made by recording the arguments behind selecting each of the components and describing the cohort parameters, reference standards and optional standardising parameters.

Step 2 - Component Outputs (performance metrics from each component)

Action: Collect/calculate performance metrics from each component.

Outcome: For achievement and rank performance metrics refer to: Output, Reference, and Deviation.

Procedure: Performance metrics are recorded for both achievement and rank. The 'achievement output' and 'rank score output' are established as well as descriptive data for each component parameter, i.e. references and deviations (Table 1). This is completed for

each ‘entity’ (individuals or groups) in the cohort. When referring to separate entity’s subscripts *‘i’* and *‘j’* are used.

Achievement performance metrics

Output:

The **‘achievement output’** *(O_A)* is quantified directly from the metric of interest crude measure (continuous scale) and reflects the output *being evaluated*.

Reference:

The **‘achievement reference’** *(R_A)* is the metric of interest crude measure from a previous time point and is used as the metric of comparison as framed by the comparator defined in step 1. The achievement reference and output are measured by the same metric on the same continuous scale.

Deviation:

The **‘achievement deviation’** *(D_A)* is the deviation of crude measures across the observation period or other collection of entity crude measures that constitute a deviation around the achievement reference. The reference and deviation values are set from the decision to use a cohort pooled achievement standard deviation or individual entity standard deviations. The achievement outputs are assumed to follow a normal distribution, $N(\overline{A_i}, \sigma_{A_i})$, from which the reference and deviation values are set; $R_{A,i} = \overline{A_i}$ and $D_{A,i} = \sigma_{A_i}$.

Rank performance metrics

Rank scores (forming the ordinal ranking order) are transformed to a continuous scale value (transformed rank-score) using a pre-defined function, *f*, to reflect non-equidistance (magnitude difference) between different entities based on the continuous metric that

established the ranking order. Lower and upper 'reference limiters' are applied to establish the transformed rank-score range. The reference limiters represent the range of minimum and maximum reference scores and establishes a range for future equivalent comparisons. Selection should consider a range beyond current reference score minima and maxima that would account for a realistic range of future reference score possibilities. The lower achievement limiter is attributed a transformed rank-score of 1 and the upper achievement limiter is attributed a transformed rank-score of 100 (Fig. 2a). In different program settings either a higher or lower achievement score may reflect the 'best' performance. This directionality is established during the stakeholder selections at the start of the statistical R-code data handling process.

Output

The **'rank score output'** *(O_{RS})* is formed by, for each i, aggregating the comparison of transformed rank-scores between entity i, $f(\rho_i)$, and each entity *j*, $f(\rho_j)$, when entity i has a better final rank (R_i) relative to entity *j* (R_j). This reflects meaningful magnitude differences between the final ranks for entities i and *j*.

For entity i the 'rank score output' *($O_{RS,i}$)* is then defined as:

$$O_{RS,i} = \sum_{\forall j \neq i} \delta(i,j) \tag{1}$$

where

$$\delta(i,j) = \begin{cases} \frac{f(\rho_j)}{f(\rho_i)}, if\ \ R_i < R_j \\ 0 \quad\quad , if\ \ R_i \geq R_j \end{cases} \tag{2}$$

220

221 Reference and Deviation

222 The underlying performance metrics for 'rank score outputs' are generated using

223 a **simulator**. The simulator generates an underlying distribution of rank score outputs based

224 on the underlying distribution of possible achievement outputs. It randomly selects one

225 achievement output for each entity, transforms them into ranks, transformed rank-scores,

226 and finally rank score outputs. The outputs are saved for each entity i and the procedure is

227 iteratively repeated, resulting in distributions of probable rank-score outputs, Φ_i, for each

228 entity i (Fig. 2b).

229 For each entity i, the '**rank score reference**', $R_{RS,i}$, and the **'rank score deviation'**, $D_{RS,i}$, are

230 chosen as:

$$R_{RS,i} = Median(\Phi_i) \tag{3}$$

231

$$D_{RS,i} = \begin{cases} Median(\Phi_i) - P_{16}(\Phi_i), if\ O_{R,i} < Median(\Phi_i) \\ P_{84}(\Phi_i) - Median(\Phi_i), if\ O_{R,i} \geq Median(\Phi_i) \end{cases} \tag{4}$$

232

233 where P_{16} and P_{84} are the 16th and 84th percentiles, respectively.

234

COMPONENT	COMPONENT OUTPUT	METRIC QUANTIFICATION DESCRIPTION
Achievement	Achievement output *($O_{A,i}$)*	Outcome of interest crude measure.
	Achievement reference *($R_{A,j}$)*	Achievement output from previous time-point. OR Mean achievement output over a time-period. OR A population standard. OR Other comparator of interest measured by the same metric and continuous scale.

	Achievement deviation *($D_{A,i}$)*	Deviation of: Individual achievement outputs over a time-period. OR Population achievement outputs from peers in cohort of interest. OR Population deviation for metric of interest.
	Lower and Upper reference limiters	Practical lower and upper reference limits. To establish a cohort range for consistent future comparative evaluations.
Rank	Final rank *(R_i)*	Entity rank in order of crude measures at evaluation time-point.
	Initial rank *(ρ_i)*	Entity rank at time-point of comparison prior to evaluation event or period.
	Transformed rank-scores $f(\rho_i)$	Ranks transformed to a continuous value. Reflecting non-equidistance between entity ranks generated from crude measure.
	Rank score output *($O_{RS,i}$)*	Magnitude difference of final rank position relative to peers. Aggregation of transformed rank-scores from entities with a final rank behind entity i.
	Rank score reference *($R_{RS,i}$)*	Median of simulated measures Φ_i for each entity.
	Rank score deviation *($D_{RS,i}$)*	Deviations of simulated rank score outputs. Absolute value of the difference between the median and either of P_{16} or P_{84} of simulated measures Φ_i for each entity.

Table 1. Component parameters and how they are quantified for an entity 'i'.

Step 3 – Component indicators (normalised component outputs)

Action: Component outputs transformed to normalised measures reflecting the magnitude change in achievement and rank.

Outcome: z-achievement and z-rank.

Procedure: Normalised *z*-scores are calculated based on the component outputs. Z-achievement and z-rank indicators are established as proportional measures of output deviation from their corresponding references.

For the achievement component:

$$z_{A,i} = \frac{signum \cdot \left(O_{A,i} - R_{A,i}\right)}{D_{A,i}} \quad (5)$$

where

$$signum = \begin{cases} 1, if\ lower\ achievement\ scores\ are\ better \\ -1, if\ higher\ achievement\ scores\ are\ better \end{cases}$$

For the rank component:

$$z_{RS,i} = \frac{O_{RS,i} - R_{RS,i}}{D_{RS,i}} \quad (6)$$

Step 4 - Composite Score

Action: Aggregated component indicators with optional weighting.

Outcome: Performance outcome score (POS)

Procedure: The component indicators are combined in a weighting procedure based upon a user pre-defined setup of their respective relevance. An argument for the choice of weights for each respective component is formulated. The POS is established together with an explanation of the component weighting. When applying the option of proportional weighting to the component indicators the larger the importance of the component, the higher the weighting factor: w_A for achievement and w_R for rank. The weighting is performed on centred z-scores for achievement $c_{A,i}$ and rank $c_{R,i}$ respectively:

$$c_{A,i} = z_{A,i} - M_A \quad (7)$$

$$c_{R,i} = z_{R,i} - M_R \quad (8)$$

where

$$M_A = \underset{\forall i}{\text{mean}}\left(z_{A,i}\right) \quad (9)$$

$$M_R = \underset{\forall i}{\text{mean}}\left(z_{RS,i}\right) \tag{10}$$

After the weighting, the score is shifted back so that the centre of the composite score reflects the centre of the achievement indicator, so, the composite score for *entity* i is then defined as

$$POS_i = \frac{w_A c_{A,i} + w_R c_{R,i}}{w_A + w_R} + M_A =$$

$$= \frac{w_A(Z_{A,i} - \underset{\forall i}{\text{mean}}(Z_{A,i})) + w_R(Z_{R,i} - \underset{\forall i}{\text{mean}}(Z_{R,i}))}{w_A + w_R} + \underset{\forall i}{\text{mean}}(z_{A,i}) \tag{11}$$

POS-T application example

The context is 36 OECD countries (excluding countries with incomplete data publicly available) that have set out to reduce their CO_2 emissions over the 10-year period 2006 – 2015 in the global reduction of greenhouse effect (Table 2). Four entities (Sweden, Mexico, Norway, and Luxembourg) were hypothetically to be evaluated with regard to three performance goals:

Output goals:

Goal 1. “To evaluate reduction in CO_2 emissions per capita over the period 2006 - 2015”.

Goal 2. “To evaluate change in international ranking with respect to CO_2 emissions per capita”

Outcome goal:

Goal 3. “To evaluate reduction performance relative to peers regarding CO_2 emissions per capita over the period 2006 - 2015”.

To develop performance outcome evaluation measures for these goals using the POS-T, the comparative evaluation cohort is first described. The four countries (entities) will have their performance evaluated against the cohort of 36 OECD countries. The comparator framing the evaluation was defined as a comparison of performance over time.

CONTEXT	DESCRIPTION
Entities to evaluate	Four countries (Sweden, Mexico, Norway, Luxembourg)
Comparison cohort	36 OECD countries
Comparator	CO_2 emissions per capita over the period 2006 - 2015
Quantifiable domains (components)	Internal comparison - Achievement External comparison - Rank

Table 2. Contextual features framing the evaluation.

Example step 1 - Components (Quantifiable domains) and Parameters

Tonnes of CO_2 equivalent (“CO_2 emissions”) was chosen as the achievement output measure for the evaluation of Goal 1 (Table 3). A standardising parameter of ‘per capita’ was applied to the achievement component to permit direct comparison between cohort countries on a per capita basis. The achievement reference was defined as the average annual CO_2 emissions per capita during the 10-year period 1996-2005 and the achievement output for comparison was defined as CO_2 emissions per capita in the year 2015. The OECD world ranking table position was chosen as the rank output measure when evaluating Goal 2. To evaluate Goal 3, the POS-T was used to develop a POS that depicts the performance outcome regarding emission change over time for evaluation in the context of self-comparison and comparison with peers.

COMPONENT	PARAMETER	METRIC
Achievement	Output measure	Tonnes of CO_2 equivalent 2015
	Standardising parameter	Per Capita
	Reference	10-year annual output average 1996 - 2005
	Distribution	Individual entity annual variations 1996 - 2005
Rank	Output measure	OECD ranking table 2015
	Reference	Rank of median achievement references

306 **Table 3.** Components, Parameters, and metrics used to populate the component outputs.

307

308 **Example step 2 - Component Outputs (performance metrics from each component)**

309 Data were managed following the stepwise process detailed in the R-code software

310 provided (Data S1). Data for each component were collected [42] and presented in file format

311 (Data S2 and S3). Lower and upper reference limiters were set beyond the minimum and

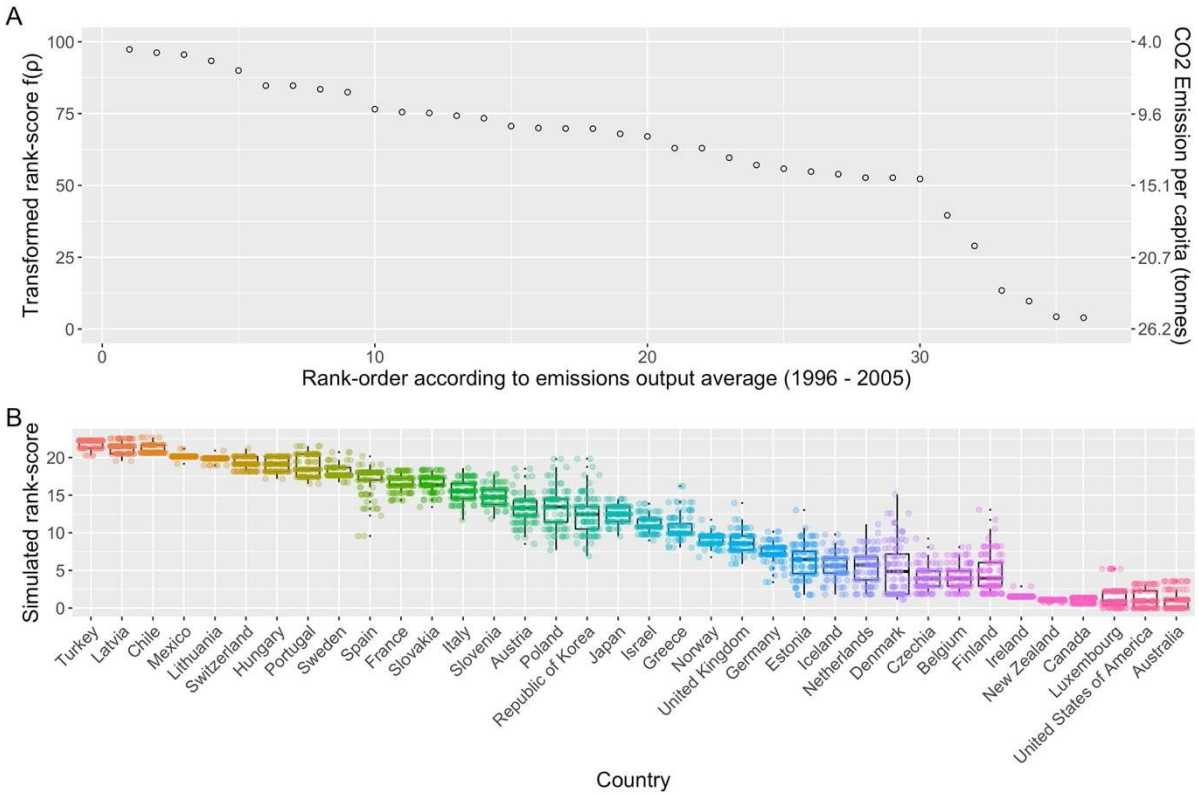
312 maximum CO_2 emissions per capita outputs. Individual entity standard deviations were

313 chosen for use in the simulations. Application of the POS-T R-code to the performance data

314 produced automated performance outputs and descriptive statistics. The following

315 component outputs were generated (Fig. 2b and Table 4).

316



317

318 **Fig 2. A. Entity rankings, B. Simulated rank scores (example data). A**. Initial rank-order (ranked by average

319 annual CO_2 emissions per capita 1996 - 2005) (ρ_i) vs average CO_2 emission per capita 1996 - 2005 (tonnes) and

320 transformed rank-score f(ρ). **B.** Simulation (example with 100 iterations shown for illustrative purposes) of

321 potential rank score outputs (Φ_i) for each entity based on underlying distribution of achievement output. Box

322 plot showing 16th and 84th percentiles.

323

COMPONENT	COMPONENT OUTPUT	SWEDEN	MEXICO	NORWAY	LUX.
Achievement	Achievement output ($O_{A,i}$)	5.47	5.74	10.47	18.17
	Achievement reference ($R_{A,j}$)	7.90	5.48	12.23	24.06
	Achievement deviation ($D_{A,j}$)	0.34	0.10	0.14	2.78
	Lower and upper reference limiters	Lower reference limiter 4 Upper reference limiter 26			
Rank	Final rank (R_i)	1	3	24	33
	Initial rank (ρ_i)	9	4	21	34
	Transformed rank-score out of 100 ($f(\rho_i)$)	82.44	93.36	62.97	9.72
	Rank score output ($O_{RS,i}$)	26.44	21.32	7.14	2.23
	Rank score reference ($R_{RS,i}$)	17.65	20.14	8.61	0.86
	Rank score deviation ($D_{RS,i}$)	1.01 (16.63 – 18.66)	1.28 (18.86 – 21.42)	0.05 (8.56 – 8.66)	1.38 (-0.53 – 2.27)

324 **Table 4.** Component outputs for the four stakeholders derived from Step 2 of the POS-T. (Units of achievement

325 = tonnes of CO_2 equivalent per capita. Rank score output = proportional score gained from ranking ahead of

other countries. Rank score reference = median score from simulation based on achievement descriptive statistics. Rank score deviation = simulation outputs based on 16th and 84th percentiles. LUX. = Luxembourg.)

Goal 1. "To evaluate reduction in CO_2 emissions per capita over the period 2006 - 2015".
Component output: Three of the four countries reduced their raw CO_2 emissions per capita. Luxembourg saw the largest reduction of the four example countries and largest reduction compared to the full OECD cohort (-5.90 tonnes per capita), followed by Sweden (13th overall; -2.43 tonnes per capita), and Norway (21st overall; -1.76 tonnes per capita). Mexico saw an increase in CO2 emissions (30th overall; +0.26 tonnes per capita).

Goal 2. "To evaluate change in international ranking with respect to CO_2 emissions per capita over the period 2006 - 2015
Component output: Three of the four countries improved their ranking. Sweden (=3rd largest rank shift overall: +8 places), Luxembourg and Mexico (=10th largest rank shift overall: +1 place) relative to all 36 comparison countries. Norway fell in ranking (=27th largest rank shift overall: -3 places) reflecting having not reduced their emissions per capita to the same level over the observation period as the comparison cohort.

Example step 3 – Component indicators (normalised component outputs)

The component outputs were transformed to normalised measures demonstrating internal and external magnitude change relative to the achievement and rank reference standards respectively (Table 5.). The three countries that saw a reduction in raw CO_2 emissions per capita demonstrated positive internal magnitude change (z-achievement: Sweden: +6.85, Norway: +12.92, Luxembourg: +2.12). One country saw an increase in raw CO_2 emissions per

capita demonstrating negative internal magnitude change (z-achievement: Mexico: -2.56). Three of four countries improved their ranking demonstrating positive external magnitude change (z-rank: Sweden: +8.69, Mexico: +0.92, and Luxembourg: +1.00). One country regressed in ranking demonstrating a negative external magnitude change (Norway: -27.73).

INDICATOR	SWEDEN	MEXICO	NORWAY	LUXEMBOURG
z-achievement $(z_{A,i})$	6.85	-2.56	12.92	2.12
z-rank $(z_{RS,i})$	8.69	0.92	-27.73	1.00

Table 5. Component indicators (normalised component outputs) derived from Step 3 of the POS-T.

Example step 4 - Composite score (POS)

The selected weighting ratio for achievement and rank was set at 1:1. The relative magnitude of change (component indicators) for both component outputs were combined, resulting in a composite score (POS) (Table 6).

Outcome goal:

Goal 3: "To evaluate reduction performance relative to peers regarding CO_2 emissions per capita over the period 2006 - 2015".

Composite score: The combined relative magnitude of change for the component indicators showed a positive development of performance for Sweden (+10.19), Luxembourg (+3.98), and Mexico (+1.60), and a negative development of performance for Norway (-4.99).

Summary information gained for performance evaluation (4 stakeholder countries highlighted)

COUNTRY	CRUDE CO_2 EMISSIONS CHANGE (<0 = reduction in CO_2 emissions)	CRUDE RANK CHANGE (>0 = improved rank)	MAGNITUDE OF EMMISSIONS CHANGE (z-achievement)	MAGNITUDE OF RANK CHANGE (z-rank)	PERFORMANCE OUTCOME SCORE (POS)
Switzerland	-1.63	2	17.32	3.37	***12.76***
Sweden	-2.43	8	6.85	8.69	***10.19***
United Kingdom	-4.46	8	7.89	7.48	***10.10***
United States of America	-4.63	0	12.72	0.00	***8.78***
Ireland	-4.70	3	8.57	3.61	***8.51***
Belgium	-4.14	6	8.34	2.86	***8.02***
Italy	-2.56	3	9.29	1.51	***7.81***
Hungary	-1.18	0	9.73	0.09	***7.33***
France	-2.21	-1	10.27	-0.52	***7.29***
Canada	-3.20	-1	9.48	-1.00	***6.66***
Finland	-4.55	11	4.75	3.45	***6.52***
Denmark	-5.59	11	3.93	3.12	***5.94***
Greece	-2.38	3	4.73	2.25	***5.91***
Slovakia	-1.85	-1	7.65	-1.00	***5.74***
Slovenia	-1.80	-1	6.82	-0.90	***5.38***
Czechia	-2.46	1	4.98	0.71	***5.26***
New Zealand	-2.19	0	5.23	0.00	***5.03***
Australia	-2.98	0	5.93	-1.00	***4.88***
Netherlands	-2.50	0	3.13	0.21	***4.09***
Luxembourg	-5.90	1	2.12	1.00	***3.98***
Israel	-1.02	-1	3.44	-0.42	***3.93***
Germany	-1.88	-2	3.47	-0.81	***3.75***
Spain	-1.98	-1	2.95	-0.33	***3.73***
Austria	-1.43	-3	2.88	-0.82	***3.45***
Portugal	-1.14	0	2.01	0.04	***3.45***
Poland	-0.41	-5	0.66	-1.70	***1.90***
Japan	-0.40	-4	2.83	-4.35	***1.66***
Mexico	0.26	1	-2.56	0.92	***1.60***
Estonia	0.24	-6	-0.26	-1.57	***1.50***
Latvia	0.79	0	-3.28	0.15	***0.85***
Lithuania	0.66	-4	-1.65	-2.91	***0.14***
Iceland	0.46	-6	-0.93	-4.36	***-0.22***
Chile	1.03	-3	-5.31	-2.13	***-1.30***
Republic of Korea	2.92	-12	-4.05	-3.40	***-1.31***
Turkey	1.42	-4	-7.94	-3.71	***-3.40***
Norway	-1.76	-3	12.92	-27.73	***-4.99***

373 **Table 6:** OECD countries ordered highest to lowest by the POS including component crude output variation
374 and component magnitude of change over the observation period. (Crude CO_2 emissions = tonnes of CO_2
375 equivalent per capita).
376

Discussion

This study set out to develop a scoring template that combines internal and external measures of performance by alignment of measurement scales which represent meaningful magnitudes of change. The resulting POS-T adheres to the principle of providing the stakeholders governing applied programs a means to report performance outcomes with convincing face-validity [37,43,44]. We exemplified application of POS-T in an evaluation of CO_2 emission reduction amongst OECD member countries. Flexible and transparent evaluation methods oriented towards stakeholders and usefulness have repeatedly been asked for in the environmental sciences [45,46]. To ensure that POS-T produces scores useful for stakeholders, an inductive (discovery) approach was found best suited.[47,48] This approach aligns with the principles of design thinking [49] where the emphasis is placed on defining the problem to be solved through the needs of the stakeholders involved.[43,44]. In the following, the main features of the POS-T are discussed considering the CO_2 emission example and directions are outlined for future research.

The measures and methods available to report performance results delimit a stakeholder's capacity to evaluate applied programs.[38] In their governance, accomplishment has been described as the gap between expected and actual output or the deviation of the output from an industry standard.[11] The POS-T provides in Step 2 methodology to establish a *magnitude* of this gap for both the achievement and rank components, i.e. for internal and external comparisons. In step 3, ready-to-combine component indicators are formed by normalisation of the component outputs produced. The resulting component indicators describe magnitudes of change, i.e. the gaps between expected and actual results or change over time for internal and external comparisons. Producing the relative magnitude of

change from a point of reference for both components (achievement and rank) rather than binary measures alone provides those evaluating the outcome with greater context. For example, on a binary scale, Sweden, Luxembourg, and Norway each demonstrated a *reduction* in CO_2 emissions per capita (positive achievement outcome) relative to their own reference standard in 2005. However, when accounting for the external context, Norway during the evaluation period slipped down the ranking table from 21st to 24th (negative rank outcome) due to that other countries reduced their relative emissions by greater amounts. Conversely, Mexico gained a ranking place from 4th to 3rd (positive rank outcome) even having had a small increase in crude CO_2 emissions due to that other countries close in rank had relatively larger increases in crude CO_2 emissions. Proportional output measures relative to self-comparison and comparison with others in a chosen cohort provides a context for stakeholders to better frame and evaluate an outcome against expectation and describe the overall accomplishment.

Integration of self-comparison and rank change magnitudes adds complexity to program evaluation indicators. Maintenance of face-validity in such composite scores requires measurement system transparency.[50] The POS-T supports transparency and face-validity by offering evaluators semantic clarity regarding the components of the integrated composite score. Stakeholders evaluating performance using the POS-T will base their assessments on normalised indicators of any measured or pre-existing method of reporting achievement output for internal (within-individual) comparisons. The crude achievement outputs can in any circumstance be ranked [51] and the normalised indicators computed for rank changes and external comparisons (between individuals). The normalised indicators are calculated in a standardised manner, i.e. for 'achievement' by subtracting the mean from an individual

raw score and then dividing the difference by the standard deviation. The mean and standard deviation are based on individual, or population standards as chosen in step 1 by the stakeholder evaluating the performance. Normalised indicators are calculated for 'rank' by transforming the ordinal scale to continuous relative values before applying the same normalisation process to a series of simulated rank outputs. Such normalised indicators are well-known and are broadly used in, for instance, global health settings for comparative evaluations of development processes at individual and population levels, e.g. in the child growth area.[52,53] In the example application of POS-T on CO_2 emissions, the normalised 'self-comparison' indicator showcases that Luxembourg's crude emissions reduction from the reference of 2005 is achieved in the context of a broader distribution of the annual fluctuation in emissions by Luxembourg compared to Sweden. In essence, Sweden's emissions reduction achievement is at face value more substantial in the context of self-comparison due to the narrower distribution of annual emissions fluctuations. The magnitude of this achievement is demonstrated by a higher achievement indicator. Regarding external comparisons, both Luxembourg and Mexico gained one place on the ranking table, yet the normalised rank indicator calculated in step 3 shows that the magnitude of Luxembourg's gain is greater than Mexico's. This is due to that Mexico's emissions per capita were very close in volume to those with similar rank, the effect being that a change in rank may occur even from small changes in emissions output resulting in a smaller rank indicator for the same crude rank change. By using the continuous data that formulates the ranking order, context and magnitude is apportioned to the component indicator in step 3 representing the rank change. Internal and external measures presented as a magnitude of change against a reference and accompanying distribution provide meaningful context to the performance. The selection of the lower and upper reference

limiters provides an important step when applying the statistical code in establishing consistent comparators and context for future equivalent evaluations. Performance outcomes presented this way can be used to observe longitudinal performance trends within an individual entity or relative to a population as well as measuring a single performance against expectation.

Some limitations to the use of the POS-T are important to consider. In experimental evaluation research, influence from external factors is controlled in the study design. Emulating an experimental design in observational performance evaluations in practice settings would require information on all confounding factors.[54] Application of the POS-T does not per se assure that the POS reflects causal effects of the program, and consideration of confounding factors is always needed when interpreting POS scores in practice settings. Moreover, it should be taken into regard that the simulation process used in POS-T to determine the rank score deviations for each entity uses the reference and its standard deviation as assumptions in the calculation. The outcome from each simulation may thus vary slightly. This effect is minimised by always running an adequate number of simulations on each occasion. Furthermore, when the data available to calculate the achievement deviation is limited, a decision must be made regarding what to use as the achievement deviation for comparison with the achievement output. The preferred option is to use the achievement deviation unique to each entity. However, an option is to use the cohort population standard deviation as this broadens the dataset to calculate deviations and improves its reliability. This may be a satisfactory solution when the comparative data sets between entities in the cohort have similar deviations. If this is not the case, an option may be to use the largest or smallest deviation in the cohort. The flexibility in selecting

components, standardising parameters and weighting of the component indicators opens the composite score to variability in its robustness. Testing for robustness is recommended and aided by the level of transparency described by the evaluator in selecting optional features in corresponding steps of the POS-T framework.

The POS-T in its current form can be applied to any program governance setting. A POS can be determined for single entities at multiple time points to assess performance trends or for multiple entities at a single time point to assess performances relative to a population standard or to peers. The component indicators can also be used to evaluate each component in isolation. The rank simulation process in isolation may furthermore be utilised to determine probabilities of performance outcome. Further development of the POS-T will include development of the statistical code to include the evaluation of performance in settings where achievement is not readily quantified, e.g. when it mainly is established through head-to-head contests.

Conclusion

The POS-T endorses face-validity in real-world program evaluations by that the resulting POS reflects a meaningful magnitude of performance outcome with regards to self-comparison and comparison with peers. The template is presented with statistical software for creating scoring systems and is exemplified by evaluation of CO_2 emissions reduction amongst 36 OECD member countries. Forthcoming research will involve application of the POS in different applied performance evaluation settings.

Materials and Methods

Construction principles

Construction of the POS-T employed an iterative approach to solution design that prioritises application of the final template in real-world program governance settings. Its practical use was further supported through the parallel construction of a statistical framework and software for score development in applications.[49] A design panel was composed for the construction consisting of scientists and practitioners (n=5) with backgrounds in epidemiology, public health and sport settings, organisational development, statistical methods, and experimental design. A composite score development model was used to guide the construction process (Fig. 3).

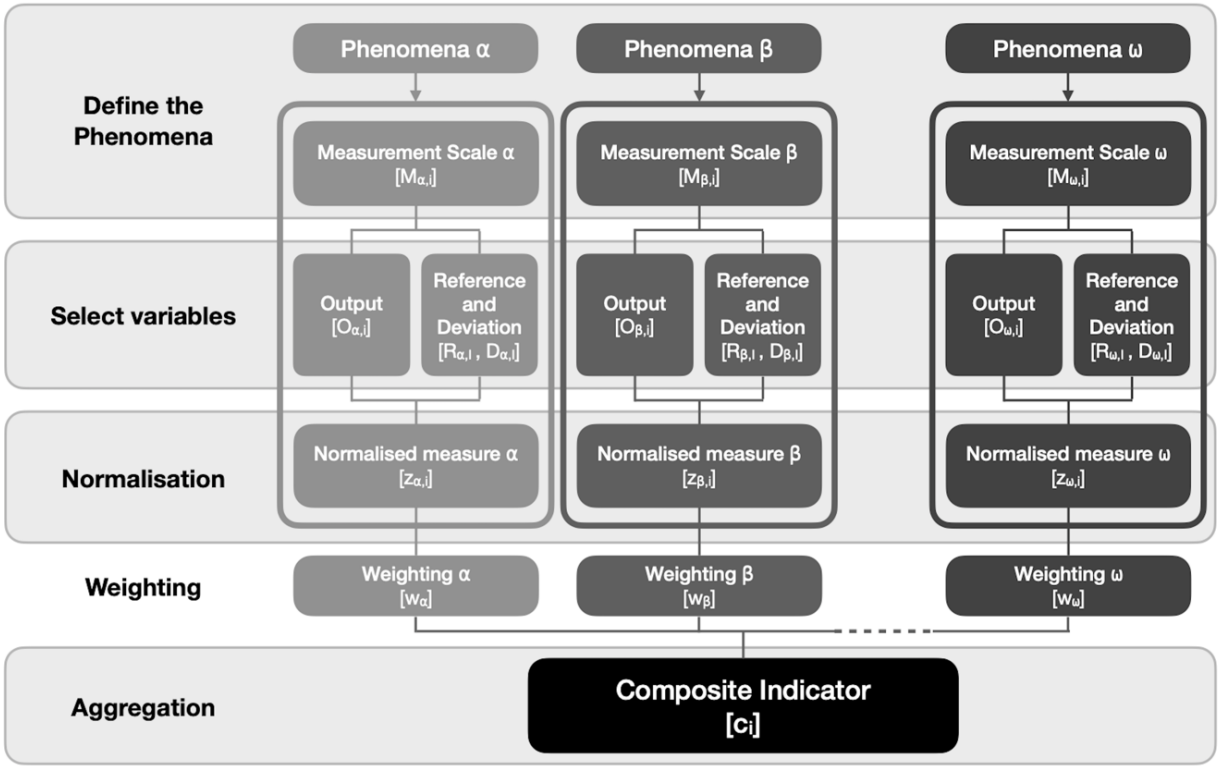


Fig. 3. Composite score development model with five development steps and three example phenomena.

Template construction process

The design panel met via an online meeting platform weekly over a twelve-month period and discussed the POS-T development in the context of four cardinal steps and one optional step depicted in the construction model (Fig. 3). In each development model step, the design panel employed an iterative process applying varying methods of data analysis and representation in the template to identify potential inconsistencies or errors in the composite score. These were identified, discussed, and addressed at each online meeting until panel consensus on user application was reached. Consensus required agreement on the stepwise process necessary for the user when defining the context of POS-T evaluation.

Once the process was established and incorporated into the POS-T consensus on the maintenance of the aggregated composite score face-validity was obligatory.

Statistical code was written using R programming language to automate the methodological outputs of steps 2-4 (Data S1). Data comprising the variables outlined in step 2 of the development model were systematically organised in data files using Microsoft Excel 2016 (Data S2 and S3). The current version of the R-code was written to use with discrete cohorts that comprise all entities in sequential ranking order for analysis.

POS-T application example

The resulting POS-T was finally applied in an evaluation of reduction of greenhouse gas emission among 36 OECD countries. The OECD has collected and reported annual greenhouse gas emissions between 1996 and 2015 for a cohort of 36 countries producing an annual emissions ranking table.[42] The OECD data set was used to showcase performance evaluations based on a composite score for comparison within either of the singular reporting metrics; tonnes of CO_2 emissions equivalent per capita, or the emissions ranking table. To showcase the template and the statistical framework, the design panel took on the virtual role of an international stakeholder commission in the environmental protection area. The final statistical framework was exemplified by applying the R-code to the 36 country OECD data set and analysing the performance outcomes of four countries: Sweden, Luxembourg, Mexico, and Norway.

References

1. "Performance Evaluation Methods: Measurement and Attribution of Program Results", (Treasury Board of Canada, Secretariat, Toronto, 1998).
2. A. DeGroff, M. Schooley, T. Chapel, T. H. Poister, Challenges and strategies in applying performance measurement to federal public health programs. *Evaluation and program planning* **33**, 365-372 (2010).
3. "Performance measurement and evaluation: Definitions and relationships." (U.S. Government Accountability Office, GAO-11-646SP, Washington DC, 2011).
4. P. de Lancer Julnes, Performance measurement: An effective tool for government accountability? The debate goes on. *Evaluation* **12**, 219-235 (2006).
5. M. Lebas, Performance measurement and performance management. *International Journal of Production Economics* **41**, 23-35 (1995).
6. K. E. Newcomer, Using performance measurement to improve programs. *New directions for evaluation* **1997**, 5-14 (1997).
7. T. Rantala. "Operational level performance measurement in university-industry collaboration", thesis, LUT University, Finland (2019).
8. B. Milstein, S. Wetterhall, C. E. W. Group, A framework featuring steps and standards for program evaluation. *Health Promotion Practice* **1**, 221-228 (2000).
9. J. P. Koplan, R. Milstein, S. Wetterhall, Framework for program evaluation in public health. *MMWR: Recommendations and Reports* **48**, 1-40 (1999).
10. L. G. Suter, C. E. Barber, J. Herrin, A. Leong, E. Losina, A. Miller, E. Newman, M. Robbins, H. Tory, J. Yazdany, American College of Rheumatology white paper on performance outcome measures in rheumatology. *Arthritis care & research* **68**, 1390-1401 (2016).
11. W. D. Savedoff, "Governance in the health sector: a strategy for measuring determinants and performance", (Policy Research working paper; no. WPS 5655, The World Bank, 2011).
12. L. Laurian, J. Crawford, M. Day, P. Kouwenhoven, G. Mason, N. Ericksen, L. Beattie, Evaluating the outcomes of plans: Theory, practice, and methodology. *Environment and Planning B: Planning and Design* **37**, 740-757 (2010).

13. A. Neely, M. Gregory, K. Platts, Performance measurement system design: a literature review and research agenda. *International journal of operations & production management* **15**, 80-116 (1995).
14. B. P. Raysmith, J. Jacobsson, M. K. Drew, T. Timpka, What Is Performance? A Scoping Review of Performance Outcomes as Study Endpoints in Athletics. *Sports* **7**, 66 (2019).
15. "The world health report 2000: health systems: improving performance", (World Health Organization, 2000).
16. D. D. Dill, M. Soo, Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems. *Higher education* **49**, 495-533 (2005).
17. F. Derrien, O. Dessaint, The effects of investment bank rankings: Evidence from M&A league tables. *Review of Finance* **22**, 1375-1411 (2018).
18. S. G. Armesto, M. L. G. Lapetra, L. Wei, E. Kelley, "Health care quality indicators project 2006 data collection update report.", (OECD Health Working Papers, No. 29, OECD Publishing, Paris, 2007).
19. M.-J. Huang, M.-Y. Chen, K. Yieh, Comparing with your main competitor: the single most important task of knowledge management performance measurement. *Journal of Information Science* **33**, 416-434 (2007).
20. A. Mehrpouya, R. Samiolo, Performance measurement in global governance: Ranking and the politics of variability. *Accounting, organizations and society* **55**, 12-31 (2016).
21. T. R. Oliver, Peer Reviewed: Population Health Rankings as Policy Indicators and Performance Measures. *Preventing chronic disease* **7**, (2010).
22. K. Keasey, P. Moon, D. Duxbury, Performance measurement and the use of league tables: some experimental evidence of dysfunctional consequences. *Accounting and business research* **30**, 275-286 (2000).
23. M.-H. Huang, A comparison of three major academic rankings for world universities: From a research evaluation perspective. *Journal of Library & Information Studies* **9**, (2011).
24. S. Schütte, P. N. M. Acevedo, A. Flahault, Health systems around the world–a comparison of existing health system rankings. *Journal of global health* **8**, (2018).
25. J. Hurst, M. Jee-Hughes, "Performance measurement and performance management in OECD health systems.", (OECD Publishing, Paris, 2001).

595 26. R. S. Kaplan, Strategic performance measurement and management in nonprofit organizations.
596 *Nonprofit management and Leadership* **11**, 353-370 (2001).
597 27. R. Kaplan, D. Norton, The Balanced Scorecard—Measures that Drive Performance. *Harvard Business*
598 *Review*, 71-79 (1992).
599 28. J. Fahlén, The trust–mistrust dynamic in the public governance of sport: exploring the legitimacy of
600 performance measurement systems through end-users' perceptions. *International journal of sport*
601 *policy and politics* **9**, 707-722 (2017).
602 29. I. O'Boyle, D. Hassan, Performance management and measurement in national-level non-profit sport
603 organisations. *European Sport Management Quarterly* **14**, 299-314 (2014).
604 30. H. Pham, B. G. Sutton, P. J. Brown, D. A. Brown, Moving towards sustainability: A theoretical design of
605 environmental performance measurement systems. *Journal of Cleaner Production* **269**, 122273
606 (2020).
607 31. N. Carter, P. Day, R. Klein, *How organisations measure success: the use of performance indicators in*
608 *government* (Psychology Press, 1995).
609 32. R. Jacobs, M. Goddard, P. C. Smith, Composite performance measures in the public sector. *Center for*
610 *Health Economics Research Paper* **16**, (2007).
611 33. I. C. Davies, Evaluation and performance management in government. *Evaluation* **5**, 150-159 (1999).
612 34. "Beyond roe-how to measure bank performance.", (*Appendix to the report on EU banking structures*,
613 European Central Bank, Frankfurt, 2010).
614 35. L. J. White, *The credit rating agencies and their role in the financial system.* (Oxford University Press,
615 London, 2018).
616 36. T. Seeley, "Cumminuty based organization (CBO) survey results: Outcome evaluation in voluntary and
617 not-for-profit organizations.", (The Muttard Fellowship, Edmonton, 2003).
618 37. D. Guyadeen, M. Seasons, Evaluation theory and practice: Comparing program evaluation and
619 evaluation in planning. *Journal of Planning Education and Research* **38**, 98-110 (2018).
620 38. E. D. Peterson, E. R. DeLong, F. A. Masoudi, S. M. O'Brien, P. N. Peterson, J. S. Rumsfeld, D. M.
621 Shahian, R. E. Shaw, ACCF/AHA 2010 position statement on composite measures for healthcare
622 performance assessment: American College of Cardiology Foundation/American Heart Association

Task Force on performance measures (writing committee to develop a position statement on composite measures). *Journal of the American College of Cardiology* **55**, 1755-1766 (2010).

39. R. Jacobs, M. Goddard, P. C. Smith, How robust are hospital ranks based on composite performance measures? *Medical care*, 1177-1184 (2005).

40. M. Nardo, M. Saisana, A. Saltelli, S. Tarantola, A. Hoffman, E. Giovannini, "Handbook on Constructing Composite Indicators: Methodology and user guide.", (OECD Publishing, Paris, 2005).

41. M.-K. Song, F.-C. Lin, S. E. Ward, J. P. Fine, Composite variables: when and how. *Nursing research* **62**, 45 (2013).

42. "Greenhouse gas emissions by source" Internet. OECD Publishing; 2014. Available from: https://www.oecd-ilibrary.org/content/data/data-00594-en.

43. A-M. R. McGowan, C. Bakula, R. S. Castner, "Lessons Learned from Applying Design Thinking in a NASA Rapid Design Study in Aeronautics ", in *58th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*. (2017), pp. 0976.

44. M. Greene, "Systems Design Thinking: Identification and Measurement of Attitudes for Systems Engineering, Systems Thinking, and Design Thinking", thesis, University of Michigan, (2019).

45. V. J. Schwanitz, Evaluating integrated assessment models of global climate change. *Environmental modelling & software* **50**, 120-131 (2013).

46. M. Haasnoot, W. Van Deursen, J. H. Guillaume, J. H. Kwakkel, E. van Beek, H. Middelkoop, Fit for purpose? Building and evaluating a fast, integrated model for exploring water policy pathways. *Environmental modelling & software* **60**, 99-120 (2014).

47. J. D. Gould, C. Lewis, Designing for usability: key principles and what designers think. *Communications of the ACM* **28**, 300-311 (1985).

48. K. Dorst, The core of 'design thinking'and its application. *Design studies* **32**, 521-532 (2011).

49. T. Brown, Design thinking. *Harvard business review* **86**, 84 (2008).

50. A. Rae, C. Wong, Monitoring spatial planning policies: towards an analytical, adaptive, and spatial approach to a 'wicked problem'. *Environment and Planning B: Planning and Design* **39**, 880-896 (2012).

51. M. Mohsin, A. Rasheed, H. Sun, J. Zhang, R. Iram, N. Iqbal, Q. Abbas, Developing low carbon economies: An aggregated composite index based on carbon emissions. *Sustainable Energy Technologies and Assessments* **35**, 365-374 (2019).
52. J. M. Wit, J. H. Himes, S. Van Buuren, D. M. Denno, P. S. Suchdev, Practical application of linear growth measurements in clinical research in low-and middle-income countries. *Hormone research in paediatrics* **88**, 79-90 (2017).
53. M. Leung, N. Perumal, E. Mesfin, A. Krishna, S. Yang, W. Johnson, D. G. Bassani, D. E. Roth, Metrics of early childhood growth in recent epidemiological research: a scoping review. *PloS one* **13**, e0194565 (2018).
54. M. A. Hernán, J. M. Robins, Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology* **183**, 758-764 (2016).

Acknowledgments

Funding

The authors acknowledge that they received no funding in support for this research.

Author contribution

BPR provided conceptual basis and primary author, OD developed the statistical code and refined conceptual thinking to practical outputs, TT guided the projects and provided major edits to manuscript, MKD and JJ contributed to conceptual development and manuscript edits.

Competing interests

The authors declare that they have no competing interests.

Data and material availability

All data are available in the main text or the supplementary materials.

Figure and table legends

Fig. 1. POS-T signifying four cardinal and one optional (weighting) data handling levels.

Table 1. Component parameters and how they are quantified for an entity 'i'.

Table 2. Contextual features framing the evaluation.

Table 3. Components, Parameters, and metrics used to populate the component outputs.

Fig 2. A. Entity rankings, B. Simulated rank scores (example data). A. Initial rank-order (ranked by average annual CO_2 emissions per capita 1996 - 2005) (ρ_i) vs average CO_2 emission per capita 1996 - 2005 (tonnes) and transformed rank-score $f(\rho)$. **B.** Simulation (example with 100 iterations shown for illustrative purposes) of potential rank score outputs (Φ_i) for each entity based on underlying distribution of achievement output. Box plot showing 16^{th} and 84^{th} percentiles.

Table 4. Component outputs for the four stakeholders derived from Step 2 of the POS-T. (Units of achievement = tonnes of CO_2 equivalent per capita. Rank score output = proportional score gained from ranking ahead of other countries. Rank score reference = median score from simulation based on achievement descriptive statistics. Rank score deviation = simulation outputs based on 16^{th} and 84^{th} percentiles. LUX. = Luxembourg.)

Table 5. Component indicators (normalised component outputs) derived from Step 3 of the POS-T.

Table 6: OECD countries ordered highest to lowest by the POS including component crude output variation and component magnitude of change over the observation period. (Crude CO_2 emissions = tonnes of CO_2 equivalent per capita).

Fig. 3. Composite score development model with five development steps and three example phenomena.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S1POST.r
- S2OECDfoundationdata.xlsx
- S3OECDrankingdata.xlsx