Pairwise Comparisons as a Scale Development Tool for Composite Measures

Ginevra Floridi[1, 2] & Benjamin Lauderdale[3]

[1]*Nuffield College and Leverhulme Centre for Demographic Science, University of Oxford*

[2]*King's College London*

[3]*University College London*

[Accepted manuscript version]

Summary

Composite scales are widely used for measuring aggregate social science concepts. These often consist of linear indices obtained as the weighted sum of a set of relevant indicators. However, selecting coefficients (or weights) that reflect the substantive importance of each indicator towards the concept of interest is a difficult task. We propose a method for the generation of linear indices for aggregate concepts based on pairwise comparisons. Specifically, we ask a group of subject-matter experts to perform a series of pairwise comparisons, with respect to the concept of interest, between profiles displaying different combinations of indicators. This allows us to estimate coefficients for each indicator that provide a linear approximation to how experts make the pairwise evaluations. As we show, the method makes it straightforward to assess intercoder reliability, while being a more accessible task than directly asking experts for coefficients. We demonstrate our method with an application to the concept of "productive ageing", including a cross-cultural comparison of weighting schemes derived from group of Italian and a group of South Korean experts on this concept.

Keywords: Composite index; Linear index; Weighting scheme; Weighting and aggregation; Robustness; Content validity.

Address for correspondence: Ginevra Floridi, Nuffield College, 1 New Road, OX1 1NF Oxford, United Kingdom. **Email:** ginevra.floridi@nuffield.ox.ac.uk

1. Introduction

Composite measures are widely used in social science research to quantify and analyse concepts that are aggregate summaries of multiple indicators, and which cannot be captured by studying their component attributes separately (Greco et al., 2019; OECD, 2008). The most common type of composite measure is a *linear index* m_i for units i, where the aggregate summary is a linear function of continuous and/or categorical indicators I_{ij}.

$$m_i = \sum_j b_j \cdot I_{ij}$$

Sometimes the indicators I_{ij} are standardized and the coefficients b_j are specified to be positive and sum to 1. In such cases, the coefficients are described as *weights*, as their magnitude then reflects the relative contribution that each individual indicator has towards the construction of the index m_i for a set of units i and indicators j (OECD, 2008). Composite scales are used to measure a wide variety of concepts for a wide range of different types of units. Country-level indices of this type include the Fragile States Index (FFP, 2017) and the Global Peace Index (IEP, 2020), while the Index of Multiple Deprivation is a sub-national example (Dibben et al., 2007). At the individual level, examples of composite measures include the Beck Anxiety Inventory (Steer & Beck, 1997) and the Protean Career Index (Baruch, 2014). The vast majority of such measures take linear form, as strong theoretical arguments for non-linear aggregation of indicators are seldom provided by the creators of these measurement strategies.

The generation of a composite scale involves, first, selecting indicators of the concept to be measured, and validating those indicators based on their relevance for that concept (Boateng et al., 2018). This form of "content validation" is usually done by consulting experts, i.e. people with expertise in the concept (Boateng et al., 2018; Hardesty & Bearden, 2004). Once a set of items is selected for inclusion, the main task is to aggregate items, typically in the additive form shown above (OECD, 2008). Aggregation requires assigning coefficients b_j to each indicator that express their relative *partial* associations with the concept, as the linear form implicitly holds constant the values of other indicators. The choice of these weighting schemes is a delicate problem as, depending on the context, there may be important ethical or normative implications in the choice of a set of coefficients over another (Permanyer, 2011).

Researchers often rely on data-driven weighting schemes, using techniques such as principal components analysis (PCA) and factor analysis (FA) (Greco et al., 2019). Since these are based on the amount of (co-)variation that each indicator explains in the data, data-driven approaches are most appropriate where the concept of interest is a latent property of the subjects under study (e.g., the health status of individuals) that plausibly is the strongest common causal determinant of the indicators. However, for many applications this is implausible, and we should instead approach the problem as one of specifying a *normative weighting scheme.*

Normative weighting schemes, based on decisions about the coefficients to be assigned to each indicator, are most appropriate in cases where the concept to be measured exists in the minds of analysts as a summary of the indicators, rather than in the data generating process for those indicators. For example, the extent of democracy in a country is reasonably understood as a summary of the presence or absence of a set of institutional features, not as a cause of that presence or absence. Common normative weighting schemes include the “equal weighting” approach, where all indicators are first standardised and then receive the same weight, or the grouping of indicators into equally weighted sub-scales, which gives individual indicators weights inversely proportional to the number of indicators in their sub-scale. Such weighting decisions are implicit and often arbitrary, which makes it difficult to know whether the weights adequately reflect the relative importance of each indicator towards the concept of interest. More rarely, normative schemes are based on explicit decisions about coefficient values or weights by researchers, stakeholders or relevant experts (Hoskins & Mascherini, 2009; IEP, 2020; Saaty, 1977). Such schemes can help in the generation of valid scales, as they make the subjectivity behind the weighting process explicit. However, they also exert significant cognitive demands on the decision makers, and may become unmanageable as the number of indicators increases (Greco et al., 2019). It is particularly difficult to directly assign coefficients where multiple indicators convey common information, and thus the partial associations of the indicators with the concept of interest given a chosen set of indicators are potentially very different from the unconditional associations of each indicator and the concept of interest. Such approaches may lead to inconsistent or biased results in cases where the participatory audience does not clearly understand the supervision framework or how their coefficient specifications interact in the presence of indicator collinearity (OECD, 2008).

The stakes of choosing appropriate coefficients or weights are whether the scaling or ranking of units with respect to the concept of interest might be substantially altered by adopting different weighting schemes (Permanyer, 2011). It is therefore considered good practice to assess the robustness of any scale by estimating the extent to which the ranking of measured objects that arises from the choice of a given set of coefficients or weights is sensitive to variations in its values (Greco et al., 2019; Permanyer, 2011). This is sometimes done through pairwise comparisons of the objects to be measured with respect to the concept of interest (e.g. IEP, 2020).

In this paper, we propose a method for deriving indicator coefficients towards the construction of a composite scale that inverts this “robustness check” procedure of checking the ranking of units by instead taking pairwise comparisons of units as its starting point. Starting from a pre-established set of relevant indicators for a concept of interest, we run a survey experiment on experts. This takes the form of a series of pairwise comparisons between profiles displaying different combinations of items with respect to the concept. The experiment is designed to enable us to estimate the coefficients that best approximate the relative value attached by experts to each indicator when making pairwise evaluations. The resulting scale is straightforward to assess for intercoder reliability using multilevel modelling. Compared to “implicit” normative weighting schemes such as the equal weighting approach, it maximises validity with respect to the concept to be measured; compared to more “explicit” normative schemes that directly ask experts about numerical weights, it considerably facilitates the decision-making task. In the next section of the paper, we discuss the necessary assumptions as well as design considerations for the experiment. In the following section, we illustrate our approach with an application to the concept of “productive ageing” and make comparisons to commonly used approaches to measuring this concept. We conclude by discussing potential applications as well as the limitations of our method.

2. Theoretical background

Many social science concepts are pragmatically defined as summaries of relevant indicators rather than as representations of underlying quantities that cause a set of indicators to vary together. An example of this is the extent to which a country is democratic, which is more naturally understood as a summary of institutions and practices in that country with respect to the concept of "democracy", rather than a latent variable of countries which is part of the causal process that generates patterns of institutions and practices across different countries (Coppedge et al., 2011). For such pragmatically defined concepts, normative weighting schemes relying on participatory approaches are most appropriate, because they allow for an expression of the relative importance of the indicators from a scholarly, societal or other normative viewpoint (Saisana et al., 2005). This gives rise to the issue of who should be making normative judgements about the relative contributions of indicators towards the construction of a measured scale. In principle, the participatory audience should be part of a community of stakeholders with a legitimate interest in the concept (Saisana et al., 2005). These may include experts, policymakers, or the general public (Greco et al., 2019).

In cases where a novel composite measure is being generated, decisions about content validity for the inclusion of items in the scale are typically made by a panel of experts (Boateng et al., 2018; Hardesty & Bearden, 2004; Lawshe, 1975). One procedure for doing this is to have a group of five to ten experts rate each item on a scale from "extremely relevant" to "not relevant" for the concept. Items that do not reach a minimum level of agreement among experts about their relevance are modified or eliminated from the scale (Lawshe, 1975). Examples range from the Fragile States Index (FSI, 2020) to the Protean Career Index (Baruch, 2014) to indices of pain in preschool children (Suraseranivongse et al., 2001). Reliance on experts is particularly helpful when the concept is based on a theoretical framework that is not commonly known to people outside the field of study. A history of publications or similar evidence of relevant research on the phenomenon of interest are typically used as criteria for the selection of relevant experts (Grant & Davis, 1997). If the subject-matter experts are perceived as true experts, then it is unlikely that there is a higher authority to challenge the content validity of the measure (Lawshe, 1975). By the same reasoning, the validity of a normative weighting scheme is maximised by reliance on a panel of experts for the derivation of weights (Saisana et al., 2005). Of course, for any social science concept, relevant stakeholders may also be represented by the subjects under study, as in the case of employees in a job performance scale (Lawshe, 1975). For instance, Dibben and colleagues (2007) administer a conjoint experiment to a sample of English residents to help inform the specification of weights in the Index of Multiple Deprivation for England, although they do not use coefficients directly estimated from the experimental data.

As outlined below, our application relates to an academic social science concept. Thus, in order to maximise validity in the linear index, we rely on a panel of academics with expertise in that concept. However, given the straightforward nature of the participatory task we propose for the elicitation of coefficients, the method illustrated here can easily be applied to situations where individuals with no specific expertise in the concept of interest (e.g., the general public) are to be involved in defining how a target concept is measured.

2.1 Target concepts and linear indices

The class of measurement problems we consider are those in which analysts aim to measure a target concept μ_i for a set of units i as a composite measure of already measured indicators I_{ij}, where each j is a different indicator of the target concept. The general form of a *linear index*, the most common type of composite measure, is

$$m_i = \sum\nolimits_j b_j \cdot I_{ij}$$

The goal of the measurement exercise is to specify the b_j such that the calculated measures m_i approximate the target concept μ_i as well as possible against a relevant criterion such as mean square error ($MSE = \frac{1}{n}\sum_{i=1}^{n}(m_i - \mu_i)^2$). If the μ_i were observed data, the b_j could simply be estimated using a linear regression predicting the μ_i using the I_{ij} and using the fitted values as the measures: $m_i = \hat{\mu}_i$. Here we are considering applications where we lack this kind of 'training data' to learn the best linear approximation to the relationship between the indicators I_{ij} and the target concept μ_i that we wish to measure. In such applications, the most common approach is for the analyst and/or other experts to directly specify the numerical values of all b_j. This is not a straightforward task for reasons that become clear if one contemplates the structure of the linear index.

The indicators I_{ij} are potentially a mix of different interval level quantities and/or binary indicators for levels of categorical variables. We note a terminological difficulty that must be navigated to discuss this class of measurement problems clearly: the multiple uses of the word "unit". We need to refer to "units" in the sense of the entities about which we want to measure something (in our application below, the older adults for whom we want to measure a level of "productive ageing"). But we also need to use "units" in the dimensional analysis sense: what dimensions do the coefficients b_j need to have such that we can construct an additive index from a set of indicators I_{ij} which each describe different kinds of quantities? The indicators in most linear indices are not commensurable: I_{i1} might be a binary indicator variable while I_{i2} might be in £s, I_{i3} a number of people, and so on. Since there is no context in which one can add a number of £s to a number of people to a binary indicator derived from a categorical survey response, it is worth beginning by noting the circumstances under which this form of linear index describes an internally consistent mathematical calculation. To avoid confusion, we will refer to this sense of "units" as "dimensional units".

In order for the summation to be a valid operation, producing an interval level measure m_i, dimensional analysis dictates that either the indicators I_{ij} already have common dimensional units that match the dimensional units of m_i and the b_j are dimensionless weights; or that each coefficient b_j has dimensional units equal to dimensional units of m_i divided by dimensional units of I_{ij}. The former case is sometimes achieved by standardising the I_{ij} to all have a pseudo-common scale of "standard deviations" or otherwise renormalising all indicators onto scales (e.g., 0 to 100) that are assumed to be commensurable. The latter case is akin to specifying elasticities or coefficients that indicate the number of units of the measured quantity m_i

associated with a one-unit change in each indicator, holding constant all possible values of the other indicators.

This language is of course familiar from the linear regression model, because it arises from the logic of this kind of linear function. It also highlights the difficulty of having experts directly specify the coefficients. Not only does each coefficient potentially take on continuously varying values, but if our goal is to best approximate the (as yet unmeasured) target concept μ_i, the optimal value of the coefficient for each indicator depends on which other indicators are included in the index as well as their coefficient values.

Our contribution in this paper is to illustrate how giving experts pairwise comparisons provides data which enable us to estimate the b_j of our index as expert-specific parameters β_j using a regression model. Instead of requiring the analyst/experts to directly specify continuously varying b_j that all depend on one another, it only requires simple comparative evaluations of the target concept μ that can be completed one at a time without reference to one another. We further discuss the relative merits of having experts directly specify the b_j versus our proposed approach later in the paper.

2.2. Single expert

The pairwise comparison tasks we consider consist of asking an expert to evaluate whether the value of the target concept μ is greater for observed unit $i = A$ or for observed unit $i = B$, where A and B are described only in terms of their indicator values. This is equivalent to asking the expert to report the sign of the difference in the levels of the target: $\mu_A - \mu_B$. The pairwise comparison task given to the experts may either ask the expert for a binary response as to whether $\mu_A < \mu_B$ or $\mu_A > \mu_B$, or (as in our application reported later) an ordered ternary response Y_{AB} as to whether $\mu_A < \mu_B$ $(Y_{AB} = 0)$, $\mu_A \approx \mu_B$ $(Y_{AB} = 1)$, or $\mu_A > \mu_B$ $(Y_{AB} = 2)$.

In the case where the expert is asked for a binary response over many such comparisons, it would be appropriate to use a conditional logit model to analyse the resulting data (McFadden, 1973). This is an example of the varying alternative choice problem for which that model was derived. However, that model does not allow for ties, which are a feature of our application. The “exploded logit” with ties (Allison & Christakis, 1994) allows for all three observed ranking patterns of two alternatives A and B, but assumes that the tied ranking arises from an ignorable failure to report the true ranking, rather than evidence that units A and B are similar. The use of this model here would be equivalent to dropping all tied responses, which is undesirable.

Instead, we follow the same distributional assumptions used in the conditional and exploded logit family of choice models, but add the assumption that the expert is providing a useful signal of approximate equivalence, using unknown thresholds for what level of difference between units A and B is detectable. This yields a choice model that can be estimated as an ordinal logistic regression, but with the explanatory variables describing varying alternatives rather than varying respondents.

We model the expert’s choice process in evaluating the relative levels of the target concept using the form of the linear index. The expert’s evaluation of the level of the target concept for unit i is assumed to be a linear function of the indicator values presented to them plus an error ε_i:

$$\mu_i = \sum_j \beta_j \cdot I_{ij} + \varepsilon_i$$

When the expert is then asked to compare the indicator profiles of two units A and B, they are being asked to evaluate μ_A versus μ_B. Their choice between the two will then be determined by the difference in their perception of the relative levels of the target concept for the two units:

$$\mu_A - \mu_B = \sum_j \beta_j \cdot \left(I_{Aj} - I_{Bj}\right) + \varepsilon_A - \varepsilon_B$$

If we assume, following the derivation of the conditional logit model (McFadden, 1973), that the ε_i have a Type-I extreme value distribution,

$$f(\varepsilon_i) = exp\left(-\varepsilon_i - exp(-\varepsilon_i)\right)$$

then the difference $\varepsilon_{AB} = \varepsilon_A - \varepsilon_B$ has a logistic distribution. Finally, we assume that the observable choice depends only on the expert's value of $\mu_A - \mu_B$, such that the expert reports $\mu_A < \mu_B$ $(Y_{AB} = 0)$, $\mu_A \approx \mu_B$ $(Y_{AB} = 1)$, or $\mu_A > \mu_B$ $(Y_{AB} = 2)$ according to the following definition:

$$Y_{AB} = 0 \; if \; (\mu_A - \mu_B) \leq \alpha_1$$

$$Y_{AB} = 1 \; if \; (\mu_A - \mu_B) > \alpha_1 \; and \; (\mu_A - \mu_B) \leq \alpha_2$$

$$Y_{AB} = 2 \; if \; (\mu_A - \mu_B) > \alpha_2$$

Because ε_{AB} follows a logistic distribution, the above assumptions describe an ordered logistic regression, where the observed outcome is the expert's pairwise comparison response and the explanatory variables are the differences $I_{Aj} - I_{Bj}$ between the indicator values of the two units A and B that are being compared, for each of the indicators entering the linear index.

The measure of interest is defined for any unit i by the original linear form:

$$\widehat{m}_i = \sum_j \hat{\beta}_j \cdot I_{ij}$$

One assumption of this procedure is that experts' assessments follow the linear aggregation model. To the extent that they do not, the procedure will estimate a linear index that approximates how those experts non-linearly aggregate indicator values across the range of indicators presented in the experiment. Another assumption of this procedure is that experts generate noisy responses, rather than perfectly reporting the sign of $\sum_j \beta_j \cdot \left(I_{Aj} - I_{Bj}\right)$. If experts in fact perfectly follow the linear aggregation model, estimation of the logistic regression may require regularisation to point-identify the coefficients and to avoid complete separation.

2.3. Multiple experts

For most social science concepts, it is implausible that two experts will apply exactly the same coefficient values when aggregating different indicators towards measuring a target concept. Thus, variation in coefficient vectors associated with different coders is both inevitable and often an object of study in itself (Greco et al., 2018). In many applications, differences between the average coefficients assigned by subgroups of coders (e.g. experts from different countries or stakeholders with different objectives) may be of substantive interest, as in the case where one wishes to assess how the relative importance of various indicators towards the operationalisation of a concept varies across contexts.

While a single expert responding to pairwise comparisons might plausibly follow the data generating process described above, or some approximation thereof, a set of experts drawn from some population will produce response patterns that are best approximated by different sets of coefficients β_j, even in the large sample limit of each coder evaluating many pairwise comparisons. Thus, it is natural to extend the ordered logistic regression described above to a multilevel model with random coefficients β_{kj} for each expert k. Using such a model, it is possible to either construct fitted values using the random coefficients for a particular expert, yielding an expert-specific measure that reflects that expert's judgments about how to aggregate the indicators, or to construct a measure based on the average coefficient that reflects a *consensus* regarding how the indicators ought to be aggregated. There are other possible variations as well, such as using interactions to define groups of experts with (potentially) different average coefficient values based on observable attributes of the experts. As in the case of our application below, we may be interested not only in the average expert evaluation of these comparisons, but also the degree of variation in those evaluations and whether disagreements reflect observable features of the experts.

2.4. Design Considerations

In designing the elicitation of expert views, it is important to keep in mind several design considerations, which depend on the goal of the exercise and what kinds of validity tests are important for the application. The simplest applications of the methodology we propose are those where one wants to precisely estimate the linear index that best approximates a single expert's implicit aggregation of the indicators. In such an instance, the data collection requires a large number of comparisons from that expert, so that the $\hat{\beta}_j$, and thus also the fitted values $\widehat{m}_i$, can be precisely estimated. Standard regression methods provide the estimation precision for both the coefficients and fitted values.

A major advantage of the approach comes where it is possible to use multiple experts, rather than relying on a single expert, as this enables the measurement exercise to pool expertise from multiple experts as well as assessing the extent to which those experts (dis)agree. In such cases, where one wants to precisely estimate the average across a population of experts, the number of comparisons that one needs to run per expert will depend on the variation in the coefficients implicitly used by the experts in their evaluations. The less the β_{kj} vary across experts k, the closer the data requirements are to the single expert case. It does not take very many experts to confirm that those experts are all following a very similar linear model, if that is in fact the case. The more the β_{kj} vary across experts k, the more data from more experts are required to precisely estimate the average expert from the broader population of similarly selected experts. Widely divergent experts limit the extent to which the multilevel model can take advantage of partial pooling to identify the average coefficient across the population of experts.

While these data requirements cannot be known ex-ante because they depend on the parameters to be estimated, it is still always the case that, with a sufficient number of responses per expert and a sufficient number of experts, we can construct expert-specific indices as well as describing the population of these. Understanding the extent of expert consensus about a social science concept is often valuable as an object of study in itself, in addition to providing a validation check on the intercoder reliability of the scale that has been derived in comparison to the scales that might have been derived using the opinions of another set of similarly selected experts or coders.

3. Application

3.1. Productive ageing

Productive ageing is defined as older adults' participation in activities that produce goods and services, or develop other people's capacity to do so, whether for pay or not (Bass & Caro, 2001). In this application we adopt a commonly used operationalisation of productive ageing as participation in activities that have economic value or, equivalently, that would have to be paid for if older adults did not perform them (Morrow-Howell et al., 2001). We thus consider four activity domains: paid work; volunteer work for charities, religious or political organisations; grandchild care; and informal care or household help to adults, including family members, friends and neighbours. While alternative definitions exist that consider broader sets of activities (e.g. Fernández-Ballesteros et al., 2011), narrow definitions are more widely used as they facilitate comparison and replication (Morrow-Howell et al., 2001).

Productive ageing represents an ideal application for the method developed here, for four main reasons. First, productive ageing is a pragmatically defined concept, in the sense that it is defined by researchers to summarise observed data, rather than intended to represent a latent variable that is causally generating correlations among a set of indicators. This is evident from the fact that correlations in time spent working, in informal care and in volunteering will necessarily reflect older adults' time allocation between different activities, and they may therefore be negative even though, theoretically, all activities positively indicate productive ageing, the concept of interest.

Second, productive ageing is a well-established concept, for which relevant indicators have been identified, are included in many ageing surveys, and are broadly accepted by those who study the concept (Hank, 2011; Morrow-Howell et al., 2001; Strauss & Trommer, 2018). This allows us to "skip" the important step of selecting items for the purposes of illustrating this application, and to focus on the weighting of such items, which is of primary interest for this study. We note that, were we measuring a new concept, some form of content validation of the items would need to be performed beforehand (Boateng et al., 2018). For this application, all the relevant indicators of "productive ageing" (paid work; volunteering; grandchild care; informal care) are available in cross-national comparative ageing surveys such as the Health and Retirement Study (HRS) in the United States (US); the Survey of Health, Ageing and Retirement in Europe (SHARE); the English Longitudinal Study of Ageing (ELSA); and the

Korean Longitudinal Study of Ageing (KLoSA). We use two of these datasets for the generation of our scale.

Third, despite the wide use of the concept in social demography and gerontology, the measurement of productive ageing so far has largely been based on weighting schemes with little or no theoretical foundation. Among the empirical studies that measure the concept by aggregating indicators of the four activity domains, by far the most common approaches consist in summing up the number of activities (Baker et al., 2005) or the number of hours (Herzog et al., 1989; Loh & Kendig, 2013) of productive involvement. Other studies have built productive ageing indices that rank subjects based on type, diversity and frequency of participation (Glass et al., 1999). Still, no attempt is made to assign a value to each activity and, as a general problem with these types of aggregations, individuals with very different forms and intensities of involvement end up being clustered together in the same group or percentile of the distribution. A way of aggregating activity indicators that explicitly gives a relative weight to each of them is to assign activities a monetary value by estimating equivalent wages for each activity, or the amount of money that would be needed to purchase equivalent services on the market (Fernández-Ballesteros et al., 2011; Herzog & Morgan, 1992). However, such methods have been criticised as inappropriate on the grounds that older people's participation has value beyond monetary terms, which has limited their use in practice (Morrow-Howell et al., 2001). Alternatively, some researchers have adopted data-driven measurement methods, which avoid difficult measurement questions by "letting the data decide". For example, Paúl, Ribeiro and Texeira (2012) make use of PCA to identify and aggregate indicators of "active ageing" in Portugal. However, as we have argued, there is no reason to expect that the weights derived from these methods will result in an adequate measure for pragmatically defined concepts like productive (or active) ageing. In our empirical application below, we show how badly they can go awry. Partly as a result of the difficulties in weighting and aggregation towards a "productive ageing" index, most empirical research on the topic so far has resorted to analysing activities as separate dependent or independent variables, thus avoiding the measurement problem by giving up on directly studying the concept of interest (Hank, 2011; Hinterlong et al., 2007).

Fourth, productive ageing makes for an interesting application of our method because the extent to which different activities are considered "productive" – i.e., the coefficient on each activity indicator – is likely to vary across contexts depending on factors such as social policies and cultural norms around families and intergenerational relations (Chen et al., 2016). There is the possibility that relative rankings of adults with identical activity profiles in terms of "productivity" might differ across countries, making it difficult to compare the level of productive ageing across societies (Chen et al., 2016). The method we develop in this study can be used to compare operationalisations of a concept across coders and groups of coders, which we demonstrate by comparing evaluations of productive ageing between a group of Italian and a group of South Korean academics. Italy and South Korea make good cases for comparison. On the one hand, in both countries, productive ageing is topical in light of demographic ageing (OECD, 2017; Rouzet et al., 2019). On the other hand, the academic discourse on productive ageing in Italy focuses on the role of older adults in increasing the productive capacity of their family members, for instance by facilitating young mothers' labour force participation through grandchild care (Arpino et al., 2014; Bratti et al., 2018). In Korea, the growth-oriented policy focus, combined with patriarchal cultural values around the family, imply that unpaid family care may not be considered a socially recognised productive accomplishment, and that conceptualisations of productivity may focus more strongly on activities performed outside the household (Lee & Lee, 2014).

Our main purpose in this application is to demonstrate the method we propose. At the same time, this application represents a useful "proof of concept" for productive ageing research. As argued above, given that productive ageing is pragmatically defined as a summary of indicators, the validity of a productive ageing scale is maximised when relying on normative judgements to assign the indicator weights (Saisana et al., 2005). Therefore, our expert-derived scales provide a useful benchmark against which we test commonly used scales in productive ageing research using the same indicator data.

3.2. Data

The first step for data collection was the generation of "productivity profiles" of older adults participating to different extents in paid work, volunteering, grandchild care and help or care to sick or disabled adults. We took the data for the generation of profiles from the KLoSA (http://survey.keis.or.kr/eng/klosa/klosa01.jsp) and from the Italian sample of the SHARE (http://www.share-project.org/) at baseline. These surveys contain information on various socio-demographic characteristics of older people in each country, and also include modules on respondents' participation in different productive roles. The target population of KLoSA at baseline consists of individuals aged 45 and above in 2006, excluding younger spouses and people living in institutions (KEIS, 2014). The first wave of SHARE targets all Italians aged 50 and above and not living in an institution in 2004, and their spouses regardless of age (Börsch-Supan & Jurges, 2005). We restricted our samples to respondents in both surveys aged 50 and above at baseline, excluding younger spouses. KLoSA has a sample size of 10,248 individuals, while the Italian SHARE sample consists of 2,558 respondents.

KLoSA and SHARE contain similar information on respondents' participation in paid work, volunteering for charities, religious and political organisations, provision of care to grandchildren, and provision of informal care or household help to adults. However, the two surveys differ in how frequency of participation in each activity is categorised. In KLoSA, paid work, grandchild care and informal care are measured in self-reported hours per week, and frequency of volunteering is measured on a scale from "nearly every day" to "never". In SHARE, by contrast, only paid work is measured in weekly hours, and all other activities are measured using frequency scales. Table 1 shows our categorisation of frequencies for each activity, separately by survey. Based on these categories, we derived two separate coding tasks, one using the KLoSA categories and the other one using the SHARE categories.

We used the Shiny package in R to build an interactive web application that presents coders with a comparison of two profiles of older adults, A and B, described by their frequency of participation in each of the four productive activities under study. The profiles A and B were sampled with equal probability from the set of unique observed profiles. For each pair, the coder is asked to select whether 'A is more productive than B', 'A and B are similarly productive', or 'B is more productive than A' based on A's and B's productivity profiles. The coder’s selection and the activity profiles of both individuals in the pair are then saved as an observation in our dataset.

Table 1. Frequency categories for each activity in the KLoSA and SHARE tasks

	KLoSA	**SHARE**
Paid work	Never 1-10 hours/week 11-20 hours/week 21-30 hours/week 31-40 hours / week More than 40 hours/ week	Never 1-10 hours/week 11-20 hours/week 21-30 hours/week 31-40 hours / week More than 40 hours/ week
Volunteer for charities, religious or political organisation	Never Less than once per month 1-3 times per month 1-3 times per week Nearly every day	Never Less than once a week Once or twice a week About every day
Grandchild care	Never 1-10 hours/week 11-20 hours/week 21-30 hours/week 31-40 hours / week More than 40 hours/ week	Never Less than once a month Once or twice a month Once or twice a week About every day
Informal care or help to sick or disabled adults	Never 1-10 hours/week 11-20 hours/week 21-30 hours/week 31-40 hours / week More than 40 hours/ week	Never Less than once a month Once or twice a month Once or twice a week About every day

We collected data from five Korean and six Italian academics, who are described in anonymised form in Table 2. We recruited experts by initially contacting academics whose curriculum vitae and publication history indicate a research interest in productive ageing in the context of their country of origin. Some of the respondents were also able to suggest other colleagues to recruit. We asked each academic to keep in mind the definition of productive ageing relative to her or his own country when taking part in the coding task, regardless of whether they were performing the task containing the KLoSA or the SHARE categories. The Korean academics completed the task between July and August 2017, and the Italian academics completed it between October and December 2017.

Table 2. Coders' characteristics and dates for the conjoint task, by country

Coder	Country of PhD	Country of institutional affiliation	Date of coding
South Korean experts			
K-1	United States	Republic of Korea	03.07.2017
K-2	United States	Republic of Korea	11.07.2017
K-3	United States	Republic of Korea	12.07.2017
K-4	United States	Republic of Korea	20.07.2017
K-5	United States	Republic of Korea	16.08.2017
Italian experts			
I-1	Italy	Italy	22.10.2017
I-2	Italy	Italy	23.10.2017
I-3	United Kingdom	United Kingdom	23.10.2017 & 11.12.2017
I-4	Italy	Italy	13.11.2017
I-5	Italy	Spain	15.11.2017
I-6	Germany	Germany	01.12.2017

All the Korean and three of the Italian experts (I-4, I-5 and I-6) performed comparisons exclusively on the KLoSA categories. Two Italian academics (I-1 and I-2) performed comparisons exclusively on the SHARE categories, and one Italian academic (I-3) performed the task with both sets of categories on different dates. Table 3 shows the number of pairwise comparisons made by each expert, by country and task completed. The highest number of comparisons made was 145 and the lowest was 51. Our final sample consists of 1,021 pairwise comparisons, 683 of which performed on the KLoSA and 338 of which on the SHARE task.

Table 3. Number of comparisons by country, task and coder (total = **1021**)

Country	**Italy**							**Korea**				
n	648							373				
Task	**SHARE**			**KLoSA**				**KLoSA**				
n	338			310				373				
Coder	**I-1**	**I-2**	**I-3**	**I-3**	**I-4**	**I-5**	**I-6**	**K-1**	**K-2**	**K-3**	**K-4**	**K-5**
n	82	145	111	70	75	65	100	101	51	65	104	52

3.3. Model

As outlined above (section 2.2), we model the choices made by experts using ordinal logistic regression models for the choice between 'A is more productive than B', 'A and B are similarly productive', and 'B is more productive than A'. The predictors that enter the model are constructed solely from the attributes of A and B. We construct dummy variables X_A and X_B from the activity levels for A and B respectively, omit the "never" category for each activity, and then define the matrix of predictors for the ordinal logistic regression $X_{BA} = X_B - X_A$, a matrix consisting of values -1, 0, and 1. This means that each coefficient in the resulting regression corresponds to an additive effect (on the log-odds of B being considered relatively more productive than A) of B moving from never engaging in an activity to a higher level of that activity or of A moving from that higher level to never, holding constant both A and B's other activities. For our analysis pooling multiple coders, we hierarchically model the coefficients for each coder for each indicator category as normal draws from a "consensus" coefficient with estimated variance.

Having estimated the coefficients for each indicator category, we use these to generate a measure of productive ageing for each respondent in KLoSA or SHARE by calculating $\hat{\beta}X_i$ given that respondent's observed set of indicators. This yields a cardinal measure of productive ageing that reflects the relative importance that the experts implicitly place on different indicator categories in their pairwise comparisons. This measure is on a log-odds scale defined by the expert's choices. The usual arguments for translating the log-odds into odds do not apply in this context because we are not ultimately interested in the effect of activities on productivity. We are interested in the extent to which participating in a certain activity with a certain frequency – as opposed to not – leads experts to judge a profile as relatively more productive than another such profile. Since our original motivation was to construct a linear index, it makes sense work with βX_i rather than $\exp(\beta X_i)$.

We compare our expert-derived productive ageing scales to those obtained using weighting schemes that have been used in the literature on productive engagement (Baker et al., 2005; Paúl et al., 2012). The comparisons are aimed at assessing two commonly used weighting approaches against the benchmark set by our expert-derived normative scales. First, we obtain a scale by summing up the number of activities that older individuals in each survey perform. This is a widely used strategy in productive ageing research, particularly for those analysing surveys such as SHARE and KLoSA where not all activities are reported in hours per week (Table 1). The comparison of our expert-derived scales with the equal weighting approach allows us to assess the extent to which our experts value limited participation in multiple roles as opposed to intense participation in a single role.

Second, we compare our scale to measures obtained using data-driven methods of aggregation that are only based on the degree of co-variation among activity indicators in the data. We treat paid work, volunteering, grandchild care and informal care as ordered categorical variables, using the same frequency categories as those used for the coding task and described in Table 1. For each survey, we generate a matrix of the polychoric correlations among the four ordinal variables, and perform PCA and FA on that matrix. We focus on the first principal component and the one-factor model, which is also the optimal model as suggested by the "very simple structure" criterion (Revelle & Rocklin, 1979). Similar results are obtained deriving factor loadings for a single-factor model using an ordinal response factor analysis model rather than working with the polychoric correlations. We expect the scale derived from such data-driven measurement methods to differ substantially from our expert-derived scales since, as we have argued, there is no reason to believe that the correlations among various activities are primarily induced by variation in an underlying level of "productivity" across older adults.

3.4. Results

We begin by estimating the ordinal logistic model for the experts' selections separately for each expert, and then construct the implied productive ageing scores for each respondent in KLoSA or SHARE (depending on which categories the coder used). As a test of reliability of the single expert estimates, we tabulate the Pearson correlations between these scores across all pairs of experts (Tables 4 and 5). In this context, where we aim to measure an interval level quantity for which neither the overall mean nor variance of the measure is well defined with respect to the concept, Pearson correlation coefficients are the appropriate measure of reliability. The correlation coefficients provide a measure of the extent to which applying the weighting scheme derived from the judgement of a given expert – as opposed to another – would alter the extent of productive ageing achieved by the same older adult, measured on an interval scale (expressed in log-odds). As such, the correlation matrix provides an initial indication of the robustness of the scale to different coders' weighting schemes.

Table 4 compares the four Italian and five Korean experts who coded comparisons using the indicator categories from KLoSA. Among the Italian experts (I-3 to I-6), the six pairwise correlations range from 0.91 to 0.98. Among the Korean experts (K-1 to K-5), the ten pairwise correlations range from 0.81 to 0.92. Table 5 shows that the three Italian experts who coded comparisons using the indicator categories from SHARE all generated measures that are correlated with one another at 0.94 to 0.96. This indicates a very high level of intercoder reliability: there is not much consequential variation in how the coders weighed the different indicator categories. These results provide strong evidence that the approach of having experts complete pairwise comparison tasks can be effective at generating robust scales. These high correlations resulted from an average of just 93 pairwise comparisons per coder, which was 20–30 minutes work for most of the coders.

Table 4. Correlation (ρ) of KLoSA productive ageing scores constructed from codings of each coder. Comparisons of Italian with Korean experts enclosed in thick border. Correlations of experts' scores with scores obtained from equal weighting (EW) and factor analysis (FA) in the last two columns.

	I-3	**I-4**	**I-5**	**I-6**	**K-1**	**K-2**	**K-3**	**K-4**	**K-5**	**EW**	**FA**
I-3	1.00	0.93	0.95	0.96	0.67	0.93	0.78	0.90	0.85	0.63	-0.29
I-4		1.00	0.91	0.98	0.77	0.93	0.87	0.97	0.92	0.68	-0.48
I-5			1.00	0.91	0.67	0.93	0.73	0.88	0.81	0.57	-0.35
I-6				1.00	0.76	0.94	0.85	0.96	0.91	0.68	-0.42
K-1					1.00	0.83	0.90	0.81	0.87	0.89	-0.38
K-2						1.00	0.83	0.92	0.92	0.70	-0.38
K-3							1.00	0.88	0.92	0.83	-0.43
K-4								1.00	0.89	0.69	-0.61
K-5									1.00	0.76	-0.36

Table 5. Correlation (ρ) of SHARE productive ageing scores constructed from codings of each coder. Correlations of experts' scores with scores obtained from equal weighting (EW) and factor analysis (FA) in the last two columns.

	I-1	**I-2**	**I-3**	**EW**	**FA**
I-1	1.00	0.96	0.95	0.82	0.41
I-2		1.00	0.94	0.73	0.35
I-3			1.00	0.74	0.36

Table 6 shows the coefficients from the analyses pooling all coders who performed the KLoSA and SHARE tasks, respectively. It also illustrates how the scale for comparing relative levels of productive ageing between two or more older individuals in KLoSA or SHARE can be obtained by adding up coefficients for those individuals' attributes. For example, an individual working for 31-40 hours per week and looking after grandchildren for 11-20 hours per week in KLoSA would get a score of 5.09 (= 3.77+1.32), lower than an individual working for the same amount of hours but providing care or help for a sick or disabled adult for 11-20 hours, who gets a score of 5.58 (= 3.77+1.81).

For each of the four activities, the magnitude of the coefficients on various frequencies relative to the "never" category suggests that experts' judgements are internally consistent, with higher coefficients assigned to higher frequency of participation within each activity domain, and negligible inconsistencies in the ranking of frequencies. It is interesting to notice that, for activities coded using descriptive frequencies rather than hours per week, the "almost every day" category gets by far the highest coefficient, suggesting that the way activity frequencies are reported in a survey may influence researchers' assessments of productivity. The "consensus" coefficients from the analysis pooling all the Korean and Italian coders who performed the KLoSA task give an indication of the relative importance assigned by these experts to each of the four activity domains.

To summarise the results, participation in paid work for more than 40 hours per week as opposed to never is associated with the largest increase in the log-odds of a profile being considered relatively more productive than another profile (3.93), followed by paid work participation for 31 to 40 hours per week (3.77). Provision of informal care is the second-ranked activity overall. The coefficients on looking after grandchildren for more than 40 hours per week and on volunteering for charities, religious or political organisations every day, as opposed to never participating in each activity, have similar magnitudes (2.29 and 2.23 respectively), making them the third-and fourth-ranked activities. These findings are in line with a conceptualisation of productive ageing as a reaction to concerns about the financial sustainability of pension and healthcare systems (Bass & Caro, 2001; Herzog et al., 1989). Within this framework, paid work continuation and informal caregiving may represent activities through which older people themselves "make up" for the relative increase in the number of pensioners and long-term care recipients, whereby volunteering and grandchild care may be thought as having higher consumptive or leisurely components (Arpino & Bordone, 2017). Among the three Italian coders who performed the task using the SHARE categories, paid work is also by far the most productive activity. However, these experts assign relatively greater importance to grandchild care and lower to informal caregiving than their colleagues who performed the task using the KLoSA categories.

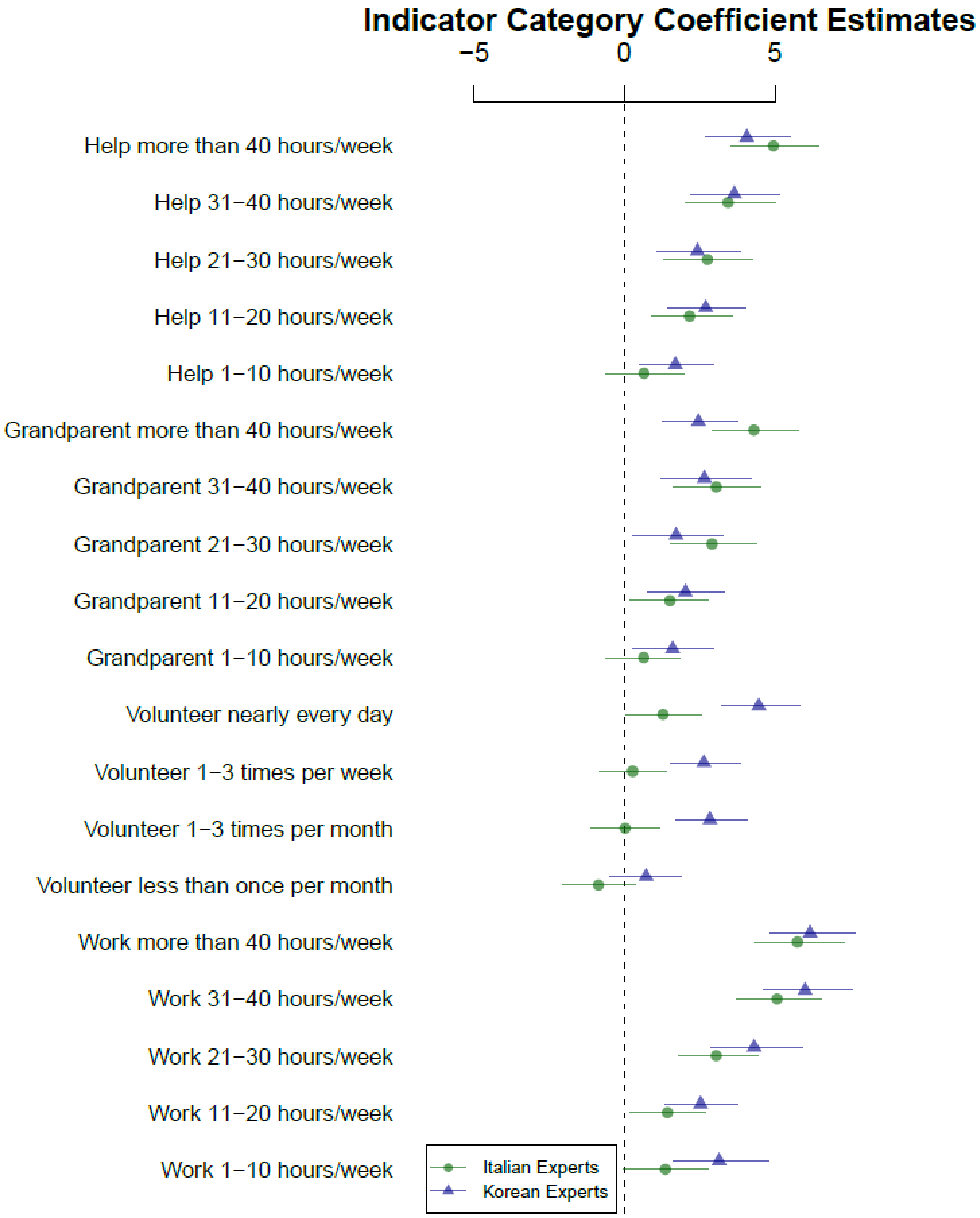
Table 6. Coefficients and standard errors from ordered logistic regression of experts' responses on the full set of activity indicators, by coding task (KLoSA vs. SHARE)

	KLoSA task	**SHARE task**
Paid work (reference: never)		
1-10 hours/week	1.44 (0.31)	0.78 (0.43)
11-20 hours/week	1.31 (0.23)	2.47 (0.43)
21-30 hours/week	2.39 (0.27)	3.55 (0.46)
31-40 hours/week	3.77 (0.28)	5.05 (0.50)
More than 40 hours/week	3.93 (0.26)	5.21 (0.51)
Volunteering (reference: never)		
Less than once/month	0.18 (0.22)	
1-3 times/month	0.99 (0.20)	
1-3 times/week	0.93 (0.18)	
Nearly every day	2.23 (0.25)	
Less than once/week		0.95 (0.30)
Once or twice/week		1.10 (0.31)
About every day		2.33 (0.37)
Grandchild care (reference: never)		
1-10 hours/week	0.59 (0.25)	
11-20 hours/week	1.32 (0.26)	
21-30 hours/week	1.45 (0.32)	
31-40 hours/week	1.77 (0.31)	
More than 40 hours/week	2.29 (0.24)	
Less than once/month		0.43 (0.38)
Once or twice/month		0.44 (0.40)
Once or twice/week		1.61 (0.34)
About every day		3.45 (0.43)
Informal care or help (reference: never)		
1-10 hours/week	0.79 (0.23)	
11-20 hours/week	1.81 (0.26)	
21-30 hours/week	1.86 (0.28)	
31-40 hours/week	2.57 (0.31)	
More than 40 hours/week	3.08 (0.28)	
Less than once/month		0.32 (0.31)
Once or twice/month		0.71 (0.34)
Once or twice/week		0.95 (0.32)
About every day		2.77 (0.37)
Intercepts		
-1 | 0	- 1.03 (0.12)	- 1.17 (0.20)
0 | 1	1.02 (0.12)	0.92 (0.19)
Number of observations	683	325
Number of coders	9 (5 Korean, 4 Italian)	3 (3 Italian)

Going beyond the consensus estimates, when we compare Italian and Korean experts to one another, we see greater evidence of disagreement. The twenty “cross-cultural” pairwise correlations in the individual scales enclosed in the thick border in Table 4 range from 0.67 to 0.97. Some of these are substantially lower than any of the “within-cultural” correlations discussed above, giving an initial indication that there may be some systematic differences between the coefficients that the Korean and Italian coders put on at least some indicator categories. In order to understand these differences, we estimate a hierarchical model that pools the data from the nine coders who completed comparisons using the KLoSA indicator categories. In this model, we assume that Italian and Korean experts are drawn from different populations of experts, each with a common mean coefficient for each indicator category. In Figure 1, we plot the estimates for the “consensus” scales of Italian versus Korean experts.

The coefficient estimates from the hierarchical model indicate that, while the differences in the evaluation of paid work and informal caregiving are small, there is some evidence of differences in the relative importance of full-time (i.e., more than 40 hours per week) grandchild care provision between the Korean and Italian coders. The largest difference is the much higher importance assigned to volunteer work by Korean experts versus Italian experts. These coefficient differences are the primary explanation for the observed pattern of lower pairwise correlations in the scores generated from the responses of experts from different countries. While not large in an absolute sense, these differences illustrate our expectation that the relative coefficients assigned by experts to various productive roles may partly depend on the socio-cultural context to which the definition of productive ageing is applied.

Figure 1. Coefficient estimates for Italian versus Korean experts coding using the KLoSA indicator categories.



The productive aging scores elicited through the coding task can be compared to the scores obtained through alternative normative and data-driven methods of aggregation on the same set of activities. In the commonly used weighting scheme based on summing up the number of activities in which older adults participate, all activities are assigned equal weight, ignoring the frequency with which they are performed. This is equivalent to a linear index in which coefficients are equal and positive for all non-zero levels of all activities, and zero otherwise. Tables 4 and 5 report the correlations between the expert-derived scores and the equal weighting scores in the "EW" columns. For the KLoSA task (Table 4), these correlations range from 0.57 to 0.89 and are generally lower than the correlations of experts' scores with one another. The same is also true for the SHARE coding task, as shown by the correlations under the "EW" column in Table 5, which range from 0.73 to 0.82. With the exception of expert K-1, all expert-derived scales are more strongly correlated with one another (regardless of country of origin) than with the equal weighting scale. The relative values of the coefficient estimates plainly illustrate that experts value some activities (e.g. paid work) as more productive than others (e.g. grandchild care), value higher frequencies of participation as corresponding to higher levels of productivity (Table 6), and do so with a high degree of intercoder reliability.

Lastly, we compare our expert-derived scales to those obtained through data-driven methods of weighting and aggregation. Table 7 shows the factor loadings for single-factor models obtained by performing PCA, FA and an ordinal factor analysis model on the KLoSA and SHARE data, respectively. The standardised factor loadings represent the correlation of each activity with a latent variable, or factor, which summarises (co)variation in the data. The results clearly indicate that the loadings obtained from factor analysis are unlikely to reflect the relative importance of each activity towards productive ageing. In the Korean dataset, the single factor is not positively associated with participation in all four activities, with paid work having a negative association with all the other activities. This is likely to reflect the fact that, in Korea, paid work participation is intensive, and often incompatible with participation in unpaid or family activities (Yang, 2011). For Italian SHARE respondents, we do find a single factor that is positively correlated with higher frequencies of participation in all four activities. However, paid work participation is assigned the lowest weight (i.e. the smallest factor loading) among all activities, suggesting that the latent factor that best explains variation in the data is at most weakly related to productivity as the concept is understood by the experts.

Table 7. Standardised factor loadings for each productive activity for the one-factor model using i) principal components analysis ii) factor analysis iii) Markov Chain Monte Carlo ordinal factor analysis, KLoSA and SHARE data

	PCA on polychoric correlation matrix	FA on polychoric correlation matrix	MCMC ordinal factor analysis
		KLoSA (n = 10,254)	
Paid work	– 0.783	– 0.703	– 0.723
Volunteering	+ 0.305	+ 0.118	+ 0.117
Grandchild care	+ 0.757	+ 0.468	+ 0.755
Informal care & help	+ 0.342	+ 0.149	+ 0.169
		SHARE (n = 2,508)	
Paid work	+ 0.237	+ 0.100	+ 0.160
Volunteering	+ 0.607	+ 0.285	+ 0.291
Grandchild care	+ 0.627	+ 0.357	+ 0.349
Informal care & help	+ 0.738	+ 0.640	+ 1.239

Unsurprisingly, the correlations between the scores derived for each expert through the coding task and the factor scores are low, as shown in the “FA” columns of Tables 4 and 5. For the KLoSA data, the correlations range between 0.29 and 0.61 in absolute value (which sign to use is ambiguous because of the reversed loading on paid work), while for the SHARE data they range between 0.35 and 0.41. Given how much lower these correlations are than those within expert scales and between each expert and the equal weighting approach, it is clear that the correlations among the four activities are unlikely to reflect their substantive association with the concept of productive ageing. This is unsurprising for reasons we have already mentioned but nonetheless highlights the importance of adopting normative weighting schemes for concepts that are meant as pragmatic summaries rather than as reflecting a latent factor that causally generates the observed indicators.

These results indicate some interesting directions for future research in productive ageing. They strongly suggest that data-driven approaches such as FA or PCA are best avoided for obtaining productive ageing scales. Equal weighting schemes are more useful, with the important caveat that they do not reflect the greater value that experts from both countries place on paid work and informal caregiving. Assuming that our expert-derived scales represent a valid benchmark in terms of weighting schemes for productive ageing in Italy and Korea (Lawshe, 1975; Saisana et al., 2005), an interesting exercise would be to replicate previous empirical studies in productive ageing research that use alternative weighting methods in the same contexts, and compare the results to those obtained using the scale developed here. Furthermore, productive ageing scales may be generated for (and compared across) different contexts, by making use of the family of cross-national harmonised ageing studies of which SHARE and KLoSA are part as the source of indicator data (see https://g2aging.org/ for information on the available surveys). Given that the substantive focus of this article is on the development of a method for scale construction, however, in our discussion below we focus broadly on the value added of the method and its social science applications, as well as its limitations.

4. Discussion

In this paper we have proposed an approach for the derivation of indicator coefficients towards the construction of linear indices for measuring aggregate social science concepts. Our method takes the comparison of units with respect to the concept by subject-matter experts (or other appropriately selected coders) as its starting point. It allows us to construct measurement scales based on single coders' weighting schemes, as well as "consensus" scales for groups of coders. The method we propose offers several advantages over a variety of commonly used weighting approaches.

Compared to weighting schemes based on data-driven methods such as PCA or FA, our method allows for normative expressions of the relative importance of various indicators towards a concept. This is necessary for social science concepts defined as summaries of relevant indicators rather than as representations of latent variables that cause – or plausibly approximate the causal process by which – indicators to vary together (Saisana et al., 2005). In our application, the comparison of our expert-derived productive ageing scales with those obtained using data-driven methods demonstrates that measurement methods based on the amount of co-variation among a set of indicators are best avoided when dealing with such pragmatically defined concepts. Having established the need for normative weighting schemes, we argue that our method represents an excellent compromise between fully "implicit" normative schemes such as the equal weighting approach, and more "explicit" methods based on direct numerical assessments by experts.

Relative to the equal weighting approach, our method increases validity with respect to the concept to be measured. Equal weighting describes the relationships between indicators and the measure entirely through crude exclusion/inclusion decisions: indicators get either zero weight, or a weight inversely proportional to the number of included indicators. While this may sometimes yield an acceptable approximation, it is preferable to use information from relevant stakeholders (in our application, subject-matter experts) to describe the relationship more flexibly. Experts or relevant stakeholders are, by definition, highly reliable sources of knowledge about the relative importance of different indicators (Lawshe, 1975), and "implicit" normative schemes such as equal weighting fail to make use of such knowledge.

Relative to "explicit" normative schemes, our approach minimises the burden on experts by giving them a quickly repeatable task that directly relates to the target concept that is the aim of the measurement task. Existing methods for eliciting normative weighting schemes from expert coders include the Budget Allocation Process (BAP) (Hoskins & Mascherini, 2009) and the Analytic Hierarchy Process (AHP) (Saaty, 1977). Although these methods are often invoked as the most appropriate for the generation of valid scales (Greco et al., 2019; Saisana et al., 2005), their use has been limited because it is difficult, in practice, to directly elicit weights from experts. It has been shown that decision-makers find direct numerical weighting difficult when assessing units based on more than one attribute or indicator (Hainmueller et al., 2015). Pairwise coding tasks such as the one we propose allow one to study many indicators at the same time and evaluate which of those indicators make units more or less reflective of a target concept. Research comparing different methods of eliciting preferences to a behavioural benchmark (derived from the outcome of a referendum) shows that designs based on pairwise comparisons come remarkably close to the behavioural benchmark, and are preferable to rating a single profile (Hainmueller et al., 2015). Compared to BAP and AHP, our method gives the participatory audience a more accessible task. Our experts need only make binary/ordinal response comparisons of two units rather than jointly specifying a potentially large number of interval level coefficients, the correct specification of which all depend on one another. All that is required to derive coefficients is having enough repetitions of the pairwise comparison task, so our method is not dependent on the availability of experts with both a deep knowledge of the relevant concept and also a good intuition for linear functions. In fact, it may easily be applied to situations where one wishes to crowdsource coefficients about the relative importance of different items towards a concept of interest where one is interested in how non-expert members of the public think about that concept (Benoit et al., 2016; Carlson & Montgomery, 2017).

Finally, relative to other existing methods for the generation of composite measures, our method makes it straightforward to assess inter-coder reliability. It allows us to compare operationalisations of a concept across coders and groups of coders, with many potential applications in social science research. Here, we wish to highlight three classes of such applications. First, as pointed out above for productive ageing, much empirical research on composite social science concepts resorts to analysing indicators as separate variables, restating the research questions in terms of the indicators rather than the concept (Hank, 2011; Hinterlong et al., 2007). To the extent that this strategy is motivated by difficulties faced by researchers around weighting and aggregation, our method solves the issue by enabling researchers to obtain valid scales in an efficient manner. This facilitates the study of composite concepts and, importantly, allows one to compare measures of such concepts across groups of individuals, contexts and over time. Second, our method is appropriate when one wishes to compare the operationalisation of a given concept across space. For instance, in the case of productive ageing, our method allows one to compare the relative value of different activities as assessed by Italian and Korean experts, which may be used to compare the extent of productive ageing across societies using different cultural standards. As such, our method facilitates research that goes beyond the measurement and use of a concept towards the analysis of its operationalisation. Third, assuming that reliance on judgements from experts or relevant stakeholders maximises validity (Lawshe, 1975; Saisana et al., 2005), our method allows to assess existing scales, such as those obtained using equal weighting or data-driven methods.

The method we have proposed can be usefully applied for measurement in a variety of social science settings, whenever the aim is to obtain a linear index that reflects the relative importance of indicators towards some concept of interest. Examples of potential applications include such diverse scales as the economic or democratic development of countries (Coppedge et al., 2011), the tourism sustainability of regions (Mikulić et al., 2015), or individuals' responsibility over shaping their own career path (Baruch, 2014). Importantly, what these concepts have in common is their definition as a summary of existing indicators rather than as a latent property of the subject under study that causes variation in the indicators.

In practice, if one wishes to generate a scale for a newly developed concept, we follow previous literature in recommending the selection of a pool of five to ten experts for content validation

(Bonsang et al., 2018), followed by reliance on those experts for the pairwise coding task we have proposed. “Real-world” data from surveys or registers provide a good starting point for the generation of the coding task. As we have shown, potential applications of the method also include cases where one wishes to test existing scales against the expert-derived scales, in which case indicators can be the same as those used for the generation of other commonly used scales. For the construction of the coding task, we recommend assigning each unique profile in the data equal probability of being selected for the pairwise comparison, as this maximises variation while avoiding implausible indicator combinations or excessive repetition of the same profiles. As outlined in section 2.4, the optimal number of repetitions of the coding task depends on the amount of variation in the implied coefficients across coders. The more the coefficients vary across coders, the more data will be required to precisely estimate a “consensus” weighting scheme. In our application we have shown that the method performs well with an average of 85 pairwise repetitions per coder, which is the work of around 20-30 minutes. If one wanted to construct a scale using a very large number of indicators, it might be unwise to show coders profiles including all of those indicators at once, although recent tests on conjoint experiments using pairwise comparisons suggest that respondents can cope with more indicators than one might fear (Bansak et al., 2021). If the number of indicators became very large, one might instead show random subsets of indicators for each pairwise comparison, and then rely on modelling to bridge the information about the relative importance of different indicators into a common scale.

There are limitations to acknowledge regarding the methodology that we propose. The first of these relates to indicator availability and selection. In our example, we took the data for the generation of profiles from widely used surveys. This allowed us to obtain comparisons over plausible profiles, while disregarding information on all other characteristics of the profiles, such as age or gender, which could have potentially introduced biases. The underlying assumption is that the definition of the concept of interest is independent of characteristics that are unrelated to the set of indicators included. We can think of instances where this is not the case: in the example of productive ageing, definition of productivity may be thought to differ by, for instance, gender or age. However, if that was the case, then these characteristics could have easily been included in the coding task. Moreover, as we have shown by comparing profiles generated using two different surveys, the way indicators are coded may have an impact on the relative weight assigned by coders. This can be considered more broadly as a limitation of the available data used to derive the indicators.

A second important kind of limitation is that the pairwise comparison method may encourage or discourage certain approaches to coding among coders, though we do not think it is obvious which way such biases would go. For instance, coders might be inclined to look at the indicator they think is most important and then only use the other categories as tie-breakers. Relatedly, depending on how the coders proceed, it may make sense to model the responses differently than we have done. Our analysis assumed a logistic additive response model with no interactions between indicators because this matches the linear index form that is so frequently used by composite measures. However, the coders might have followed coding rules that are poorly described by that model, putting higher or lower importance on particular combinations of indicators. Interactions may be present when the coders’ evaluation criteria are not mutually preferentially independent, in which case multi-criteria approaches that take interactions into consideration would better reflect the coders' choices about the relative weight of the indicators (Angilella et al., 2016). With enough pairwise codings, more complex response functions could be estimated, resulting in non-linear indices that potentially incorporate interactions between the indicators. However, getting sufficient data to reliably recover these is likely to exhaust coders' patience, with limited benefits for the measurement of many concepts.

Finally, we acknowledge that there is an extensive theoretical literature formalising the definition and conceptualisation of concepts (e.g. Guttman, 1959; Saris & Gallhofer, 2004) that we do not directly engage here. Very few of the creators of the composite measures, which are the subject of our analysis, engage the theoretical basis for the concepts that they are measuring at this metaconceptual level. Given this, our aim in this paper is to provide a pragmatic tool for translating concepts as analysts understand them into linear indices as they typically aim to measure them.

References

Allison, P.D., and Christakis, N. (1994). Logit models for sets of ranked items. *Sociological Methodology, 24*, 199-228.

Angilella, S., Corrente, S., Greco, S., and Słowiński, R. (2016). Robust Ordinal Regression and Stochastic Multiobjective Aceptability Analysis in multiple criteria hierarchy process for the Choquet integral preference model. *Omega, 63*: 154-169.

Arpino, B., & Bordone, V. (2017). Regular provision of grandchild care and participation in social activities. *Review of Economics of the Household, 15*, 135-174.

Arpino, B., Pronzato, C., & Tavares, L. P. (2014). The effect of grandparental support on mothers' labour market participation: An instrumental variable approach. *European Journal of Population, 30*, 369-390.

Baker, L., Cahalin, L., Gerst, K., & Burr, J. A. (2005). Productive activities and subjective wellbeing among older adults: The influence of number of activities and time commitment. *Social Indicators Research, 73*, 431-458.

Bansak, K., Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2021). Beyond the breaking point? Survey satisficing in conjoint experiments. *Political Science Research and Methods, 9*, 53-71.

Baruch, Y. (2014). The development and validation of a measure for protean career orientation. *The International Journal of Human Resource Management, 25*, 2702-2723.

Bass, S. A., & Caro, F. G. (2001). Productive Aging: A conceptual framework. In N. Morrow-Howell, Hinterlong, J. and Sherraden, M. (Ed.), *Productive Aging: Concepts and Challenges*. Baltimore & London: The Johns Hopkins University Press.

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Review, 110*, 278-295.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health, 6*(149).

Börsch-Supan, A., & Jurges, H. (2005). *The Survey of Health, Ageing and Retirement in Europe - Methodology*. Mannheim: Mannheim Research Institute for the Economics of Ageing (MEA).

Bratti, M., Frattini, T., & Scervini, F. (2018). Grandparental availability for child care and maternal labor force participation: Pension reform evidence from Italy. *Journal of Population Economics, 31*, 1239-1277.

Carlson, D., & Montgomery, J. M. (2017). A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *American Political Science Review, 111*, 835-843.

Chen, Y. C., Wang, Y., Cooper, B., McBride, T., Chen, H., Wang, D., Lai, C.Y., Montemuro, L.C., & Morrow-Howell, N. (2016). A research note on challenges of cross-national aging research: An example of productive activities across three countries. *Research on Aging, 40*, 54-71.

Coppedge, M., Gerring, J., Altman, D., Bernhard, M., Fish, S., Hicken, A., Kroenig, M., Lindberg, S.I., McMann, K., Paxton, P., Semetko, H.A., Skaaning, S., Staton, J., &

Teorell, J. (2011). Conceptualizing and measuring democracy: A new approach. *Perspectives on Politics, 9*, 247-267.

Dibben, C., Atherton, I., Cox, M., Watson, V., Ryan, M., & Sutton, M. (2007). *Investigating the impact of changing the weights that underpin the Index of Multiple Deprivation 2004*. London: Department for Communities and Local Government.

Fernández-Ballesteros, R., Zamarrón, M. D., Molina, M. Á., Schettini, R., Díez-Nicolás, J., & López-Bravo, M. D. (2011). Productivity in old age. *Research on Aging, 33*, 205-226.

FFP. (2017). *Fragile States Index and CAST framework methodology*. Washington, DC: The Fund for Peace.

Glass, T., Mendes De Leon, R., Marottoli, R. A., & Berkman, L. F. (1999). Population based study of social and productive activities as predictors of survival among elderly Americans. *British Medical Journal, 319*, 478-483.

Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. *Research in Nursing & Health, 20*, 269-274.

Greco, S., Ishizaka, A., Matarazzo, B., & Torrisi, G. (2018). Stochastic multi-attribute acceptability analysis (SMAA): An application to the ranking of Italian regions. *Regional Studies, 52*, 585-600.

Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research, 141*, 61-94.

Guttman, L. (1959). A structural theory for intergroup beliefs and action. *American Sociological Review, 24*, 318-328.

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences, 112*, 2395-2400.

Hank, K. (2011). Societal determinants of productive aging: A multilevel analysis across 11 European states. *European Sociological Review, 27*, 526-541.

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of Business Research, 57*, 98-107.

Herzog, A. R., Kahn, R. L., Morgan, J. N., Jackson, J. S., & Antonucci, T. C. (1989). Age differences in productive activities. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 44*, 129-138.

Herzog, A. R., & Morgan, J. N. (1992). Age and gender differences in the value of productive activities. *Research on Aging, 14*, 169-198.

Hinterlong, J., Morrow-Howell, N., & Rozario, P. A. (2007). Productive engagement and late life physical and mental health: Findings from a nationally representative panel study. *Research on Aging, 29*, 348-370.

Hoskins, B. L., & Mascherini, M. (2009). Measuring active citizenship through the development of a composite indicator. *Social Indicators Research, 90*, 459-488.

IEP. (2020). *Global Peace Index 2020: Measuring peace in a complex world*. Sydney: Institute for Economics & Peace.

KEIS. (2014). The Korean Longitudinal Study of Aging. [Data files and codebooks]. Retrieved from https://survey.keis.or.kr/eng/klosa/klosa01.jsp

Lawshe, C. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575.

Lee, O. E. K., & Lee, J. (2014). Factors associated with productive engagement among older South Koreans. *Journal of Social Service Research, 40*, 454-467.

Loh, V., & Kendig, H. (2013). Productive engagement across the life course: Paid work and beyond. *Australian Journal of Social Issues, 48*, 111-137.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Economics*, New York, Academic Press, 105-142.

Mikulić, J., Kožić, I., & Krešić, D. (2015). Weighting indicators of tourism sustainability: A critical note. *Ecological Indicators, 48*, 312-314.

Morrow-Howell, N., Hinterlong, J., Sherraden, M., & Rozario, P. (2001). Advancing research on productivity in later life. In N. Morrow-Howell, Hinterlong, J. and Sherraden, M. (Ed.), *Productive Aging: Concepts and Challenges*. Baltimore & London: The Johns Hopkins University Press.

OECD. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. Paris: OECD Publishing.

OECD. (2017). *Pensions at a glance 2017: OECD and G20 indicators*. Paris: OECD Publishing.

Paúl, C., Ribeiro, O., & Texeira, L. (2012). Active ageing: An empirical approach to the WHO model. *Current Gerontology ad Geriatrics Research, 2012*, 10.

Permanyer, I. (2011). Assessing the robustness of composite indices rankings *Review of Income and Wealth, 57*, 306-326.

Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research, 14*, 403-414.

Rouzet, D., Sánchez, A. C., Renault, T., & Roehn, O. (2019). *Fiscal challenges and inclusive growth in ageing societies*. Paris: OECD Publishing.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Methematical Psychology, 15*, 234-281.

Saisana, M., Saltelli, A., & Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society: Series A, 168*, 307-323.

Saris, W. E., & Gallhofer, I. (2004). Operationalization of social science concepts by intuition. *Quality and Quantity, 38*, 235-258.

Steer, R. A., & Beck, A. T. (1997). Beck Anxierty inventory. In C. P. Zalaquett & R. J. Wood (Eds.), *Evaluating stress: A book for resources*. Lanham, MD: Scarecrow Press.

Strauss, S., & Trommer, K. (2018). Productive ageing regimes in Europe: Welfare state typologies explaining elderly Europeans' participation in paid and unpaid work. *Journal of Population Ageing, 11*, 311-328.

Suraseranivongse, S., Santawat, U., Kraiprasit, K., Petcharatana, S., Prakkamodom, S., & Muntraporn, N. (2001). Cross-validation of a composite pain scale for preschool children within 24 hours of surgery. *BJA: British Journal of Anaesthesia, 87*, 400-405.

Yang, Y. (2011). No way out but working? Income dynamics of young retirees in Korea. *Ageing and Society, 31*, 265-287.