ESCoE ECONOMIC STATISTICS CENTRE OF EXCELLENCE

A COLLABORATION WITH Office for National Statistics

Composite Indices: Reflections on Recent UK Experience and Lessons for Public Service Productivity Measurement

Richard Heys

ESCoE Discussion Paper 2024-18

December 2024

ISSN 2515-4664

DISCUSSION PAPER

Composite Indices: Reflections on Recent UK Experience and Lessons for Public Service Productivity Measurement
Richard Heys
ESCoE Discussion Paper No. 2024-18
December 2024

Abstract

It is an age-old debate in statistics and measurement: when to provide a relatively large number of series of data (a dashboard) to permit the user to extract a common signal themselves, and when to create an artificial 'single-number' index through some process of amalgamation or aggregation such that a single index provides a simple (some say over-simplified) line of sight for users on whether matters have improved, worsened, or stayed constant.

This paper reflects on recent experience in the UK, in relation to four specific instances: the construction of Gross and Net Inclusive Income, the development of Public Service Productivity estimates, the National Wellbeing Dashboard and the creation of the Health Index to consider what these examples teach us, to consider how we can move forward in terms of when composite indices are valuable / useful and when they are not.

The paper concludes that a composite index has value when the weights used to derive it represent societal preferences or a societal reflection of value, and reviews alternative approaches which have been proposed or implemented.

Keywords: Composite indices, well-being, public services, productivity, GDP

JEL classification: C18, C65, D78, E01, I31

Richard Heys, Office for National Statistics
richard.heys@ons.gov.uk

Published by:
Economic Statistics Centre of Excellence
King's College London
Strand
London
WC2R 2LS
United Kingdom
www.escoe.ac.uk

Composite Indices: Reflections on Recent UK Experience

and lessons for Public Service Productivity Measurement

Richard Heys

Office for National Statistics[1]

Executive Summary

It is an age-old debate in statistics and measurement: when to provide a relatively large number of series of data (a dashboard) to permit the user to extract a common signal themselves, and when to create an artificial 'single-number' index through some process of amalgamation or aggregation such that a single index provides a simple (some say over-simplified) line of sight for users on whether matters have improved, worsened, or stayed constant.

This paper reflects on recent experience in the UK, in relation to four specific instances: the construction of Gross and Net Inclusive Income, the development of Public Service Productivity estimates, the National Wellbeing Dashboard and the creation of the Health Index to consider what these examples teach us, to consider how we can move forward in terms of when composite indices are valuable / useful and when they are not.

The paper concludes that a composite index has value when the weights used to derive it represent societal preferences or a societal reflection of value, and reviews alternative approaches which have been proposed or implemented.

JEL Codes: C18, C65, D78, E01, I31

[1] The views in this paper are those of the author. The author would like to thank Cliodhna Taylor, Skevi Pericleous, Eleanor Rees, Craig McLaren and two anonymous referees for insightful comments, however, any errors remain the authors.

1. Introduction

Quality statistics meet the needs of users. However, different users have different needs and different degrees of competence in interpreting and understanding the data statisticians put before them. For as long as there has been statistics there has been a debate: how much to simplify or 'spoon-feed' users, or conversely, how much to trust them to draw what they need to find from within the data themselves. Those who support the provision of composite indices see value in ease and speed of use, simplicity of interpretation and more straightforward communication: those who oppose the use of composites see valuable information being drowned beneath an overarching message which, in the worse circumstances, is imposed on the data through the method of aggregation. There will probably never be a universal winner in this debate: the answer depends on circumstance, culture, history, and a variety of other factors.

This paper, therefore, will focus on a narrower question, namely: what are the circumstances whereby we can maximise the potential value of composite indices, by minimising the twin risks of 'variable omission' and 'artificial signal / signal suppression'?

This paper breaks this down into four parts: (2) definition of terms, (3) a theory within which to tackle the question above, (4) a review of the four recent developments at the UK Office for National Statistics which are of relevance, namely the construction of Gross and Net Inclusive Income, the development of Public Service Productivity estimates, the National Wellbeing Dashboard and the creation of the Health Index, and finally (5) conclusions.

2. Definitions

Let us imagine a dataset **D** formed of three data time series (D^1, D^2, D^3), which at time t can be presented in one of two ways. The first way is as a dashboard (B_t) where B_t is the simple set of the three data values:

$$B_t = \{D^1_t, D^2_t, D^3_t\} \quad (1)$$

All data is maintained at all times and is equally visible to users.

Now let us imagine a composite (C_t). C_t is also derived from the set (D^1, D^2, D^3), but through a process of linear aggregation, a single variable is created:

$$C_t = \alpha^1{}_t D^1{}_t + \alpha^2{}_t D^2{}_t + \alpha^3{}_t D^3{}_t \quad (2)$$

To do this, a series of weights (α^1,α^2,α^3) is created and applied. In this example, we would permit the value of α to change through time, and C_t is created through a simple additive process. This permits a simple formula to be developed without loss of generalisation.

One can easily think of Gross Domestic Product as being in this form[2] through the use of money value weights (i.e. prices (p)) to derive each of the output, income and expenditure measures of GDP[3]:

$$GDP_t = Y_{1t} = \sum{}^{i}{}_{1} p_{it} f(K_{it},L_{it}) = r_t K_t + w_t L_t = p_{Ct}C_t + p_{It}I_t + p_{Gt}G_t + (p_{Xt}X_t - p_{Mt}M_t) \quad (3)$$

This framework is obviously applicable to datasets of more components (D^n), but is clearly also heavily dependent on the choice of the set ($\alpha^1 \ldots \alpha^n$). The two risks mentioned above broadly relate to the potential for bias to be invited into the composite either through (1) variable omission, (2) artificial signal, or (3) collinearity.

Variable Omission

This argument relates to the omission of a variable (D^x) which has pertinent bearing on the policy issue under consideration – using a simple example, excluding the manufacturing sector and its output from GDP.

In this case, let us assume a four-sector economy, comprising manufacturing (M), services (S), Government (G) and Construction (C), such that the true composite Y_t would be composed as follows[4]:

$$Y_t = \alpha^m{}_t M_t + \alpha^s{}_t S_t + \alpha^c{}_t C_t + \alpha^g{}_t G_t \quad (4)$$

[2] Recognising that the process of aggregation is actually through the application of chain volume measures which are not strictly additive.

[3] Where Y represents output (GDP), K represents capital assets, L represents labour, r is the rate of return on capital, w is average wages and salaries, C is consumption, I is investment, G is government expenditure, W is exports, M is imports. The second component of the formula represents the output approach where f(..) can be aggregated across a number of industries. The third component is the income measure, and the fourth is the expenditure approach.

[4] In this case, α represent p. We use α to demonstrate the generic nature of this example.

In the instance of variable omission, the following formula would be used, excluding M:

$$Y_t^* = \alpha^s_t S_t + \alpha^c_t C_t + \alpha^g_t G_t \tag{5}$$

Clearly, the impact of such an omission becomes obvious and immediately implies that the need for a clear framework justifying the inclusion and exclusion of component series is required. We will discuss this further in section (3).

Artificial Signal or Signal suppression

This argument relates to the suppression of signal contained in the data set **D**, now generalised to contain data series ($D^1 \ldots D^n$) through the choice of a set ($\alpha^1 \ldots \alpha^n$) which places an 'artificially' strong or weak weight on the data series **D** such that the composite index fails to react to pertinent signal contained within an individual series, or indeed a number of series. Again, using the simple example from GDP, let us imagine we now include manufacturing output into GDP, but we give it an artificially low weight such that movements in the series do not impact the headline composite.

Returning to equation (4), one can clearly see that as α^m_t approaches zero, the value of Y_t will converge with the value Y_t^* in equation (5)[5]. This paper will label this 'strong signal suppression' in that some artificial value for α has caused this suppression. 'Strong' therefore would capture instances of some form of subjective intervention in the setting of the weights.

This is in opposition to what this paper will refer to as 'weak signal suppression'. Weak signal suppression occurs where a data series (D^o) is an outlier in terms of movement / volatility or some other important statistic characteristic, but where the value of Y_t is independent of the behaviour of D^o despite the use of a set of weights which have not experienced subjective intervention. An example would be where in equation (4) α^m_t approaches zero because manufacturing's weight in the economy was approaching zero as a fraction of the economy. In this circumstance volatile growth or a markedly different growth pattern from that experienced by the rest of the economy would become invisible within the headline measure. In this example, of course, weak signal suppression is not necessarily a problem if within the overarching framework this does not create a change in eliciting the signal inherent in the dataset: one outlier may not reveal a flaw in the index, but the next section considers the theory we can use to determine this.

[5] Obviously, this would imply to any other functional form which over-simplified the relationship and dynamics between the individual series, D, and the composite, C.

Collinearity

Another issue apart from variable omission and signal suppression/artificial signal is where the data series ($D^1 \ldots D^n$) within **D** demonstrate strong relationships. Clearly this has two implications; firstly this impacts the selection of variables for the composite C, but secondly there is a question of efficiency in terms of the number of variables which need to be included in the composite C. This provokes important questions around the degree to which data series could serve as proxies for other data series, and the nature of the relationship between the different D; is it linear for example? This may impact our appetite for the development either of certain composites or the policy validity of the final measure. Nevertheless, it is likely there is a balance to be struck. One would wish all the D going into C to be relevant and therefore would be likely to have some relationship between themselves.

This implies a requirement for a clear final objective **O**, which serves as the definition of the topic of interest which defines the scope of any dataset (**D),** such that the rule for inclusion of D_i in **D** can be expressed in terms of **O.** This is considered below.

3. Theory

As Rosen (1991) shaped this point:

> *'With regard to models, the justification for a composite indicator lies in its fitness for the intended purpose and in peer acceptance.'*

The two challenges of variable omission and signal suppression illustrate the need for a theoretical basis upon which to assemble both the dataset ($D^1 \ldots D^n$) and the weight-set ($\alpha^1 \ldots \alpha^n$). Sharpe (2004) describes the challenge well:

> *'The aggregators believe there are two major reasons that there is value in combining indicators in some manner to produce a bottom line. They believe that such a summary statistic can indeed capture reality and is meaningful, and that stressing the bottom line is extremely useful in garnering media interest and hence the attention of policy makers. The second school, the non-aggregators, believe one should stop once an appropriate set of indicators has been created and not go the further step of producing a composite index. Their key objection to aggregation is what they see as the arbitrary nature of the weighting process by which the variables are combined.'*

The two main guidance documents in this space (OECD (2008) and UNECE (2019)) both focus on these issues (alongside other factors, such as imputation of missing data). OECD (2008) is particularly clear that:

'A composite indicator is formed when individual indicators are compiled into a single index on the basis of an underlying model'[6]... and… *'[t]he quality of a composite indicator as well as the soundness of the messages it conveys depend not only on the methodology used in its construction but primarily on the quality of the framework and the data used.'*[7]

3.1. Dataset composition

As such we will consider first the choice of indicators to form the dataset **D.** Referencing OECD (2008) again:

'The selection criterion should work as a guide to whether an indicator should be included or not in the overall composite index. It should be as precise as possible and should describe the phenomenon being measured, i.e. input, output or process. Too often composite indicators include both input and output measures'[8]... and ... *'the type of variables selected – input, output or process indicators – must match the definition of the intended composite indicator.'*[9]

In short, the intellectual basis for inclusion should be derived from a rationale that the data forms part of a coherent ecosystem where inclusion is justifiable in the narrative which the composite is looking to speak to. So, an economic composite would not include variables such as measures of grief or perceptions of housing quality unless the definition of economic spanned into such social matters, or there was a clear 'theory of change' which explained why changes in these variables would be expected to impact on the headline measure of interest. The need to distinguish between different categories / classes of data is of importance here:

- One may wish to distinguish between inputs, outputs, and outcomes in this context, as we shall see with the public service productivity estimates described below,
- One may wish to differentiate between flows and stocks in terms of using accounting data, similarly with current and constant price data, or,

[6] Page 15
[7] Page 17
[8] Page 22
[9] Page 23

- One may wish to consider whether to mix together ‘goods’ and ‘bads’ into a dashboard or index: is a change which made a ‘bad’ ‘worse’ of equivalent/greater/lesser importance that something which causes a ‘good’ to ‘improve’?

Importantly, this is a recommendation which applies as much to dashboard *or* composite indicators, as it relates to the assembly of the dataset **D**, not the reporting mechanisms derived from it.

This implies a requirement for a clear final objective **O**, which serves as the definition of the topic of interest which defines the scope of **D,** such that the rule for inclusion of D_i in **D** can be expressed as:

$D_i \in \mathbf{D}$ if $\mathbf{O} = f(D_i)$ where $i = 1....n$ (6)

Whilst algebraically complete, equation (6) opens a question of degrees: how strong a relationship between Di and **O** is required to justify inclusion? This essentially is the challenge dashboard approaches have routinely faced: set the threshold too low and one ends up with hundreds of items, set the threshold too high and only a small number of factors may feature and one runs increasing risks of variable omission. Similarly, as touched on above, we remain faced with the challenge of collinearity; take two factors D_a and D_b that both would meet the requirement in equation (6), but which also demonstrate which have a high degree of correlation.). Is there value in the inclusion of both if, in combination, they add no or little explanatory value, or the inclusion of one may distort the impact of the other? In the pursuit of brevity in the methods proposed, it may be sufficient to undertake standard tests of collinearity or to amend equation (6) to present a stepwise argument whereby inclusion is conditional on the existing membership of **D**, (here described as $\mathbf{D_1...D_n}$), where F(…) is an acceptable benchmark value for the degree of correlation between D_i and **O**, which is itself set objectively by the degree of correlation observed for the existing set **D**:

$D_i \in \mathbf{D}$ on condition of $(\mathbf{O} = f(D_i \mid \mathbf{D})) \mid (f(D_i) >= F(D_i \mid \mathbf{D_1...D_n}))$ where $i = 1....n$ (7)

The obvious requirements in this equation is for the threshold or benchmark inherent in F(…) to be tightly defined as this will act to reduce ambiguity and marginal cases. If this is objectively set by the values of f(…) for the existing included variables then in effect one can ensure that only metrics better correlated with **O** join **D**, so the inclusion of D_i is dependent on the relationship of D_i and **O**.

Clarity around this inclusion strategy, however, opens new possibilities: there is nothing to prevent the development of composite indices (C_i) from a panel of variables ($D_i \ldots D_n$) to feed into a higher-level composite index. In other words, it is possible for $C_i \in \mathbf{D}$ if $f(C_i) > F(C_i | \mathbf{D_1 \ldots D_n})$. This may have particular advantages whereby the 'local' composite, C_i, is able to utilise a local weighting methodology where it is challenging to find a universal weighting approach which would work across the full span of a higher-level composite, allowing a simpler process of aggregation overall.

Principal component methods

In addition to composite indices (C_i) forming a part of **D**, there is nothing to prevent the down-selection from elements of **D** to form a subset to be used to develop either a dashboard or higher-level composite **C** to reduce collinearity. This down-selection could take the form of a principal-component analysis to optimise the weights given to different elements (D_i) of **D**, and hence identifying which D_i is either the best correlate of **O**, or the best representative within the class of D dataseries; i.e. those related to health, or housing etc, in terms of the overall correlation performance of the class. This notion of classes is important in relation to the national Wellbeing Dashboard addressed below.

3.2. Weights

Considering the question of weights compels us to tackle the concept of value, as surely we wish to use a structure whereby those components of greater value (i.e those elements which better correlate with O) are given the greater weight, whether one is designing a composite metric or a dashboard. The necessity is to find a mechanism to expose, across a society, the relative value that people as a collective place on different outcomes or outputs in an unbiased fashion. In effect, we are looking to identify a mechanism to expose people's preferences in a format that is not subject to the importation of a bias or *a priori* opinions from the collecting body or agent. Essentially this presents three options:

- Subjective, albeit generally expert-led processes to derive weights. Obviously, under the definition above these are discarded as being too open to the perception of bias, even if the experts are able to derive an optimised system.
- 'Objective' weights drawn from within the dataset, using '*endogenous'* methods based on statistical methods such as Principal Component Analysis (PCA), as well as frequency-based weights. It should be noted Dutta, Nogales, and Yalonetzy (2021)

demonstrate this may result in the resulting indices violating two fundamental measurement properties: **monotonicity** and **subgroup consistency**. Monotonicity states that if the experience of an individual worsens in any indicator, then the overall experience of the society to which this individual belongs, should not, *ceteris paribus*, improve[10]. Subgroup consistency requires that changes in the overall population should reflect the changes happening at the smaller population subgroup level. For instance, if the outcome of interest in a particular region of a country improves, while all other regions remains unchanged, then subgroup consistency implies that the overall outcome metric in the country should not decrease[11]. In short, it is possible to observe instances where, for example, poverty in a group may increase, but if the weights move by a relatively larger factor towards groups observing constant poverty rates, this can result in the poverty metric appearing to improve. Given these concerns this paper therefore focuses on exogenous methods[12].

- 'Objective' *exogenous* weights which allow a clear line of sight on unbiased preferences. There are two broad methods that reveal people's view either through their actions (revealed preference) or directly reported through some selection methodology, and which can act as a numeraire within our structure. In this section, we will briefly discuss five possible mechanisms: the use of **prices** as weights, the use of **time** as weights , the use of **voting** or **surveys** to report directly, the use of **legislative and regulatory decisions** as a proxy for social preferences, and finally placing the value in terms of a **common numeraire** derived for this purpose such as a quality-adjusted life year (QALY) or wellbeing adjusted life year (WELLBY) (as per Layard and De Neve (2023)) and using the change in these metrics multiplied by the monetary value given to these as the weights.

Prices: Probably the most significant differentiator between economic statisticians and social statisticians is the economist's acceptance of prices as a valid mechanism for exposing revealed preferences on the basis that the 'invisible hand' of the market allows all consumers to express their preferences to reach an optimal outcome where society has maximised wellbeing through

[10] This is a particular instance of the concept of suppression described above – where the framework does not just suppress the signal contained in a particular data series, but through the application of the weighting system derive a perverse outcome where it appears the situation improves whilst it is actually getting worse.
[11] Again, this is the same instance of suppression, but viewed via the geographical axes rather than the individual data series.
[12] Additional research by the same authors suggest hybrid (combining exogenous and endogenous) approaches may also present challenges.

the raising and lowering of prices until no-one's outcome can be improved without it being to the detriment of another. Under this model, prices reflect societal preferences and relative weights in terms of wellbeing. Let us assume people value an increase in 0.1 on a 0-10 SWB scale equivalently, *whatever source it comes from*. Let us assume people have sufficient information to say that 0.1SWBs is worth £x, whatever value we may want to put on x. Let us assume that people's actions are rational and informed by this knowledge. It therefore would appear evident that the average person would use their disposable income to maximise their SWB, and hence if some activity / service / good provides them with 0.2SWBs of benefit, they would be willing to spend £2x to access it.[13]

If we all do this, then those who achieve the most SWBs from any given activity / service / good are most likely to be willing to purchase it[14], noting the impact of the income constraint. The market should therefore act as an efficient mechanism for distributing these activities, services, and goods, after controlling for income. And hence, if one accepts this story than the prices set in the market (let's call them, in the jargon, 'market prices') are independently set and unbiased estimators of relative value, then one can use these as an objective set of weights which fully incorporate the collective wisdom of all of society and delivers a clean mechanism for removing subjective decisions from our policy-relevant composite index proxy for SWB.

Except, of course, we already know there are to key constraints which are likely to suggest a simple linear relationship may be very difficult to reveal. Firstly, SWB is driven by all sorts of factors that are not available on the market and hence, an over-strong reliance on market prices may either exclude or under-value the benefits we all receive.[15] Secondly, it would not be unusual for economists to expect that the activities / services / goods discussed above may reveal diminishing marginal returns to SWB, so our hypothetical 0.2 SWBs of benefit per unit may not be a constant and may fall with time, or rather with existing social welfare. This is likely to be of particular import when considering the public service productivity measurement issues discussed below, although this need not invalidate aggregation across the wider dataset **D**.

We therefore need to reflect on three issues: 1) what to do with those items that don't have a market price, 2) what to do with those items whose market price excludes some part of its value

[13] Subject to the consideration of diminishing marginal returns. Obviously in this context we are forced to rely on revealed average preferences.
[14] Noting that incomes define the ability to purchase, so SWB would not necessarily be optimised within a market where incomes do not correlate with SWB optimisation.
[15] Formal proofs of this are available from, for example the works of Dasgupta and others.

(positive or negative) and 3) what to do with those items who demonstrate strong diminishing returns at quantities we may expect to have to address? There are a number of potential options to consider.

Dasgupta (2021) provides a framework to consider this through the development of what he terms 'accounting prices'[16], under which he looks to add (subtract) the value of externalities to deliver a price that would allow an 'inclusive' accounting model to cover not just the economy but other stocks and flows relating to natural and human capital, which are currently excluded from national accounts, and social capital, which is generally lacking a quantitative measurement framework in money terms. This model raises intriguing possibilities, but is clearly data intense and so it is sensible to consider further options.

Time: Similarly, Coyle and Nakamura (2019) have raised interesting questions about using time, as collected in time-use surveys as a weighting mechanism. This avenue takes time and people's allocation of time as the numeraire whereby people allocate a finite amount of time budget as they allocate a finite monetary budget. This has the added advantage of being more 'democratic' and less 'plutocratic': all citizens have an equal quantity of time to allocate, whereas those with a higher income have greater weight when using prices[17] (.).

Again, however, the same challenge remains as when using prices: how to account for items that affect wellbeing but do not impact people's usage of time [18], either large (climate change), or small (heating your home). Global warming happens 24/7; it may change people's decisions at the margin, but the generally pervasive nature of this phenomenon may mean the scale of impacts is either so widespread or too subtle to be lost amongst a detailed recording of time on primary and secondary activities. Equally, heating your house does not have a time component, unless one reflects the time value of the labour necessary to earn sufficient to meet the price of heating – i.e. converting heating costs into their time equivalent)

Finally, we need to consider the treatment of productivity gains (i.e. technologies which permit individuals to use less time on an activity / undertake two activities at once): if less time is used, does that suggest these activities are now 'less valued'? Intuitively a baseline period

[16] In usual pricing methods, these would change through time and in practice address the diminishing returns point.
[17] Noting that the aggregate weights can be given by the average of budget shares to develop a democratic income measure, for which this does not hold
[18] Noting this can be available in some Time Use Surveys

would need to be identified such that if more activities are undertaken this would be recorded as more value being delivered within a set time budget.

Voting: Democratic societies look to reflect the preferences of the citizenry through the recording of ballots on a periodic cycle to enable a direct revelation of the opinions of all members of society. Whilst this may appear a strong mechanism to extract an understanding of trade-offs, there are a number of challenges, setting aside Arrow's impossibility theorem:

- Timing: Elections can be infrequent or in some countries, non-existent, or corrupted. In such cases, these may serve as a poor measure of public opinion, at the very least in the periods between elections.
- Conflicting viewpoints: Some countries have elections for different tiers of government and how to assess the signal about views when the same electorate elects different candidates with different slates of policies in local or national elections would rely on some form of subjective assessment that one election or another was superior in some fashion and a better representation of people's true views. Recognising people formally vote for individuals not policies, and politicians can change their minds, and indeed there can be time-inconsistency issues at play, how to reflect this would also need to be considered.
- Franchise: Obviously, only those eligible to vote can participate. This normally precludes the young, may preclude those in prison or with certain types of criminal records, and definitely excludes those who are yet to be born. Many elections exclude some combination of foreign nationals, depending on exact political rules around eligibility and residency.
- Slate of policies: Whilst political parties put forward a set of policies, this does not imply the relative importance or weight attached to each or that individual voters may support all the positions taken.
- Variability: Elections can lead to a wholesale change in the dominant position as one government is elected and another defeated. This is probably not reflective of a wholesale change in the value system in a country.

Surveys: Surveys could offer a more timely and nuanced perspective on this topic, but in posing the question to respondents one would need to fall back on some form of relativistic weighting scheme to frame the question, although contingent valuation methods and discrete choice models do exist (Coyle and Manley (2021) do provide methods for tackling this question which merit serious consideration).

Legislative and regulatory decisions: One approach that should, *in theory,* approximate societal weights and priorities are legislative and regulatory decisions – the direct application of societal preferences as revealed through ministerial action. If we take such actions as a proxy for the views of voters, given they voted for the said ministers, exploiting the relative weights in the key performance indicators caught in inspectorate inspection regimes for different services, as set by ministers, may reveal relative preferences within that particular set of variables. This approach is of high relevance for public service productivity below, where it has been used to provide a valid approximation of the relative importance of different metrics. However, it is unlikely this approach would have merits across more than limited and discrete areas of activity.

Derived numeraires: Would it not be better, if wellbeing is the metric of interest, to weigh different drivers of wellbeing by the impact they have on the statistical wellbeing-year (WELLBY), as advocated by Layard and De Neve (2023)? In the UK context, where HM Government has estimated that 1 additional WELLBY point on a 10-point scale equals £10,000-£16,000[19], then one could also have inter-changeability between WELLBYs and monetary terms[20] such that:

$$\Delta WELLBY = \Delta Income\ x\ \delta WELLBY/\delta Income \qquad (8)$$

Such an approach would close the loop more directly between the policy drivers of SWB and SWB itself, but may 'lock-in' effect sizes when these may actually vary across time or space. Nonetheless, this is undoubtedly an approach well worth considering and one which allows different policy initiatives to be valued *either* in monetary or wellbeing units in a fashion that allows wider comparability.

It is, of course worth noting that in the absence of a valid weighting structure, applying a neutral regime of equal weights (Laplace (1814)) may appear to have some advantages in that it does not disproportionately over-weight some aspects through the subjective choices made,

[19] This value is derived from the statistical and economic work relating to a value of a Quality-adjusted life-year (QALY), which gives a value of £60,000 for a year in good health. The WELLBY value is derived from the fact that the difference from a life worth little more than death and good health appeared to be around 5-6 WELLBY points.
[20] The author would like to thank Fabrice Murtin (OECD) for the insightful comments that deriving monetary values such as these 'c*an be tricky due to income measurement issues and attenuation biases.'*

although it obviously has equal and opposite costs, that is it disproportionately over-weights other factors by omission, as the allocation of even weights is in itself a subjective choice, and one which may be deliberately omitting any information from data actually possessed. Such a weighting regime, save in the singular case where all weights are equal, cannot be correct, and given that is the one case where we must presume, if we are debating a need for weights at all, does not exist, the Laplace approach clearly offers nil value: given the definition of the problem, it is clearly a biased approach masquerading as an exercise in unbiased practice.

As such, we can summarise the theoretical considerations as:

- Datasets **D** need to have an underlying intellectual rationale driven by a 'theory of change' where there is a sensible a priori reasoning to believe the inclusion of any specific factor D_i can be justified as there being an unambiguous relationship with the headline matter of inquiry or interest, in this case **O**.
- Weight sets need to objectively reflect societal preferences through a mechanism which can philosophically be shown to have a relationship with societies views of such matters, and which is of relevance to the datasets under consideration – so market prices are clearly valid for the market economy, but may be a poor choice for social and environmental data, specifically where we know market prices omit key information relating to externalities, or indeed there may be no market prices.

Using these two points as guidelines, we can now review the four recent applications described above to consider how these have been applied in these instances.

4. Recent Examples

4.1. Inclusive Income

Gross Inclusive Income (GII) and its sibling Net Inclusive Income (ONS 2023a) are composite indices drawn from the same methodology as the National Accounts use to produce Gross Domestic Product (GDP) and Net National Disposable Income (NNDI). In relation to the two methods issues raised above – the composition of **D** and the weighting method, GII and NII utilise well-established approaches. **D** is drawn from a theoretical framework based around the National Accounts principles of the asset and production boundary. The National Accounts define a set of assets (productive capital) which lie within a specified 'asset boundary'. Alongside these, all the resultant flows which result from human interaction with these assets deliver the sum of flows of output, income and expenditure within the 'production boundary'.

Using this method, the Inclusive Income indices expand the asset boundary to account for human and natural capital, and commensurately expands the production boundary to align to the output arising from this wider class of assets. As these flows of benefits are not all strictly ‘output’ the framework is re-conceptualised to align to the concept of consumption and proxied by equivalent income flows commensurate with the flow of consumption from the results the interaction of these assets with human activity or as part of natural actions.

Importantly these indices, under this regime, form a strong theoretical match for the concept of economic welfare, that is the absolute total of the flows of benefits received from humanity’s interaction with the economy and the other structures and ecosystems, primarily unpaid work and the environment which deliver flows of benefits which can be conceptualised in such a framework. Factors such as subjective wellbeing and purely social factors may not be best considered within such a framework and as such are excluded.

With regards to weighting, the Inclusive Income indices are weighted using market price money metrics. This is not to say that alternative money price metrics could not be considered; as proposed by Dasgupta (2021), accounting prices, loosely defined as shadow prices taking account of externalities, would potentially be a superior alternative weighting mechanism in the Inclusive Income framework, and indeed, Dasgupta would argue a more appropriate one.

The value of this methodology is, whilst dashboards deliver great value by exposing users to many different types of data, bringing this range of information into a single measure is, if appropriately weighted, capable of quickly revealing the trade-offs between components of the headline metric. So, if GDP increases but at a cost to the environment and the services it delivers citizens, this can be seen by the growth rate being lower than that of GDP, and maybe even negative. Equally it can reveal the importance of the market economy within a wider measure of economic welfare.

4.2. National Wellbeing Dashboard

The UK National Wellbeing Dashboard (ONS 2023) was first launched in 2012 and revised in 2023. It currently forms 60 data-series (in our terminology $D_1 \ldots D_{60}$) evenly divided between 10 domains or classes. These are:

- Personal Wellbeing
- Our Relationships
- Health

- What we do
- Where we live
- Personal Finance
- Education and Skills
- Economy
- Governance
- Environment

Each measure is presented independently as a metric in and of its own right with an overarching narrative supplied. There is an ongoing discussion around the complexity of extracting a signal from this number of data series, and so there is consideration of either undertaking a principal agent analysis of each domain to identify the 'lead metric', or the development of a composite index per domain which would allow the presentation of a streamlined set of 10 indicators covering the domains. Clearly such a strategy would benefit from the points raised in this paper.

4.3. Public Service Productivity

In 2003 the UK National Statistician commissioned a review from Sir Tony Atkinson to review the measurement of public services in economic statistics. The Atkinson Review (2005) identified the key issue as the absence of market prices: without a price which could move to reflect the value society places on the quality of a good or service, a simple quantity measure of output may fail to adequately reflect value added: a school which failed to educate children would be counted equally to a school where all the students achieve top marks, or two hospitals delivering identical operations would be scored the same whether in one all the patients were cured and in the other they all died. In the same way that a brickmaker in the private sector can charge a positive price for whole bricks and zero for broken bricks, the need to find a proxy for quality was identified as the key requirement.

Atkinson proposed revising measures of output to account for the value with an adjustment for the change in the quality of the outcomes delivered which could be attributed to the service provision: so, a school which taught 1% fewer students than the year before, but saw them achieve 2% better grades on average would see its quality adjusted output increase by 1%[21]. To estimate, at any time period t, whole public service Productivity (Y_t), a cost weight activity

[21] (ΔQuantity Change + ΔQuality Change) = (-1% + 2%) = 1%

index is used to bring the different services (y_t) together as a composite index as follows, where c representing the cost-weighting attributed to that service:

$$Y_t = c_{1t}(y_{1t}) + c_{2t}(y_{2t}) + c_{3t}(y_{3t}) + \ldots.. \quad (9)$$

Within each individual service (y_i)[22], we can implement a further decomposition, where N represents quantity, and L represents quality:

$$y_{it} = c_{1t}(N_{1t}.L_{1t}) \quad (10)$$

Two issues with the assembly of either y or Y are clearly apparent:

- Firstly, the whole point of this exercise is that the cost of production of these services acts as a poor proxy for the value they create. If this was not the case, there would be no need to apply the quality adjustment L. Therefore, to use cost (c) to weight together the components when *a priori* cost is seen as an insufficient metric is clearly problematic. Whilst having pleasant characteristics of availability, comparability and accuracy, this remains an area of potential improvement. One approach which have been proposed is to incorporate some assessment of the impact on final outcomes into the weighting approach, either through using quality per unit cost weights (L_{nt}/c_{nt}) or the impact on outcomes driven by outputs ($\delta L/\delta N$). An important question here, which is also addressed below in relation to the Health Index is whether the weighting scheme should be independent of the values of the data series contained within **D.** This paper will reflect on this in its conclusions.
- Secondly, within each service, y[23], there is the implicit assumption that for all services the implicit weight on N and L are both 1. That is a one percentage point increase in quality is equivalent to a one percentage point increase in quantity. The question whether this is a valid assumption is an important one. Let us consider an example from education. Let us imagine we have a thousand students in their GCSE year, where GCSE is a terminal qualification from the compulsory schooling system[24]. Let us assume these students have achieved the average level of qualification – as measured in GCSE points – let us say they each attain, on average 40 GCSE points. Let us compare an increase in the cohort size of 1% with the impact of a marginal increase (1%) in the GCSE points attained by the existing cohort.

[22] Equivalent to D in the notation used above.
[23] Which is a form of subsidiary composite
[24] This is no longer strictly the case, but is a suitable hypothetical for the example.

- Cohort growth of 1% = 10 students, each attaining 40 GCSE points = 400 extra GCSE points
- Quality Growth of 1% = 1,000 students each achieving, on average, an extra 0.4 GCSE points = 400 extra GCSE points

At this level this appears comparable and hence a valid assumption, but alternatively if one considers the financial rate of return to GCSEs, the question becomes whether average earnings for 10 extra students is greater or lesser that the marginal wage differential for two cohorts averaging 40 points and 40.4 points multiplied by 1,000 students? If one makes any assumption of diminishing marginal returns to GCSE points (so the 99th is worth little more than the 98th while the 11th is worth considerably more than the 10th), the result may be that the value of such an increase in the cohort size outweighs the value of the marginal benefit of additional GCSE points. This is obviously an empirical question but one which also depends on the question whether the quality of educational services can be measured simply by an indicator of labour market returns, as opposed to educational satisfaction, the benefits to society in terms of value development, or any gains in social capital or intangible heritage capital.

So, how common are diminishing marginal returns in terms of the delivery of public services? Given any form of rationing, which one must assume in some way inherent in the delivery of services free at the point of delivery[25], then one must assume that those cases with the greatest degree of need will receive the service first: the most serious cancer patients will receive treatment before those whose cancer is either at an earlier stage of development, or considered more benign. This strongly suggests that either quality adjustments should receive *lower* weights as quantity increases or that the quality adjustment should itself decrease as the value of the service diminishes. That is, the average quality effect should be higher for the first thousand cancer treatments than for cancer patients 10,001-11,000, under probably unrealistic assumptions that cases are prioritised according to the potential gain in QALYs, as opposed to issues like time on the waiting list. Waiting lists also provide a second reason why we can probably set this concern aside in most cases. At any time, there is a number of cases on a backlog and whilst in a perfect world without a backlog early intervention may be considered optimal, saving the future cost of expensive operations and hence preferable, once one has a backlog the need ethically to address severe cancers is obvious, so it is likely that diminishing

[25] Otherwise demand would be essentially unlimited.

returns are a second-order concern. Having noted this, given the condition of public services in most countries, it is prudent to set this complexity to one side.

Health Index

The ONS Health Index provides an index for the period 2015-2021 compiled from a range of health and healthiness metrics. Its purpose is to '*provide… a picture of health in its broadest sense recognising the importance of health outcomes, risk factors and the social, economic, and environmental drivers to support health to improve now and in the longer term*.'[26]

Normalised to 100 in 2015, with all components normalised to 100 at the same point, the Health Index has three domains: Healthy People, Healthy Lives and Healthy Places, each comprised of a number of sub-domains. With 56 indictors across these three domains, the complete domain map comprises:

- Healthy People
    - Difficulties in Daily Life
    - Mental Health
    - Mortality
    - Personal Wellbeing
    - Physical Health Conditions
- Healthy Lives
    - Behavioural Risk Factors
    - Children and Young People
    - Physiological risk factors
    - Protective Measures
- Healthy Places
    - Access to Green Space
    - Access to Services
    - Crime
    - Economic and working conditions
    - Living Conditions

[26] ONS (2023c)

Drawn from publicly available data, which had a '*reasonable certainty that the data would continue to be available into the future*'[27], with data gaps completed using imputation, the data components used in the Index made '*comparable with other data in the index…[using] … standardisation and normalisation*.' Factor analysis was used to determine which indicators fitted best into which domains and 'to decide what weights indicators should be given, that is, how important they are in measuring health.' This is a key distinction to the normal methodology of preparing a composite index, in that strictly a composite can be described as one where the weighting regime is established on an *a priori* basis, even if the values change, for example as prices adjust. The Health Index, through its use of a principal component approach, allows the weighting prioritisation to change based on the values of the components themselves, and with the ability for weights to change in previous periods as the latest data is taken into account.

Does this matter, or can we still view the Health Index as a 'composite index' despite this difference? This paper argues we should as the value is in comparing and contrasting different methodologies to assembly data to create single-number index outputs. The relative merit of weighting methodologies decided in advance, irrespective of the values contained in the series contained in **D**, and methodologies which allow α to vary with the data values of **D** is a core question.

The OECD Handbook discusses the following core steps in the construction of composite indicators:

Component	*Inclusive Income*	*Public Service Productivity*	*Health Index*
Theoretical framework. A theoretical framework should be developed to provide the basis for the selection and combination of single indicators into a meaningful composite indicator under a fitness-for-purpose principle	*Inclusive Income utilises the conceptual framework underpinning the National Accounts*	*PSP utilises a conceptual framework aligned to National Accounts*	*Health index draws a conceptual framework of factors affecting health. This aligns strongly to existing wellbeing dashboards*
Data selection. Indicators should be selected on the basis	*Data are drawn from pre-existing satellite accounts,*	*Data are acquired for a full range of*	*Factor analysis used to identify*

[27] ONS(2022)

of their analytical soundness, measurability, country coverage, relevance to the phenomenon being measured and relationship to each other. The use of proxy variables should be considered when data are scarce	*each with its own underlying framework and data validation and verification procedures.*	*public services to deliver complete coverage.*	*dataseries for inclusion*
Imputation of missing data. Consideration should be given to different approaches for imputing missing values. Extreme values should be examined as they can become unintended benchmarks.	*Imputation, where required is undertaken within existing national accounts frameworks using triangulation across multiple datasources.*	*Where output data cannot be acquired, the 'inputs=outputs' methodology is used to deliver a proxy for output on the basis cost can proxy for value.*	*Imputation undertaken at local levels before aggregation.*
Multivariate analysis. An exploratory analysis should investigate the overall structure of the indicators, assess the suitability of the data set and explain the methodological choices, e.g. weighting, aggregation	*Indicator structures are underpinned by national accounts fundamentals.*	*The primary challenge in this regard is attribution: if more than one service contributes to an outcome, how should these be weighted across services. Implicitly cost is used to proxy for value.*	*Factor analysis used to identify relationships*
Normalisation. Indicators should be normalised to render them comparable. Attention needs to be paid to extreme values as they may influence subsequent steps in the process of building a composite indicator. Skewed	*Implicit in the use of common market price / exchange value terms.*	*Implicit in the use of common market price / exchange value terms.*	*Applied*

data should also be identified and accounted for.			
Weighting and aggregation. Indicators should be aggregated and weighted according to the underlying theoretical framework. Correlation and compensability issues among indicators need to considered and either be corrected for or treated as features of the phenomenon that need to retained in the analysis	*Weighting uses market price weights, drawn from societal preferences.*	*Weighting uses cost weight in market price weights terms, as a proxy for societal preferences.*	*Weights used based on factor analysis to determine* 'how important they are in measuring health'
Robustness and sensitivity. Analysis should be undertaken to assess the robustness of the composite indicator in terms of, e.g., the mechanism for including or excluding single indicators, the normalisation scheme, the imputation of missing data, the choice of weights and the aggregation method	*The scale and importance of human capital clearly has a driving impact, but this is not a surprise.*	*The assumption that cost can proxy for value is at odds with the underlying conceptual framework (if it did we would not need to derive this composite.*	*Subject to external testing in development.*
Back to the real data. Composite indicators should be transparent and fit to be decomposed into their underlying indicators or values	*Decompositions presented.*	*Decompositions presented.*	*Decompositions presented.*
Links to other variables. Attempts should be made to correlate the composite indicator with other published indicators, as well as to	*Comparisons to GDP and NNDI*	*Strong relationships to public sector inputs and outputs in national accounts*	*Primarily used in health environment*

identify linkages through regressions		*and public sector finance statistics.*	
Presentation and Visualisation. Composite indicators can be visualised or presented in a number of different ways, which can influence their interpretation.	*Growth, levels, decompositions etc*	*Growth, levels, decompositions etc*	*Growth, levels, decompositions etc*

5. Discussion and Conclusions

These three examples provoke a number of questions which are worth consideration, with specific reference to each metric:

Inclusive Income and Health Index

- Timeliness of data

Public Service Productivity.

- Timeliness of data
- Do all the quality adjustments utilised present diminishing marginal returns to scale?
- The impact of Case-Mix
- Quasi-market or regulated prices
- Are *a priori* cost weights a sensible approach?
    - o How to weight preventative services, that is those which diminish demand for other services?
    - o How to weight latent capability?
    - o How to weight hard to measure services?

Timeliness of Data

All three composites struggle with issues of timeliness of input data. Clearly, late delivery of key variables compels one of two responses: firstly significant time lags between publication and the reference period or secondly, the application of some form of nowcasting technology to allow more timely production. Thus, if we are to consider nowcasting, the question of consistency between the methods applied *between* different composites is as valid a question as consistency *within* the specific composite. That is, whilst we need internal consistency within any specific composite to present a meaningful metric, we also need to consider the degree to which methods consistency between, a number of composites is required.?

This leaves one open to two approaches; the first is uniformity; the application of a generic and standardised method which can be applied to a variety of composites (see for example Kapetanios and Papaillas (2023)), or the development of highly tailored and specialised nowcasting methods to apply to each individual composite in an effort to best reflect it particular nature. There is a mid-point on this spectrum where a standard method is applied to tailored data can be considered, which attempts to retain the benefits of both methods as far as possible. The application of the Kapetanios and Papaillas (2023) method allows a more timely production of quality-adjusted quarterly estimates, albeit requiring a full year's worth of non-quality adjusted output data from the previous year. Whilst therefore a significant improvement on the two-year lag inherent in the previous version, the development of additional methods clearly presents an opportunity to side-step this significant hurdle to user engagement and deployment of the metric.

Simultaneously, Inclusive Income has reviewed the capacity to use the Kapetanios and Papaillas (2023), and envisage a feasible deployment if necessary. However, post Covid-19 the accelerated production of actual input data has appeared viable and hence effort has been deployed to utilise this approach, not least because the application of a nowcasting methodology which inherently uses recent data to train the algorithm is extremely challenging when recent data presents a highly significant social and macro-economic event which will inevitably distort underlying relationships the nowcast will rely on.

Diminishing Returns

An obvious question is whether the quality adjustments applied behave like economic variables. In many cases they do, as more of the outcome is delivered it may demonstrate diminishing marginal returns over time. An example in this regard is education and its use of GCSE points / final exam marks as the measure of interest. If a society doubled the number of GCSE points acquired in a cohort would the quality of the outcome have doubled, or rather would we see the rate of return to GCSEs in a competitive market fall, such that the quality adjustment should be intervened in, to account for this? Note, there are several points here: qualifications are not the only output of education so we need to ensure the basket of outcomes is correctly composed. Secondly, grading policy can make a significant difference to marked results, as observed during Covid 19. One response to this may be to down-weight attainment over time, but how to shape such an adjustment is not immediately obvious, however if we wish to upweight some elements for the fact value exceeds costs, it is only fair to down-weight those elements where the inverse holds true. However, thirdly, in a National Accounts landscape, one would expect the weights implicit in chain volume linking to address this sufficiently well, although these quality adjustments are as yet not ncorporated into the UK national accounts.

Quasi-market or regulated prices

There are situations where the government applies a ‘market process’ to a service, predominantly by applying tariffs or fees to a service. In these situations it is only correct to stand back and consider whether these closely enough approximate market price or exchange values to negate the need to create a public service productivity composite, or at least to negate the need for quality adjustments for these services as the prices serve this function sufficiently.

Whilst there are numerous examples we can consider in the UK: including tribunals and courts fees, and the NHS payment scheme, the common factor all appear to share is they represent the *cost* of the service rather than the *value*. In the market sector a price would represent cost plus profit and allow the direct derivation of gross value added. Whilst public service payments of this type do not include a profit component, whilst they are useful in imputing the input value, they do not represent a sufficient mechanism to measure output without the addition of a quality adjustment. Given that there is an additional question of whether the fee applied in some services delivered by the NPISH sector, such as higher education and whether quality adjustments are similarly required there. This is outside the scope of this paper.

A Priori Cost Weights

Logically, if one is willing to accept the argument that cost is a poor enough measure of the value of these services to develop a composite measure including quality adjustment in the first place, the argument that cost is a good enough metric to weight the components of that index is clearly challenging.

It is worth flagging as per Mahony, Vernier and Weale (2024):

> *The use of cost shares is less controversial for inputs as these are generally purchased under competitive conditions. There are likely to be some deviations from this, e.g. the NHS is close to a monopsony for some inputs such as drugs, but arguably this simply offsets the market power of drug companies.*

Taking this argument as given, the concerns above need only relate to output metrics and the attached quality adjustments. Whilst cost weights are relatively easy to access, if they introduce biases into the composite we clearly need consider alternatives. Here again, Mahony, Vernier and Weale (2024) summarise the key considerations:

> *This first is how to construct the weights, here it seems that many NSIs resort to expert judgement, but it is unclear how systematic this is. The second is what overall weight should be on the quality adjustment. If this is one, e.g. as used for dividing between attainment and other adjustments in UK education, it assumes that all quality dimensions are covered which is a very strong assumption. In the Danish healthcare case … the quality adjustment is less than one, to allow for other unknown factors that affect quality. This may then lead to higher jumps in growth rates when new quality adjustments are introduced and a need to revise backwards.*

This opens us to two discreet options: the first is to down-weight the quality adjustment component to prevent the over-claiming of value due to the costs weight approach, and the second is to identify if there is a method to 'value-weight' different services. There are three detailed examples to consider:

- Preventative services
- Latent capability
- Services where it is hard to observe outputs.

Preventative services

Let us imagine a relatively low cost intervention which reduces downstream demand for an expensive upstream service. This might be in relation to fire prevention training to prevent fire service call-outs to fires, family support services which reduce demand for children's social care, tobacco cessation support reducing demand for costly cancer operations or defence spending which deters aggression and hence reduces demand for the defence infrastructure purchased.

In all the cases the cost-weighting approach clearly displays a core weakness: increasing supply of a low-cost service in a one-to-one fashion with reducing demand for a high-cost service will generally reduce quantity output (before quality adjustment) – so one extra low cost output in place of one high cost output will reduce total output, and indeed will do so until the ratio of low-cost to high cost quantities exceeds the ratio of low-cost to high-cost prices.

The clearest example of this occurred in relation to health services during the Covid-19 pandemic when the replacement of high cost operations with medium cost critical care was of a sufficient scale for the NHS to be busier than ever before but also demonstrate a significant downturn in quantity output and hence productivity (ONS 2022a). Whilst this was not a preventative service it is clear within the current composite model the same result would be obtained by investing in a function which causally reduced high cost services as a deliberate policy objective.

This penalisation is obviously ‘unfair’. Let us imagine a complex, expensive, potentially risky and personally uncomfortable procedure as an operation to remove a cancer. Socially it is clearly optimal to prevent the need for this by preventing the appearance of the cancer by helping the patient give up smoking, particularly if this is low-cost. The idea we want our metric of productivity to increase if the patient gets cancer and goes down if we prevent this cannot be correct. The appropriate response should surely be to better reflect the value of the intervention by using the ‘avoided’ cost rather than the actual cost. A simple configuration could be that, for services designated as preventative, the cost weight could be calculated as:

Cost $\text{Weight}_{\text{preventative}} = \text{Cost}_{\text{avoided service}}$.Probability(preventative intervention will reduce the use of the designated avoided service)

So, if we imagine a service which costs £300, but reduces demand for a £40,000 intervention by 5%, the cost weight which would be applied would be £2,000. Weale (forthcoming) will discuss this further .

But this use of a derived weight, driven by a perspective of the resultant value of the service, raises the question: if one should take this approach for preventative services, shouldn’t one also take the same approach with other services? There is a question of data availability here, but there are also special cases, specifically latent capability.

Latent Capability

Let us take a service which deals with infrequent but severe disasters. Covid-19 is an obvious example but the fire service is a more typical one.

The average fire engine is not in use at fire events 24 hours a day, 7 days a week. A large fraction of its time is spent in the fire station, potentially undergoing maintenance or being used for training, but at other times merely lying latent until the next time it is called on. The public, through their taxes, have paid for a level of service which includes response times which relies on some degree of latent capability. Clearly in a period with more fires it would appear that, if one only counted fire events that productivity has risen as the fire engine and the crew are deployed more frequently. Equally, if the fire service delivers an intense programme of fire prevention events and training, such that the number of fire events fell, it could appear the productivity has fallen.[28] Whilst this clearly conflates several of the issues under discussion the issue of what to do if the level of latent capability is felt to need to increase at some point in time. Let us imagine that following Covid-19 the NHS was able to make the case to sustain more empty beds / wards / unutilised staff to better cope with peak demand events of the nature of Covid-19. Our measures would currently show a step-change fall in health productivity, because we have no way of bringing latent capability into the output side, only the input side. This would be clearly problematic, *because the value which it is felt by society latent capability delivers is neither reflected in the costs* <u>*or*</u> *the costs weights.*

One would therefore need to identify a mechanism to apply a value weight within the framework. As with preventative capability, the answer appears to be the application of a probability x avoided cost methodology which one might applying a cost-benefit framework. As with preventative services, one would require a tight definition of eligibility, which is regularly reviewed. One would need an estimate of the avoided cost, and finally a probability that the identified spending allowed one to avoid the 'avoided cost', in the same way that Atkinson (2005) focused on attribution.

In methods terms the treatment of preventative and latent capability appear to simply be two sides of the same coin and hence a common approach may be worth considering.

Hard to measure outputs

Are 'hard to measure' outputs (like policing and defence) hard to measure for this reason, or for other reasons. If we consider they share common characteristics this suggest a common treatment across the three is a credible strategy?

---

[28] There may be changes in the number of fire events which are not caused by the fire service, such as the use of more fire-retardant fabrics.

The answer is sometimes: policing is hard to measure because it is a service which delivers multiple outcomes (solving crime, investigating crime, preventing crime, managing large crowds and special events, road traffic accidents, missing persons, some elements of national security, addressing organised crime etc), and as the Police Activity Survey[29] this means that only around a third of inputs are expended addressed elements of criminality. Hence fighting crime is the largest focus of inputs but it is not the majority. The answer here is accurate input and output measures, by type of service, not necessarily a new approach to cost-weighting.

Defence on the other hand clearly shares characteristics with both preventative and latent capabilities. Defence's primary output, in most circumstance is deterrence capability; i.e. The accumulation of a sufficiently large or advanced latent capability to rain destruction on a foe (primarily considered a state actor) to ensure they do not seek to force the circumstances where the UK would be compelled to do so. Defence therefore clearly fits into the same family with preventative and latent capability and hence it appears logically consistent to consider whether a similar methodology could meet need in this circumstance.

As ever, the question of data availability is central: where could we source data which would allow one to derive a measure of output distinct from the input costs, to reflect the extra value these services deliver?

One practicable approach would be to consider the business cases attached to capital investment by the Ministry of Defence. If we consider these to be labour-augmenting technology (a soldier with a tank is a more significant deterrent than a soldier with a musket), then the cost-benefit appraisals delivered in this context should deliver a first-order approximation of the additional value delivered over and above the cost of acquiring the capital asset. Whilst this approach may implicitly treat labour and intermediate consumption as 'inputs=outputs', by securing our methodology to intensely scrutinised business cases (e.g. the spend on nuclear submarines exceeds £60bn), we access data which can be externally verified and which is computed against a common methodology which is well understood. This approach may also be more widely applicable to capital projects in other services and is worthy of further consideration. Nevertheless, as a data source for preventative and latent capability this appears a sensible and practicable approach based in well-understood economic theory.

Case-Mix

[29] https://www.gov.uk/government/publications/police-survey-2022-initial-analysis

One question which was not explicitly considered in the Atkinson Review was what happens when demand conditions substantively move? Demand from citizens was effectively taken as a given variable and productivity was considered, quite rightly as a supply-side issue.

Nevertheless, economists are quite sufficiently aware that a significant enough change in the patten can cause a firm or service to not just shift along a production possibility frontier, but also to move to a new frontier either inside or outside the previous one. Covid is a classic example, the addition of a significant new demand for critical care services (medium cost) caused the NHS to reduce the volume of high-cost operations and other similar activity. The result of substituting a medium cost service for a high-cost service was a noticeable fall in output and productivity despite inputs and workloads being higher than ever before.

This is easy to understand in the preventative / latent capacity model described above. The probability that a person might die if this preventative / latent capacity was not available jumped, changing critical care from a medium cost / medium value adding service to a medium cost / high value adding service, which hence dominated high cost / high value adding services, like operations.

Conclusions

Aside from the issues relating to accessing timely data, all the big issues discussed above related to whether costs in public services productivity can proxy for value. Clearly in a model explicitly designed because the answer was 'no' we need to be cognisant of this. Cost can be assumed to, at the margin, reflect value under conditions of normal competition, which drive profits down to their normal level. When one is explicitly measuring public and merit goods, this clearly is not the case. The question, in pragmatic terms therefore becomes, for some services, can cost be an acceptable first order approximate to value. For some services, in steady-state, probably yes, at enough for use as weights, but even then if demand conditions shift substantial this needs to be reconsidered.

But there are equally cases where cost cannot be perceived as a first order approximate of value. In those instances identified as falling into this group, an alternative set of weights (in common exchange price terms) are required which deliver that first order approximation. This paper suggests probability adjusted avoided costs, (appropriately discounted) could be explored to serve this experiment, where these avoided costs are represented by the prices of the services avoided where these are available, or their own costs as used elsewhere. This approach would fall within the scope permitted by the European System of Accounts, but Weale (forthcoming) flags that these services would themselves come with impact on outcomes which could also be priced in. His work on diabetes and the impact on Quality-adjusted Life Years (QALYs) showcases exactly how this could be done.

Looking more widely, beyond public service productivity, as a practical test of the more general hypothesis outlined in this paper, alongside the other examples, we clearly see a need for practicably deliverable objective weighting regimes as the key to the production of metrics which meet wider user need and acceptance. Further research is undoubtedly required, but this paper hopes to provide a clear conceptual case for objective weights wherever possible and to provide feasible alternatives. Nevertheless, it is important to consider what researchers constructing composite indices can do to assess robustness and mitigate sensitivity to the weighting method they use. Clearly this paper does not deliver routine tests or code which can be deployed to test for the risks described above, but user engagement is key. Composite indices are generally produced to meet a particular user need and buy-in has generally been hindered by the threat of subjective weights reducing the 'cut-through' of such indices. The core argument in this paper is that objective weights are the key requirement to unlock enhanced use by policy-makers and decision-takers. As such, the key question is whether such indices, using objective weights as proposed, are finding an audience and proving beneficial.

In doing so, it is likely that we will hit points where, if a composite brings together a wide variety of data from different domains, the challenges of using a single objective weighting regime become greater, at least in the eye of the policy-maker. As such, key questions of inter-operability are visible. Can time and prices be treated inter-changeably? Can prices and WELLBYs be used side-by-side. Economists would instinctively say yes: convincing the wider community relies on strong worked examples which mirror intuition and expectations about the relationship between the resultant **C** and **O.** There is no doubt there will need to be some trial and error to reach accepted resolutions.

Bibliography

Atkinson (2005) 'The Atkinson Review: Final Report: Measurement of Government Output and Productivity for the National Accounts' Available at https://webarchive.nationalarchives.gov.uk/ukgwa/20160106003117/http:/www.ons.gov.uk/ons/guide-method/method-quality/specific/public-sector-methodology/articles/atkinson-review-final-report.pdf

Coyle and Nakamura (2019) 'Towards a Framework for Time Use, Welfare and Household-centric Economic Measurement' (ESCoE DP 2019-01). Available at https://www.escoe.ac.uk/publications/towards-a-framework-for-time-use-welfare-and-household-centric-economic-measurement/

Dasgupta (2021) 'Final Report - The Economics of Biodiversity: The Dasgupta Review' Available at https://www.gov.uk/government/publications/final-report-the-economics-of-biodiversity-the-dasgupta-review

Dutta, I., Nogales R., and Yalonetzky G. (2021) 'Endogenous weights and multidimensional poverty: A cautionary tale' Journal of Development Economics 151 (2021)

Dutta, I., Nogales R., and Yalonetzky G.(forthcoming) 'Multidimensional Poverty Indices: How Much Should We Listen to Experts?' Working Paper

Kapetanios, G., & Papailias, F. (2023) 'Nowcasting Unpaid Production Activity and Quality Adjustments in Public Service Productivity' (ESCoE DP 2023-20), Available at https://www.escoe.ac.uk/publications/nowcasting-unpaid-production-activity-and-quality-adjustments-in-public-service-productivity/

Laplace, P.-S. Marquis de (1814) 'Essai philosophique sur les probabilites'

Layard, R., and De Neve, E. (2023) 'Wellbeing: Science and Policy' Cambridge University Press

OECD (2008): 'Handbook on Constructing Composite Indicators' Available at https://www.oecd.org/sdd/42495745.pdf

ONS (2022) 'Health index contents and definitions' Available at https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandwellbeing/methodologies/healthindexindicatorsanddefinitions

ONS(2022a) ‘International comparisons of the measurement of non-market output during the coronavirus (COVID-19) pandemic’. Available at: https://www.ons.gov.uk/economy/economicoutputandproductivity/output/methodologies/internationalcomparisonsofthemeasurementofnonmarketoutputduringthecoronaviruscovid19pandemic

ONS (2023) ‘National Wellbeing Dashboard’ https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/ukmeasuresofnationalwellbeing/dashboard

ONS (2023a) ‘Inclusive income; 2005 to 2019’ Available at https://www.ons.gov.uk/economy/economicoutputandproductivity/output/articles/ukinclusiveincome/2005to2019

ONS (2023c) ‘Health Index methods and development: 2015 to 2021’. Available at https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandwellbeing/methodologies/healthindexmethodsanddevelopment2015to2019

O’Mahony, M., Verniers, M., & Weale, M. (2024): ‘Review of Approaches to the Measurement of Public Sector Productivity Statistics’ Economic Statistics Centre of Excellence, King’s College. London

Rosen, R (1991) ‘Life Itself: A Comprehensive Inquiry into Nature, Origin, and Fabrication of Life’ Columbia University Press

Sharpe, A. (2004) ‘Literature Review of Frameworks for Macro-indicators# Centre for the Study of Living Standards, Ottawa, Canada

UNECE (2019): ‘Guidelines on producing leading, composite and sentiment indicators’ Available at https://unstats.un.org/unsd/nationalaccount/docs/ECECESSTAT20192.pdf