Similarities and differences in the measurement of "productive ageing" across experts and contexts: A conjoint experiment approach comparing Italy and South Korea

Ginevra Floridi (presenting author)

g.floridi@lse.ac.uk

Department of Social Policy

London School of Economics and Political Science

United Kingdom

Benjamin Lauderdale

b.e.lauderdale@lse.ac.uk

Department of Methodology

London School of Economics and Political Science

United Kingdom

Paper prepared for submission to the 2019 PAA Annual Meeting

September 2018

Introduction

The use of composite measures for multidimensional concepts has become increasingly common in academic and policy research (Greco, Ishizaka, Tasiou, & Torrisi, 2018; OECD, 2008). Composite measures enable researchers to summarise information and to compare the performance of subjects – such as individuals or countries – with respect to concepts that lack more direct measures. They are easier to interpret than sets of indicators, and they facilitate communication with policy-makers and the general public (OECD, 2008). Composite measures are necessary for providing quantitative assessments and analyses of multidimensional concepts that cannot be captured by studying their constituent attributes separately. However, the construction of composite scales is not straightforward, as it requires researchers to make decisions on which indicators to include, and on how to aggregate such indicators into a scale. One of the most complex issues with the construction of composite measures is that of assigning weights to individual indicators that are reflective of their relative importance towards the concept to be measured (Munda & Nardo, 2005).

In ageing research, a prominent example of a multidimensional concept is 'productive ageing', defined as older people's participation in activities that produce services or goods that have value for others (Bass & Caro, 2001). The concept is "multidimensional" in the sense that a variety of activities might contribute to an individual's overall level of productivity, and each of these activities are easier to measure individually than is the overall concept. Despite the relevance of productive ageing in light of demographic trends in high-income countries, this multidimensional concept is difficult to formalise into a single measure that can be compared across contexts. Most quantitative studies of productive ageing treat its dimensions as separate indicators, or use arbitrary combinations of those indicators.

In this paper, we propose a method for supervised measurement that takes the form of a conjoint experiment on experts and apply it to the concept of productive ageing. We then compare assessments of the relative weights assigned to various productive activities between a group of Italian and a group of South Korean academics. We consider participation in paid work, volunteering, grandchild care and informal care for sick and disabled adults as indicators of productive ageing. To construct our measure, we take these indicators (as measured in major ageing surveys) and ask Italian and South Korean academics with a research interest in productive ageing to complete a series of pairwise comparisons on hypothetical profiles of older people participating in different combinations of these activities, and to different extents. By ranking a hypothetical profile as 'more productive',

'similarly productive' or 'less productive' relative to another such profile, the experts implicitly reveal the relative weights to place on each activity. We model responses on the full set of activities, revealing the weights assigned to them by each expert. These weights can then be used to assess the level of agreement among academics about the relationship between the indicators and the concept of interest and, ultimately, to generate a measure of productive ageing from the available indicators.

With respect to our specific application, this study represents a first attempt to generate a productive ageing scale that is responsive to the relative importance that academics put on different activities. More generally, we make a methodological contribution to the literature on composite measures by proposing a strategy for supervised measurement that is straightforward to implement and that easily allows to test for differences across experts and contexts, providing a structured way for scholars to assess agreement and disagreement about the empirical realisation of multidimensional concepts.

Background

Weights in Composite Measures

Composite measures are commonly used to operationalise concepts that are multidimensional in nature, such as wellbeing, poverty or social participation (Decancq & Lugo, 2013; Hoskins & Mascherini, 2009; Ravallion, 2011). The OECD (2008) Handbook on Constructing Composite Indicators recommends a theory-driven approach for the generation of composite measures, from the selection of indicators to be included to their aggregation into a scale. Particular attention in this process should be devoted to weighting. Weights in composite measures express both the relative importance of each indicator towards the concept to be measured, and the trade-offs between indicators. Since weights are essentially value judgements, weighting should be done along the lines of the theoretical framework. However, in many applications, weighting decisions are poorly justified (OECD, 2008).

Weighting can be implemented through unsupervised or supervised methods. Unsupervised or data-driven methods use observed associations among a defined set of indicators to identify the measure that best explains variation in them. Examples of data-driven methods include principal components analysis, factor analysis and multivariate regression (Greco et al., 2018). These approaches have well understood limitations, the most important one being

that the amount of variation in the data that an indicator explains is unlikely to reflect its substantive importance for the concept to be measured (Saisana & Tarantola, 2002). Because it is not explicitly based on theory, the derivation of weights using data-driven approaches is not straightforward to interpret, and often lacks transparency. Moreover, the weights derived from correlation structures can change between different editions of the same measure, hindering comparability over time (Decancq & Lugo, 2013).

Supervised or participatory methods involve decisions by participants (usually subject-matter experts) that determine the weights to be assigned to each indicator towards the construction of a scale. Examples of participatory approaches include the budget allocation process, where participants are assigned a budget to distribute among various indicators according to their relative importance (Hoskins & Mascherini, 2009); and the analytic hierarchy process (Saaty, 1977), where participants are asked to compare pairs of indicators based on an ordinal preference scale, with levels ranging from 'equally important' to 'much more important'. These existing supervised methods can help in the generation of weights, as they make the subjectivity behind the weighting process explicit. However, they can exert significant cognitive stress on the decision makers, and may become unmanageable as the number of indicators increases (Greco et al., 2018). Moreover, they may lead to inconsistent or biased results in cases where the participatory audience does not clearly understand the supervision framework (OECD, 2008).

Productive Ageing: Definition and Measurement

The academic discourse on productive ageing has developed over the last thirty years as a reaction to the growing policy focus on maintaining older people's 'productivity' in the labour force in response to population ageing in high-income countries (Bass & Caro, 2001; Herzog, Kahn, Morgan, Jackson, & Antonucci, 1989). The productive ageing framework highlights the societal importance of broader forms of participation by defining productive activities as those producing goods and services, or developing other people's capacity to do so, whether for pay or not (Bass & Caro, 2001).

Narrow definitions of productive ageing only include activities that can be assigned economic value, such as paid work, volunteering, and caregiving (Hinterlong, 2008). Broader definitions also include activities that develop older people's potential to be productive, such as education, training and self-care, and some go as far as including any activity that has a

social or spiritual dimension, such as shopping, hobbies and religiosity (Fernández-Ballesteros et al., 2011; Thanakwang & Isaramalai, 2013). Broad definitions of productive ageing overlap with two alternative conceptualisations, 'active ageing' (WHO, 2002) and 'successful ageing' (Rowe & Kahn, 1997), both of which tend to be more concerned with social participation, health and the biological aspects of the ageing process.

Ideally, empirical work on productive ageing should first define and justify which activities are considered productive and then aggregate indicators of such activities into a single measure. This task is made difficult by the fact that the relative importance of each activity is not predetermined. Even before the problem of indicator availability, the relative importance of different activities may vary according to who defines the concept, and to which context the concept is being applied. For an example of the latter problem, we might imagine that the relative extent to which paid work and child care work are assessed as productive could depend on the structure of old-age pensions and child care provision in a given social context.

Because of the difficulties connected with weighting, research on productive ageing often resorts to analysing activities as separate variables. This strategy is most commonly used in studies of the effects of activity participation for health and wellbeing (Hinterlong, Morrow-Howell, & Rozario, 2007; Li, Xu, Chi, & Guo, 2013), but it is also a common procedure in studies of the predictors of productive participation, in which case activities are used as separate dependent variables (Akintayo, Hakala, Ropponen, Paronen, & Rissanen, 2016; Hank, 2011). This approach to measuring productive ageing is sometimes preferred as it does not require the researcher to attach arbitrary values to each activity. In turn, though, it does not reveal much about the extent of productive ageing achieved, as it restates the research question in terms of the indicators rather than the concept. As a solution to this problem, some studies of the health effects of participation combine multiple activities together into binary indicators of whether respondents are 'involved' or not, usually restricting the definition of involvement to those who participate with a certain frequency (Di Gessa & Grundy, 2014; Kim, Kim, & Kim, 2013). However, this coarse approach to aggregating indicators still does not take into account differences in productive roles.

Alternatives to no or simple binary aggregation are summing up the number of activities (Caro, Caspi, Burr, & Mutchler, 2009) or the number of hours (Herzog et al., 1989; Loh & Kendig, 2013) of productive involvement. These methods present complementary drawbacks: summing up the number of activities is fine for assessing participation in multiple roles, but it

is problematic as a measure of the extent of involvement, as intense participation in a single role is valued less than sporadic participation in various activities. Summing up the total number of hours solves this problem, but still assigns fundamentally different forms of participation equal weight (Bukov, Maas, & Lampert, 2002). In fact, twenty hours of involvement distributed, for instance, between volunteering and grandchild care might be less demanding or have different health effects than twenty hours of paid work or care for a disabled adult. Studies of productive ageing by Glass and colleagues (1999) and Davis et al. (2012) have attempted to build productive ageing indices that rank subjects based on type, diversity and frequency of participation. 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 extents of involvement end up being clustered together in the same group or percentile of the distribution.

A way of aggregating components that explicitly gives relative weight to each of them is to assign activities a monetary value. While the standard procedure for doing this with paid work is to consider an average wage typically given for that type of work, the monetary value of unpaid productive activities needs to be estimated, usually by calculating the amount of money that would be needed to purchase equivalent goods or services on the market (Fernández-Ballesteros et al., 2011; Herzog & Morgan, 1992). Despite representing sensible strategies for assessing the relative importance of each activity towards a measure of 'productivity', monetary valuation methods are not the only defensible kind of valuation (Morrow-Howell, Hinterlong, Sherraden, & Rozario, 2001). Older people's participation may have value beyond monetary terms, and may be especially likely to provide private goods to its recipients. For instance, activities such as grandchild care may be valued far more by the recipients than their market cost, and, because they also tend to have a consumptive component, individuals may spend considerably more time and effort on them than what is required on the market (Herzog & Morgan, 1992). In addition, even assuming that monetary values adequately reflect the importance of activities towards the conceptualisation of productive ageing, the monetary cost of an activity is undoubtedly a bad proxy for studying its predictors or its consequences for health and wellbeing.

These kinds of debates often lead researchers back to unsupervised methods, as a way of avoiding difficult measurement questions by "letting the data decide". For example, Paul, Ribeiro and Texeira (2012) make use of principal components analysis to identify and aggregate indicators of active ageing in a study of Portugal. In the resulting measure, various

indicators of health and activity are given a score proportional to the amount of (co-)variation each of them explains in the sample. However, there is no reason to expect that the weights resulting from these methods will actually be a good measure of the concept of interest; in the example below, we show how badly they can go awry.

Definition and Context for this Study

We adopt a relatively narrow definition of productive ageing as producing services or goods that have value for others, and consider paid work, volunteering, grandchild care and care for sick or disabled adults as productive activities. We exclude activities such as learning, housework and self-care because of their predominantly consumptive nature, albeit in recognition of their potential for developing older people's capacity to be productive. Narrower definitions offer a good compromise between the need, on the one hand, to make the concept relevant for policy-making in countries predominantly concerned with the economic consequences of population ageing; and that, on the other, to rectify the age and gender biases inherent in treating paid work as the only form of productive accomplishment (Herzog et al., 1989). Moreover, narrow definitions have the advantage of facilitating comparison and replication (Morrow-Howell et al., 2001).

As discussed above, existing studies of productive ageing have mostly relied upon weakly supervised or data-driven methods for the weighting and aggregation of activities into a scale. Because productive ageing is an academic concept, and the weights represent value judgements about the relative importance of each activity, measurement should ideally rely on the judgements of academics with an expertise in productive ageing. However, strong supervision is often difficult to implement in practice, as actual experts struggle to translate their expertise into direct decisions regarding the relative numerical weights of the activities. We propose a conjoint experiment approach for the eliciting of weights from experts that makes the subjectivity behind the weighting process transparent and that is straightforward to implement. Moreover, the method allows to assess agreement and disagreement among experts about the relative importance of the indicators towards the concept to be measured.

In this application, we compare formalisations of productive ageing between a group of Italian and a group of South Korean academics. Italy and Korea make good cases for comparison. In both countries, productive ageing is topical in light of demographic ageing (OECD, 2017). At the same time, there is reason to believe that scholarly assessments of the

relative importance of each activity domain towards its measurement differ. The recent academic discourse on productive ageing in Italy has developed in the context of the low provision of public and subsidised family services in the country (Saraceno, 2016). Older people who look after their grandchildren or care for disabled adults provide services that would otherwise have to be paid for, and increase the productive capacity of others by substituting for their time. In particular, recent research on older Italians has paid increasing attention to the role of grandchild care in facilitating young mothers' labour force participation (Arpino, Pronzato, & Tavares, 2014; Bratti, Frattini, & Scervini, 2017). In Korea, recent studies in social gerontology have proposed the adoption of definitions of productivity beyond paid work (Kim, 2013; Kim et al., 2013). However, as Lee and Lee (2014) argue, 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.

Because the relative value assigned to each activity may differ by sociocultural context (Chen et al., 2016), comparative studies on productive ageing are rare and mostly limited to comparing countries within the same region (Feng, Son, & Zeng, 2015; Hank, 2011). However, cross-regional comparative research is valuable as it can help untangle the relationships between sociocultural structures and older people's productive engagement. A necessary step towards making sensible comparisons is to assess the degree of scholarly agreement and disagreement about the realisation of the concept between different contexts. Expert agreement on the relative importance of productive activities towards an aggregate measure would validate cross-regional comparisons; strong disagreement would instead suggest that alternative conceptualisations should be used in different contexts.

Method

Conjoint analysis is a multivariate method of data analysis in which respondents are assumed to evaluate any object or concept as a bundle of attributes (Hair, Anderson, Tatham, & Black, 1998). In conjoint experiments (Green & Rao, 1971), respondents are asked to compare or rate profiles combining multiple attributes that vary randomly across repetitions of the task, enabling researchers to estimate the relative influence of each attribute on the resulting choice. Since its aim is to decompose respondents' preferences for different profiles into

individual indicators, conjoint analysis is often referred to as a decomposition method (Greco et al., 2018). It was first developed in relation to marketing research, and since the 1970s it has been widely used to study how consumers make trade-offs among competing products and suppliers (Green, Krieger, & Wind, 2001). More recently, conjoint experiments have been also applied to the study of attitudes in political science, as in the case of natives' attitudes towards different types of immigrants (Hainmueller & Hopkins, 2015).

In this paper we use a conjoint experiment for the measurement of a multidimensional concept, productive ageing, for which the component attributes are known, but the relative weight to be assigned to each attribute towards the construction of a scale is unknown. We consider four activity domains – paid work, volunteering, grandchild care and informal care for adults – as indicators of productive ageing. Our aim is to elicit experts' judgements about the relative importance of each activity towards the construction of a productive ageing scale. Each expert is assumed to possess knowledge of a latent scale that measures how 'productive' an older individual is based on that individual's frequency of participation in each of the four activities considered. Eliciting such a latent scale directly is difficult, as it requires experts to make explicit decisions about the quantification of the value of each activity (Green & Rao, 1971). However, the expert can more easily assess two profiles of older individuals relative to each other on the productive ageing scale based on their frequency of participation in the four activities. The scale can thus be elicited by having the expert repeatedly compare between pairs of older adults whose frequencies of participation in each productive activity vary across repetitions of the task. Conjoint analysis can be carried out either at the individual respondent level (in this application, experts) or via aggregation across respondents (Hair et al., 1998). This allows us to estimate and compare different scales for each expert, as well as a 'consensus' scale pooling responses from all the experts. Moreover, it allows us to assess whether there are differences in the conceptualisation of productive ageing between a group of Italian and a group of Korean academics, by estimating separate scales for each group.

The focus of this study is on the measurement of a multidimensional concept, rather than on attitudes or preferences towards some object or idea. By definition, an expert is someone who knows the concept well, and who can identify the relative importance of each attribute towards its measurement. Therefore, it makes sense to ask experts, and not the general public, to perform the coding task. Nevertheless, the method could also be applied to cases where

one intends to elicit a scale for a multidimensional concept from the general population via a crowdsourcing or online survey platform.

Data

The first step for data collection was the generation of 'productivity profiles' of older people 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 Korean Longitudinal Study of Aging (KLoSA) (http://survey.keis.or.kr/eng/klosa/klosa01.jsp) and from the Italian sample of the Survey of Health, Ageing and Retirement in Europe (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 activities. The target population of KLoSA at baseline consists of individuals aged 45 and above in 2006, excluding younger spouses as well as 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 (Borsch-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 to sick or disabled 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 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. 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, along with information relative to the productivity profile of both individuals in the pair, is then saved as an observation in our dataset. Conjoint experiments often use an independent randomization, but this would lead to implausible combinations of activity frequencies in our application. Thus, in order to obtain interesting comparisons and to avoid excessive repetition of the same productivity profiles across comparisons, we assign each unique productivity profile found in the surveys an equal probability of being selected in every repetition of the task.

We collected data from five Korean and six Italian academics, whose names are anonymised as listed in Table 2. We recruited experts by initially contacting academics whose curriculum vitae and publication history indicate a research interest in productive or active ageing. 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 conjoint 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.

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. Table 3 shows the number of pairwise comparisons performed by each expert, by country and task completed. We obtained on average of 93 comparisons per expert; the highest number of repetitions performed 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.

Model

We model the choices made by the 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 from the randomly assigned attributes of A and B. We construct dummy variables X_A and X_B from the

assignments for A and B respectively, omitting 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 βX_i given that respondent's observed set of indicators. This yields a cardinal measure of productive ageing that reflects the relative weights that the experts implicitly place on different indicator categories in their codings. This measure is on a log-odds scale defined by the expert 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 effects of activity indicators on the experts' responses, but rather on the measurement of a latent productive aging scale. Since it is easier to think in terms of additive scales rather than multiplicative scales, working with βX_i is preferable to working with $\exp(\beta X_i)$.

We also compare our productive ageing scale to measures obtained using unsupervised 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 conjoint 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 principal components analysis (PCA) and factor analysis (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.

Results

We begin by estimating the ordinal logistic model for the coders' selections separately for each coder, and then constructing the implied productive ageing scores for each respondent in KLoSA or SHARE (depending on which categories the coder used). As an initial test of reliability, we tabulate the correlations between these scores across coders (Tables 4 and 5)[1]. Table 4 compares the four Italian and five Korean experts who coded comparisons using the indicator categories from KLoSA. Among the Italian experts, the six pairwise correlations range from 0.91 to 0.98. Among the Korean experts, 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 highly reliable scales.

Table 6 shows the coefficients from the analyses pooling all coders who performed the KLoSA and SHARE tasks, respectively. For each of the four activities, the magnitude of the coefficients on various frequencies relative to the "never" category suggests that experts' judgements were internally consistent, with higher weight assigned to higher frequency of participation within each activity domain, and negligible inconsistencies in the ranking of frequencies. 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. Unsurprisingly, 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). Thus, the five Korean and four Italian experts who performed the task using the KLoSA categories seem to agree that paid work is the most important productivity domain. Provision of informal care is the second-ranked activity overall. Caregiving for a sick or disabled person for more than 40 hours per week as opposed to never is associated with an increase in the log-odds of being selected as relatively more productive by 3.08, and the corresponding increase for caregiving for 31 to 40 hours per week is 2.57. The coefficients on looking after grandchildren for more than 40 hours per week and on volunteering for charities, religious or

[1] In this context, where we aim to measure a latent quantity for which neither the overall mean nor variance of the scores is well defined, correlation coefficients are the appropriate measure of reliability.

political organisations every day, as opposed to never participating in each activity, are of similar magnitude (2.29 and 2.23 respectively), making them the third-and fourth-ranked activities in terms of productive ageing. 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 more importance to grandchild care and relatively less to informal caregiving than their colleagues who performed the task using the KLoSA categories.

Going beyond the consensus estimates, when we compare Italian and Korean experts to one another, we see greater evidence of disagreement. The twenty "cross-context" pairwise correlations in the individual scales enclosed in the thick border in Table 4 range from 0.67 to 0.97. Given that some of these are substantially lower than the "within-context" correlations discussed above, this is an initial indication that there may be some systematic differences between the weights 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 of which have 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 between Korean and Italian experts in the evaluation of paid work and informal caregiving are small, there is evidence that the Korean experts put more weight on volunteering and less on grandchild care than the Italian coders. In particular, the importance assigned to volunteer work is substantially higher for Korean than for Italian experts. According to the responses given by the four Italian coders, only older adults who participate in volunteer work "nearly every day" are considered significantly more productive than those who do not perform any volunteering at all. Conversely, in relation to grandchild care provision, while Italian experts assign progressively higher weight to higher frequencies of participation, Korean coders appear to assign a flat degree of credit across all non-zero frequencies, with 40 or more hours of weekly grandchild care valued as not significantly more productive than up to 10 hours per week of participation. Given the number of coders from each country we cannot be confident that these differences would be maintained in a broader population of experts. Still, these patterns are a potential explanation for the observed patterns in the pairwise correlations of

scores generated from individual coders. The differences between Korean and Italian coders in the importance assigned to volunteer work and grandchild care provision are also in line with our expectation that the relative weights assigned by experts to various productive roles may partly depend on the socio-cultural context to which the definition of productive ageing is applied. In Italy, as noted above, grandparental care may be considered particularly important for welfare generation (Arpino et al., 2014), while volunteer work may be considered more as a recreational activity. In Korea family care may be seen as an "obligation" rather than a productive accomplishment of older people (Lee & Lee, 2014). This would explain why, while those not looking after grandchildren at all are penalised as significantly "less productive" than those who do some grandchild care, spending progressively larger amounts of time in this activity is not associated with being considered significantly more productive.

The relative weights placed on each activity by the experts and elicited through the conjoint coding task can be compared to the weights obtained through unsupervised methods of aggregation on the same set of activities. Table 7 shows the factor loadings for the single-factor model obtained by performing PCA, FA and Markov-Chain Monte Carlo (MCMC) ordinal factor analysis 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 the construction of a productive ageing scale. In the Korean dataset, no single latent factor is positively associated with participation in all four activities, with paid work time having a negative association with all the other activities. Given that the number of hours present in one week is limited, it is clear that the latent factor is measuring older Koreans' time allocation across different domains rather than their degree of productivity. For Italian SHARE respondents, we do find a single factor that is positively correlated with higher frequencies of participation all four activities. However, paid work participation is assigned the lowest weight (i.e. the lowest factor loading) among all activities, suggesting that the latent factor that best explains variation in the data is at most weakly related to productivity. Given that productive ageing is defined as older people's participation in activities that produce goods or services that have an economic value, the small loading on paid work participation suggests that the latent factor is unlikely to truly reflect SHARE respondents' level of productive engagement.

Lastly, we generate factor scores for SHARE and KLoSA respondents from the single-factor model with polychoric correlations (FA), and compare the resulting 'productive ageing' scores to those elicited from Italian and Korean academics in the conjoint coding task. Unsurprisingly, the correlations between the scores assigned by each expert through the supervised conjoint experiment and the unsupervised factor scores are low, as shown in the last rows of Tables 4 and 5. For the KLoSA data, the correlations range between 0.29 and 0.61 in absolute value, while for the SHARE data they range between 0.35 and 0.41. Given how much lower these correlations are than the between-expert correlations, it is clear that the statistical associations among the four activities are unlikely to reflect their substantive correlation with a latent measure of productive ageing. This comparison highlights the particular importance of adopting some form of measurement supervision for the construction of scales whenever the indicators that make up the desired concept are jointly subject to a constraint, such as the number of hours present in one week.

Discussion

In this paper we described an experimental approach to measurement supervision that takes the form of a conjoint coding task on experts, and applied it to the concept of productive ageing with reference to Italy and Korea. The method is effective in eliciting internally consistent judgements, as demonstrated by the fact that the ordering of frequencies for the same activity is largely consistent within and across coders. The results indicate that there is a high degree of agreement among experts about the relative importance of the four different activity domains towards the construction of a productive ageing scale. Consensus estimates across experts indicate that paid work participation is valued as the "most productive" activity, followed by informal care for sick or disabled adults. A likely explanation for this is that productive ageing was developed as a reaction to concerns about the financial sustainability of pensions and healthcare systems: paid work continuation and informal caregiving may therefore represent activities through which older people themselves "make up" for the relative increase in the number of pensioners and long-term care recipients (Morrow-Howell & Wang, 2013). Volunteering and grandchild care are generally thought of as having higher consumptive or leisurely components (Arpino & Bordone, 2017), which may also explain why the expert coders implicitly view them as less intrinsically productive. While the Korean and the Italian scholars largely agree about the relative weights of paid

work and informal caregiving, Korean experts place relatively more importance on volunteering and less on grandchild care provision than their Italian counterparts. These results are in line with differences between the two socio-cultural contexts to which the definition of productive ageing is applied, and suggest some degree of caution about the use of multidimensional indices of older people's engagement in cross-national comparative research (Chen et al., 2016).

To our knowledge, this study is the first to propose a measure of productive ageing that is responsive to the relative importance that academics, who use the concept for empirical research, attach to each of its component activities. It contributes to the literature on composite measures by proposing an experimental approach for supervised measurement based on a pairwise conjoint coding task. The proposed method offers several advantages compared to the various measurement strategies most commonly employed for multidimensional concepts like productive ageing. Unlike most strong supervision methods, it does not require experts to make difficult direct assessments of the relative weights to put on different indicators, instead giving them relatively straightforward pairwise comparisons of units involving the available set of indicators. At the same time, it does not require supervision over real cases involving information beyond the indicator set, which could potentially introduce biases, and it easily allows for the testing of differences between experts, providing a structured way for scholars to assess agreement and disagreement about the empirical realisation of aggregate concepts. Compared to weakly supervised methods of aggregation that involve arbitrary weighting decisions, our approach allows to assign a weight to each indicator that is reflective of its relative importance towards the construction of a scale based on experts' judgements. The results clearly indicate that experts view some productive activities as more important than others. Thus, aggregation approaches that simply sum up the total number of activities or hours of involvement and give equal weight to all forms of participation may not adequately reflect the academic conceptualisation of productive ageing. The method also offers clear advantages relative to purely data-driven measurement strategies, as demonstrated by the comparison of our productive ageing scale to those obtained using factor analysis. In fact, our approach is responsive to the relative importance that experts put on different indicators, as opposed to the empirical correlation of those indicators. In general, weighting based on the co-variation between indicators is best avoided as a measurement strategy, especially when indicators are jointly subject to a constraint such as time.

There are some important limitations to recognise regarding the methodology that we propose. The first of these relates to indicator availability and selection. We took the data for the generation of profiles from widely used datasets on ageing. 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 productive ageing is independent of individual characteristics that are unrelated to one's participation in productive roles. However, if the definition of productivity was thought to differ by, for instance, gender or age, then these characteristics could have easily been included in the coding task. In the datasets we looked at, activities are coded using different categories, with volunteer work being the only activity categorised on a frequency scale in the Korean dataset, and paid work the only one measured in hours in the Italian dataset. If the scale on which activities are measured influences experts' judgements on the comparisons, this may constitute a threat to the internal validity of the scale. However, since ageing datasets such as KLoSA and SHARE are widely used in research on productive ageing (Hank, 2011; Lee & Lee, 2014), this can be considered more broadly as a limitation of the available data rather than one that is specific to this application.

A second kind of limitation is that the pairwise comparison method may encourage or discourage certain approaches to coding among the experts, though we do not think it is obvious which way such biases would go. One could imagine that simply showing all the indicators together implicitly indicates that they all deserve some (or even similar) weight. On the other hand, to code more quickly, coders might be inclined to look at the indicator they think is most important (in this case, likely paid work) 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, but in principle the coders might have followed coding rules that are poorly described by that model, putting higher or lower weight on particular combinations of indicators especially. With enough pairwise codings, more complex response functions could be estimated, but getting sufficient data to reliably recover these is likely to exhaust coders' patience, with limited benefits for the measurement of most concepts. Finally, if one wanted to construct a scale using a very large number of indicators, it would be unwise to show experts profiles including all of those indicators at once. 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.

To conclude, the use of a conjoint coding experiment on experts for the generation of weights for a multidimensional scale can produce estimates that are highly consistent both within and across coders. Moreover, it allows to test for inter-coder reliability in the definition and measurement of any multidimensional concept. As such, the method can be applied to a variety of different situations in which the researcher wishes to generate a measurement for a multidimensional concept and to assess inter-coder variation in the definition of a scale.

References

Akintayo, T., Hakala, N., Ropponen, K., Paronen, E., & Rissanen, S. (2016). Predictive factors for voluntary and/or paid work among adults in their sixties. *Social Indicators Research, 128*(3), 1387-1404.

Arpino, B., & Bordone, V. (2017). Regular provision of grandchild care and participation in social activities. *Review of Economics of the Household, 15*(1), 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*(4), 369-390.

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.

Borsch-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. (2017). Grandparental availability for child care and maternal labor force participation: Pension reform evidence from Italy. *Journal of Population Economics, Online first*, 1-39.

Bukov, A., Maas, I., & Lampert, T. (2002). Social Participation in Very Old Age: Cross-Sectional and Longitudinal Findings from BASE. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 57*(6), 510-517.

Caro, F. G., Caspi, E., Burr, J. A., & Mutchler, J. E. (2009). Global activity motivation and activities of older people. *Activities, Adaptation & Aging, 33*(3), 191-208. doi: 10.1080/01924780903148151

Chen, Y. C., Wang, Y., Cooper, B., McBride, T., Chen, H., Wang, D., . . . 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, 21 November 2016*, 1-18.

Davis, S., Crothers, N., Grant, J., Young, S., & Smith, K. (2012). Being Involved in the Country: Productive Ageing in Different Types of Rural Communities. *Journal of Rural Studies, 28*(4), 8.

Decancq, K., & Lugo, M. A. (2013). Weights in multidimensional indices of wellbeing: An overview. *Econometric Reviews, 32*(1), 7-34.

Di Gessa, G., & Grundy, E. (2014). The Relationship between Active Ageing and Health Using Longitudinal Data from Denmark, France, Italy and England. *Journal of Epidemiology and Community Health, 68*(3), 6.

Feng, Q., Son, J., & Zeng, Y. (2015). Prevalence and correlates or successful ageing: A comparative study between China and South Korea. *European Journal of Ageing, 12*(2), 83-94.

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*(2), 205-226. doi: 10.1177/0164027510395398

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.

Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G. (2018). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research, Online first*.

Green, P. E., Krieger, A. M., & Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces, 31*, 56-73.

Green, P. E., & Rao, V. R. (1971). Conjoint measurement for quantifying judgemental data. *Journal of Marketing Research, 8*, 355-363.

Hainmueller, J., & Hopkins, D. J. (2015). The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants. *American Journal of Political Science, 59*(3), 529-548.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). Multivariate data analysis. Upper Saddle River, New Jersey: Prentice-Hall.

Hank, K. (2011). Societal Determinants of Productive Aging: A Multilevel Analysis across 11 European States. *European Sociological Review, 27*(4), 15.

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*(4), 129-138.

Herzog, A. R., & Morgan, J. N. (1992). Age and gender differences in the value of productive activities. *Research on Aging, 14*(2), 169-198.

Hinterlong, J. (2008). Productive Engagement Among Older Americans: Prevalence, Patterns, and Implications for Public Policy. *Journal of Aging & Social Policy, 20*(2), 141-164. doi: 10.1080/08959420801977491

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*(4), 348-370.

Hoskins, B. L., & Mascherini, M. (2009). Measuring active citizenship through the development of a composite indicator. *Social Indicators Research, 90*(3), 459-488.

KEIS. (2014). The Korean Longitudinal Study of Aging. from Korean Employment Information Service http://survey.keis.or.kr/ENCOMAM0000N.do

Kim, J. H. (2013). Productive Activity and Life Satisfaction in Korean Elderly Women. *Journal of Women & Aging, 25*(1), 80-96. doi: 10.1080/08952841.2012.717850

Kim, J. H., Kim, M. H., & Kim, J. (2013). Social activities and health of Korean elderly women by age groups. *Educational Gerontology, 39*(9), 640-654.

Lee, O. E. K., & Lee, J. (2014). Factors associated with productive engagement among older South Koreans. *Journal of Social Service Research, 40*(4), 454-467.

Li, Y., Xu, L., Chi, I., & Guo, P. (2013). Participation in Productive Activities and Health Outcomes Among Older Adults in Urban China. *The Gerontologist, 54*(5), 784-796.

Loh, V., & Kendig, H. (2013). Productive Engagement Across the Life Course: Paid Work and Beyond. *Australian Journal of Social Issues, 48*(1), 111-137.

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.

Morrow-Howell, N., & Wang, Y. (2013). Productive Engagement of Older Adults: Elements of A Cross-Cultural Research Agenda. *Ageing International, 38*, 11.

Munda, G., & Nardo, M. (2005). *Constructing consistent composite indicators: The issue of weights*. Ispra: Joint Research Centre.

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.

Paul, C., Ribeiro, O., & Texeira, L. (2012). Active Ageing: An Empirical Approach to the WHO Model. *Current Gerontology ad Geriatrics Research, 2012*, 10.

Ravallion, M. (2011). On multidimensional indices of poverty. *Journal of Economic Inequality, 9*(2), 235-248.

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.

Rowe, J. W., & Kahn, R. L. (1997). Successful Aging. *The Gerontologist, 37*(4), 7.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Methematical Psychology, 15*(3), 234-281.

Saisana, M., & Tarantola, S. (2002). *State-of-the-art report on current methodologies and practices for composite indicator development*. Ispra: European Commission, Joint Research Centre.

Saraceno, C. (2016). Varieties of familialism: Comparing four Southern European and East Asian welfare regimes. *Journal of European Social Policy, 26*(4), 314-326. doi: 10.1177/0958928716657275

Thanakwang, K., & Isaramalai, S. (2013). Productive engagement in older adults: A concept analysis. *Nursing & Health Sciences, 15*(1), 124-130.

WHO. (2002). *Active Ageing: A Policy Framework*. Geneva: Who Publishing.

Tables

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

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

Table 2. Coders 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

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

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 factor analysis (FA) in the last row.

	I-3	**I-4**	**I-5**	**I-6**	**K-1**	**K-2**	**K-3**	**K-4**	**K-5**
I-3	1.00	0.93	0.95	0.96	0.67	0.93	0.78	0.90	0.85
I-4		1.00	0.91	0.98	0.77	0.93	0.87	0.97	0.92
I-5			1.00	0.91	0.67	0.93	0.73	0.88	0.81
I-6				1.00	0.76	0.94	0.85	0.96	0.91
K-1					1.00	0.83	0.90	0.81	0.87
K-2						1.00	0.83	0.92	0.92
K-3							1.00	0.88	0.92
K-4								1.00	0.89
K-5									1.00
FA	- 0.29	- 0.48	- 0.35	- 0.42	- 0.38	- 0.38	- 0.43	- 0.61	- 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 factor analysis (FA) in the last row.

	I-1	**I-2**	**I-3**
I-1	1.00	0.96	0.95
I-2		1.00	0.94
I-3			1.00
FA	0.41	0.35	0.36

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)

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

Figures

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