On Aggregating Subjective Well-Being

Niklas Kaunitz

March 8, 2017

Abstract

This paper discusses the assumptions underlying the aggregation of individual well-being. Any aggregation method is associated with *measurability* assumptions regarding the underlying well-being measure, as well as *moral philosophical* assumptions with respect to how individual well-being is weighted into a composite metric. The former is generally acknowledged more often than the latter.

We compare welfare across a set of countries, under alternative aggregation methods, and find that countries can be ranked under comparatively weak measurement assumptions, and, just as important, that the aggregation methods can be chosen so as to refrain from strong ethical preconceptions.

1 Introduction

Subjective well-being (SWB) is by now an acknowledged research area in both economics and psychology. Hitherto, the general focus has been on the determinants of *individual* well-being, such as the impact on well-being of income, unemployment, divorce, etc.[1] There has, however, been an increasing interest in performing aggregate comparisons.

[1]See Frey and Stutzer (2002) and Di Tella and MacCulloch (2006) for a summary of the literature in economics, and Diener et al. (1999) for the field of psychology. Berlin and Kaunitz (2015) is an example showing how economic conditions influence subjective well-being.

This range from suggestions of supplementing "hard" measures such as GDP with SWB data (e.g. Fleurbaey (2009) and Stiglitz et al. 2010), to the extreme of judging country welfare exclusively based on SWB (Layard 2005).

Mostly, these country comparisons of subjective well-being rank countries using only the mean. This holds true both for academic articles (as e.g. in Diener 2000), international organizations (OECD 2016), and the increasingly common policy think tanks promoting subjective well-being comparisons, such as the Earth Institute's World Happiness Report (Helliwell et al. 2015). While it is well known that using the mean of individual well-being is dependent on a measurement assumption—we must assume that individual subjective well-being is cardinally measured, and level-comparable across individuals—the implicit *philosophical* axiom in using mean-ranking is rarely discussed, or even mentioned. For example, a mean-ranking rule does not care whether an aggregate increase is caused by an improvement for the worst-off or for the happiest part of the distribution.

As a case in point, Kahneman et al. (2004) proposes using the so-called experience sampling method to arrive at individual measures of well-being which, in their opinion, lies closer to the origin of the utility concept. This is continuing a strand of innovative research by Kahneman and others on improved measures of individual SWB. In the end, the authors derive their statistic for "national well-being" by simply taking the mean of individual SWB. Granted, the explicit references to Bentham and Edgeworth indicate that Kahneman et al. consider themselves working in the utilitarianist tradition, but it is, nonetheless, remarkable that their aggregation statistic is presented as if it were a mere technicality. The authors admit that there are "of course, many assumptions underlying this formulation" (p. 432), but mention only measurement aspects.

Sechel (2014) is a notable exception. She criticizes the use of mean comparisons, and instead proposes using a head count ratio: the population share of at least "just

satisfied" individuals. However, this measure has its own problems, analogous to the head count measure in income poverty analysis: it ignores *all* distributional information above and below the chosen cutoff line. While this may seem motivated for an explicit poverty measure, it appears strange for a well-being measure. For example, with a SWB scale of 0–10 and a cutoff line at 5, Sechel's head count ratio would rank the vectors $(0, 5)$ and $(4, 10)$ as equally good. But as the latter distribution dominates the former, it appears counterintuitive that this is not reflected in the ranking. Sechel motivates this choice through a sceptical stance on interpersonal level comparisons, arguing that "just satisfied" is the SWB level that is most comparable across countries.

If we instead take both measurement and moral philosophical assumptions seriously, how can we compare different methods of aggregating well-being? We will proceed using tools from social choice and the income inequality literature. For each aggregation scheme, we will be clear about (1) what measurent assumption on "life satisfaction" is required, and (2) which normative assumptions are required. These aggregation methods will then be compared for a group of countries, taking statistical sampling noise into account when deriving rankings of countries.

The rest of this paper is structured as follows. Section 2 gives a background and discusses some basic conceptual and philosophical aspects of well-being aggregation. Section 3 introduces the formal aparatus needed for comparing measurement and aggregational assumptions in a systematic way. In section 4, this theoretical framework is applied to an empirical analysis of the Nordic countries, plus Germany and the U.K. Finally, section 5 summarizes the main points made in the paper.

2 Welfarism, utilitarianism and prioritarianism

In a seminal article, Sen (1979) coined the term *welfarism*. This denotes the dogma that, when assessing a social state, we only use information on individuals' self-perceived well-being. This is similar to what in moral philosphy is normally referred to as *hedonism*—only satisfaction, however defined, has an intrinsic moral value. Of course, this does not mean that welfare comparisons are only of interest to those willing to accept the dogma of welfarism. Indeed, even if it is not considered to be *all* that matters, for most people (although there are exceptions) self-perceived welfare surely matters to *some* extent, as a "piece of the puzzle".[2] In this paper, we will, in all practical respects, take a welfarist perspective: we will only consider individuals' self-assessed well-being, and how this can be aggregated to measures of social welfare.

Taking welfarism as given, it is an open question how to rank vectors of individual well-being. Do we care about equality of distribution in well-being, and, if so, how does the trade-off look between magnitude and equality? The answer of classical utilitarianism is to consider only the sum of individual well-being, and disregard anything else. Thus, equality of distribution plays no role at all. In applied work on subjective well-being, and generally in social choice, sum-ranking is very rarely used. Instead, the standard is rather to use its population-independent analogue—the mean.[3] Using the mean implies that we are totally indifferent to how well-being is distributed in the population; one individual improving from utility 1 to 2 is neither better nor worse than another individual improving from 99 to 100.

[2]For an example of marrying Sen's theory of capabilities (Sen 1985) and aggregate welfare comparisons, see Schokkaert (2007).

[3]The differences in moral philosophical implications between sum-utilitarianism and mean-utilitarianism are in fact quite far-reaching, for example with respect to Derek Parfit's so-called repugnant conclusion (Parfit 1984). This is, however, usually disregarded in social choice, even to the point of calling mean-ranking the "utilitarian social welfare ordering" (Bossert and Weymark 2004). A discussion of these matters is beyond the scope of this paper. See, e.g., Tännsjö (1998).

Historically, some utilitarian economists have defended *income* egalitarianism by resorting to the presumed decreasing marginal utility of income. Namely, if all individuals have identical, concave, utility functions of income, the utilitarian optimum is achieved at the exactly egalitarian distribution. Or alternatively, if all individuals have concave, but unknown, utility functions (randomly distributed), expected utility is maximized for the completely egalitarian distribution (Lerner 1944). But these arguments evade the fundamental issue, in that they seek to reduce preferences for equality to a question of utilitarian efficiency. It is easy to think of potential situations where there really are trade-offs between better-off and worse-off individuals, and where the latter do not have the privilege of higher marginal utility of income—for example in the distribution of public health care. Such considerations have led to the notion of *prioritarianism*.

Derek Parfit (Parfit 1991) illustrates his account of "the priority view" by the following example. Imagine that a family has two children, one healthy, and the other suffering from some painful handicap. The family can now move either to the city, where the second child could recieve some treatment, or to the suburb where the first child would flourish. Furthermore, we assume that the gain to the first child of moving to the suburb is greater than the gain to the second child of moving to the city. Parfit summarizes the payoffs in a matrix:

	Child 1	Child 2
City:	20	10
Suburb:	25	9

While a utilitarianist would choose the suburb, the prioritarianist may argue that the marginal improvement for the second child is more important than the potential larger gain for the first child—in other words, priority to the worst-off. However, in contrast to radical egalitarian ideals, such as maximin, prioritarianism also puts value on total well-

being. The exact trade-off between total utility and its distribution is often left unspecified. So ultimately, two prioritarianists may come to different conclusions on the example above.

When it comes to commonly held moral intuitions, my impression is that they lie closer to prioritarianism than to utilitarianism, although some conception of diminishing marginal utility of income is probably there as well. When it comes to public health care (at least in European countries), policies seem in part prioritarian rather than utilitarian. For example, there clearly exist cases of extremely expensive patients, where the marginal utility of public funds is arguably very low, and thus indefensible with utilitarian arguments. From a prioritarian point of view, however, these cases resemble the example of the two children above.

3 Measurement and social welfare orderings

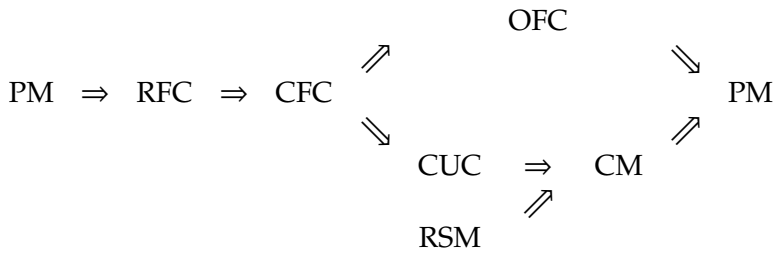
The informational content of a measure can be described by specifying the set of transformations under which the measure stays "the same"; we denote this the *set of admissible transforms*.[4] For example, for an ordinal measure, the set of admissible transforms consists of all strictly increasing functions. Similarly, we can define information restrictions for a vector of (utility) measures, $\mathbf{u} = (u_1, \ldots, u_n)$, where an admissible transform is a vector of functions. Taking again the ordinal case, we can define ordinal utility with interpersonal comparability by restricting the set of admissible transforms to all n-vectors $\boldsymbol{\phi} = (\phi^*, \ldots, \phi^*)$, where ϕ^* is strictly increasing; i.e., we restrict the strictly increasing transform ϕ^* to be the same for all individuals. Now, to say that, for instance, "OFC implies OM" is equivalent to saying that $\boldsymbol{\phi}_{\text{OFC}} \subset \boldsymbol{\phi}_{\text{OM}}$. Table 1 describes the most important examples of information restrictions and their associated sets of admissible transforms. The

[4]This section draws partly from Bossert and Weymark (2004), which is a comprehensive summary of measurement, social welfare orderings and some fundamental results. For an extensive formal treatment of measurement theory, see Roberts (1979).

Table 1: Measurability assumptions

Measurability class	Admissible transforms, $\boldsymbol{\phi} = [\phi_1(t), \ldots, \phi_n(t)]$	Interpersonal comparability (with SWB interpretation)	Implies
Ordinal measurability (OM)	ϕ_i is strictly increasing for all i	—	—
Ordinal, full comparability (OFC)	$\phi_i \equiv \phi^*$, where ϕ^* is strictly increasing	levels ("x is more satisfied than y")	OM
Cardinal measurability (CM)	$\phi_i(t) = a_i + b_i t$, where $\forall i : a_i \in \mathbb{R}, b_i \in \mathbb{R}_{++}$	—	OM
Cardinal, unit comparability (CUC)	$\phi_i(t) = a_i + bt$, where $\forall i : a_i \in \mathbb{R}$ and $b \in \mathbb{R}_{++}$	intra-individual differences ("going from state A to B, x benefitted twice as much as y")	CM, OM
Cardinal, full comparability (CFC)	$\phi_i(t) = a + bt$, where $a \in \mathbb{R}$ and $b \in \mathbb{R}_{++}$	levels, differences	all above
Ratio-scale measurability (RSM)	$\phi_i(t) = b_i t$, where $\forall i : b_i \in \mathbb{R}_{++}$	intra-individual ratios ("going from state A to B, x doubled her satisfaction while y tripled his")	CM, OM
Ratio-scale, full comparability (RFC)	$\phi_i(t) = bt$, where $b \in \mathbb{R}_{++}$	levels, all differences, all ratios (x is 2.5 times as satisfied as y)	all above
Perfect measurability (PM)	ϕ_i is the identity transform, $\phi_i(t) = t$	(no restriction)	all above

Figure 1: The logical interconnections between measurability classes



logical interconnections between these measurability classes are illustrated in figure 1.[5]

We now formalize the notion of a social decision rule. A *social welfare ordering* (SWO) denotes an ordering of social states, $\succeq$. As usual, $\mathbf{u} \sim \mathbf{v}$ is an abbreviation for "$\mathbf{u} \succeq \mathbf{v}$ *and* $\mathbf{v} \succeq \mathbf{u}$," and $\mathbf{u} \succ \mathbf{v}$ abbreviates "$\mathbf{u} \succeq \mathbf{v}$ *and not* $\mathbf{v} \succeq \mathbf{u}$." Here, we will only consider preorders on utility vectors. This means that an SWO is required to be *reflexive* ($\mathbf{u} \succeq \mathbf{u}$) and *transitive* ($\mathbf{u} \succeq \mathbf{v} \succeq \mathbf{w} \Rightarrow \mathbf{u} \succeq \mathbf{w}$). However, we do not make the (otherwise common) assumption of *completeness*; i.e., we allow for the possibility that there may be pairs of social states for which neither $\mathbf{u} \succeq \mathbf{v}$ nor $\mathbf{v} \succeq \mathbf{u}$ holds. In addition, it is common to make further assumptions on SWO:s, of which we will only consider the following ones:

- *anonymity*: if $\mathbf{u}'$ is a permutation of $\mathbf{u}$, then $\mathbf{u}' \sim \mathbf{u}$
- *strong Pareto principle*: if $\mathbf{u} \geq \mathbf{v}$, with $u_i > v_i$ for some i, then $\mathbf{u} \succ \mathbf{v}$
- *weak Pareto principle*: if $\mathbf{u} > \mathbf{v}$ ($u_i > v_i$ for all i), then $\mathbf{u} \succ \mathbf{v}$
- the *principle of population*: duplicating a distribution does not change orderings, i.e. $(\mathbf{u}, \mathbf{u}) \sim \mathbf{u}$

[5]We do not here make a distinction between cases where the underlying phenomemon is cardinal but ordinally measured, and cases where the underlying phenomenon can only be given an ordinal interpretation (and, hence, must be measured ordinally). Whatever measurability assumption is adopted, that will, thus, capture the joint effect of the phenomenon and its measurement. We are in this paper chiefly interested in the connections between assumptions, aggregation methods and rankings, not in how these assumptions can be motivated.

The anonymity axiom, sometimes called symmetry, can also be seen as an impartiality condition. The Pareto principles are variations of non-satiation conditions. Strong Pareto implies non-satiation in all arguments separately, while weak Pareto can be seen as a minimal non-satiation requirement. The term 'Pareto' is somewhat unfortunate, since when paired with anonymity, they go much further than the usual concept of *Pareto improvement* in economics. While the Pareto principles may seem innocuous, the Pareto principles may not be accepted by an extreme egalitarian. Note also that head count ratio measures, such as that in Sechel (2014), do not satisfy these axiom (as illustrated in the example given in section 1). Finally, the principle of population, also called replication invariance, may be important if we want to compare different-size populations, or if we want to examine samples (unless using population weights). Note that classical utilitarianism, where social states are ranked according to the *sum* of individual utilities, does not satisfy this axiom. (Other axioms are possible, and common in the literature, such as continuity, separability and completeness.) In many cases, the SWO is (or can be) defined by means of a function f, so that $\mathbf{u} \succeq \mathbf{v} \Leftrightarrow f(\mathbf{u}) \geq f(\mathbf{v})$. Then f is called a *social welfare function* (SWF). (The value of f itself is usually taken to carry only ordinal information.)

We now say that a vector of transforms $\boldsymbol{\phi}$ is *compatible* with a social welfare ordering $\succeq$ if

$$(\boldsymbol{\phi} \circ \mathbf{u}) \succeq (\boldsymbol{\phi} \circ \mathbf{v}) \text{ iff } \mathbf{u} \succeq \mathbf{v},$$

i.e., if $\succeq$ is invariant with respect to $\boldsymbol{\phi}$. Letting $\boldsymbol{\Phi}^*$ be the set of all vectors compatible with $\succeq$, we say that $\succeq$ *requires* $\boldsymbol{\Phi}^*$ (or $\succeq$ is *information invariant with respect to* $\boldsymbol{\Phi}^*$). For example, the standard economic concept of Pareto improvement requires only ordinal measurability (OM). The head count ratio, however, additionally requires level comparability (OFC).

Table 2: Compilation of social welfare orderings

Ordering	$U \succeq V$ iff	Req.	Complete	Pareto
HC ratio	$\mathbb{E}[U \geq c] \geq \mathbb{E}[V \geq c]$	OFC	yes	none
1st order stoch. dom.	$\forall x : F_U(x) \leq F_V(x)$	OFC	no	both
Leximin	$\exists x : [F_U(x) < F_V(x) \;\&\; \forall (y < x) : F_U(y) = F_V(y)]$, or $\forall (x) : F_U(x) = F_V(x)$	OFC	yes	weak
Mean ("utilitarian")	$\mathbb{E}[U] \geq \mathbb{E}[V]$	CFC	yes	both
Mean log	$\mathbb{E}[\log(U)] \geq \mathbb{E}[\log(V)]$	RFC	yes	both
2nd order stoch. dom.	$\forall x : \int_{-\infty}^{x} F_U(t)\mathrm{d}t \leq \int_{-\infty}^{x} F_V(t)\mathrm{d}t$	CFC	no	both

Table 2 summarizes some social welfare orderings along with their required measurability assumptions, whether they are complete orders, and whether they satisfy the Pareto principles. Conversely, starting from a particular utility measure, table 2 shows how the choice of social welfare ordering must depend on what measurability assumption we impose on the measure in question. For example, let's say we consider our utility measure to be ordinal and fully interpersonally comparable. We can then evaluate social welfare using, say, the leximin or stochastic dominance, but the mean ranking would not be logically consistent under this measurability assumption.[6]

The stochastic dominance social welfare orderings constitute an important class and will be examined a bit closer. Distribution A is said to stochastically dominate distribution B at order 1 if $F_A(x) \leq F_B(x)$ for all $x \in \mathbb{R}$. It is clear that first-order stochastic dominance (SD1) satisfies our axioms above: since it compares only distributions it trivially satisfies anonymity and the principle of population, and it is clearly reflexive, transitive and

[6]For example, letting R be the utilitarian ordering, $\mathbf{u} = (1, 1, 4)$ and $\mathbf{v} = (0, 0, 9)$, we see that $\mathbf{v}R\mathbf{u}$, but $\sqrt{\mathbf{u}}R\sqrt{\mathbf{v}}$, so the OFC transform ($\sqrt{\cdot}$, $\sqrt{\cdot}$, $\sqrt{\cdot}$) is not compatible with R.

satisfies Pareto. However, it is well-known from portfolio theory that $A \underset{1.s.d.}{\geq} B$ implies $f(A) \geq f(B)$ for *any* increasing function f. Thus, SD1 is the *minimal* SWO given our axioms, in the following sense: if A first-order stochastically dominates B then $A \succeq B$ for *any* SWO satisfying our (very modest) axioms.[7] By comparing the definitions of the head count ratio and SD1 in table 2, we see that $A \underset{1.s.d.}{\geq} B$ implies that A (weakly) dominates B for any possible head count measure. It is apparent that first-order stochastic dominance is an extraordinary powerful property. We could also modify SD1 to only consider well-being up to some specified threshold t, if we are more concerned with those least satisfied. This would be equivalent to considering all head count ratio measures at or below t.

We now return to prioritarianism. As discussed in section 2, prioritarianism acknowledges a trade-off between total welfare and egalitarianism, but does not unequivocally specify a solution. At this point we introduce the concept of a *Pigou-Dalton transfer*. A Pigou-Dalton transfer is a transfer from a better-off individual to a worse-off individual, leaving their relative ranking unchanged. The so-called Pigou-Dalton transfer condition, or the Pigou-Dalton principle, states that a Pigou-Dalton transfer must always increase social welfare. Now, if we interpret prioritarianism as putting higher weights on worse-off individuals, it appears that any prioritarian SWO must satisfy the Pigou-Dalton principle. A general class of such orderings are those SWOs corresponding to concave social welfare functions; an example is the mean of log utility, $\mathbb{E}[\log(U)]$ (see table 2). In general, this class require the stronger measurement assumption of ratio-scale full measurability (RFC). Another prioritarian example, requiring only CFC measurability, is the class of single-parameter Gini social welfare orderings introduced in Donaldson and Weymark (1980) (a generalized utilitarian SWO with positional weights). The egalitarian component can be strengthened further by adding the requirement of *transfer sensitivity*: a Pigou-Dalton

[7] Fomulated differently: Let C be the class of all SWOs that satisfies Pareto and anonymity. Then SD1 $= \bigcap C$.

transfer increases welfare more if it takes place lower down in the distribution (Shorrocks 1987). Although somewhat less intuitive than the Pigou-Dalton principle, this puts even more weight on the well-being of those worst-off. Transfer sensitivity is implied by any SWF that is concave with positive third derivative (i.e., with convex first derivative).

So far, implementing prioritarianism seems to be somewhat arbitrary. Luckily, second-order stochastic dominance can assist us here. In order to define stochastic dominance at higher orders, let $D^1 = F(x)$ and

$$D^s(x) = \int_0^x D^{(s-1)}(y)\mathrm{d}y \tag{1}$$

for $s \geq 2$. Then A stochastically dominates B at order s if, for all $x \in \mathbb{R}$,

$$D_A^s(x) \leq D_B^s(x).$$

For the second order (SD2), this amounts to

$$\int_{-\infty}^x \Big[F_A(y) - F_B(y)\Big]\mathrm{d}y \leq 0$$

Note that since second-order stochastic dominance integrates under the values of y, it requires cardinal measurability. $A \underset{2.s.d.}{\geq} B$ can be interpreted as, given well-being level w, the average "unhappiness gap" to w is smaller in A than in B, for any w. In comparison, SD1 only consider the shares below w, not the average "deficit". In this way, SD2 takes also equality of distribution into account. It can be shown that $A \underset{2.s.d.}{\geq} B$ implies $A \geq B$ for any (weakly) concave SWO that satisfies Pareto and anonymity (including the mean). In fact, SD2 is the minimal SWF if we add the Pigou-Dalton transfer condition to our other axioms. This makes second-order stochastic dominance very interesting for performing

assumption-agnostic welfare comparisons from a prioritarian point of view.

Higher orders of stochastic dominance is possible. For instance, third-order stochastic dominance is the intersection of all SWOs that also satisfies tranfer-sensitivity (in addition to the assumptions already mentioned). We will, however, settle for up to the second order only.

4 Empirical application

We will here examine whether different measurement assumptions, and normative assumptions, result in different rankings of countries, depending on which ranking method is chosen. Or put otherwise: how limited normative and measurement assumptions can we "get away with" when comparing SWB distributions?

We use data from the European Social Survey (ESS), which contains around 1,000–3,000 observations per country and year. We will focus on a group of countries that are reasonably similar, in terms of language and culture: the Nordic countries, along with U.K. and Germany. The year for comparison is 2008. Since sampling is non-random in most countries, post-stratification sample weights are supplied, and we will adjust all estimation methods here to account for the weighting.

As a well-being measure we will use ESS's life satisfaction variable, which is formulated in the following way: "All things considered, how satisfied are you with your life as a whole nowadays?" The scale reaches from 0 to 10, with only the end points anchored (0="Extremely dissatisfied", 10="Extremely satisfied").

4.1 Statistical methods

For comparing single points in two distribution functions, as for all head count ratios, we use t-statistics, but implement weighted variants of these to take account of the non-

random sampling in ESS (using the supplied post-stratification weights). Mean-ranking is implemented similarly.

Leximin is implemented as stepwise (weighted) t-tests, starting from the lowest level of y and proceeding upwards until a significant difference is found, or the entire range has been examined. Here, the p-values are upwards adjusted, using a Holm-Bonferroni procedure (Holm 1979). (If we did not, the probability of making a type I error would increase with the number of test points.)

Testing for stochastic dominance is where we meet some difficulty. The definition of $D^s(x)$ in 1 can be expressed non-recursively as

$$D^s(x) = \frac{1}{(s-1)!}\int_0^x (x-y)^{(s-1)}\mathrm{d}F(y),$$

which has the sample analogue

$$\begin{aligned}\hat{D}^s(x) &= \frac{1}{(s-1)!}\int_0^x (x-y)^{(s-1)}\mathrm{d}\hat{F}(y)\\ &= \frac{1}{N(s-1)!}\sum_{i=1}^{N}(x-y_i)^{(s-1)}I(y_i \le x)\\ &= \frac{1}{N(s-1)!}\sum_{i=1}^{N}(x-y_i)_+^{(s-1)}.\end{aligned}$$

The problem of estimating the stochastic dominance relation now amounts to finding confidence intervals for the estimator $d^s(x) = \hat{D}^s_A(x) - \hat{D}^s_B(x)$. Davidson and Duclos (2000) show that under certain regularity conditions, for $K = A, B$, $\sqrt{N}\left(\hat{D}^s_K(x) - D^s_K(x)\right)$ is asymptotically normal with zero mean, and with covariance structure (where $K, L \in \{A, B\}$)

$$\lim_{N\to\infty} N\text{cov}\left[\hat{D}^s_K(x), \hat{D}^s_L(x')\right] = \frac{1}{[(s-1)!]^2}\mathbb{E}\left[(x - y^K)^{s-1}_+ (x' - y^L)^{s-1}_+\right] - D^s_K(x)D^s_L(x').$$

Here, the expectation can be consistently estimated using its sample analogue

$$\frac{1}{N}\sum_{i=1}^{N}(x - y^K_i)^{s-1}_+ (x' - y^L_i)^{s-1}_+.$$

Using this result, we can construct t-values for any point in the joint range of distributions A and B. In this paper, we use only discrete outcome variables; hence, we can examine the entire outcome space.

We now implement tests for stochastic dominance of order s in the following way: For each point, x, in the outcome space, we compute the t-value for the difference $d^s(x) = \hat{D}^s_A(x) - \hat{D}^s_B(x)$, thus obtaining a vector of p-values for testing whether the distribution functions are significantly different in each point. In the process, we weight all t-values and covariance matrices using the ESS post-stratification weights. Finally, since we are making a joint comparison of multiple inequalities, the p-values are upwards adjusted using the Holm-Bonferroni method. Now, distribution A stochastically dominates distribution B at order s if and only if (1) no $d^s(x)$ is significantly above 0, and (2) for at least one x_0, $d^s(x_0)$ is significantly below 0.[8]

[8]This procedure can be used to test stochastic dominance for arbitrary distributions. However, since we examine discrete distributions, we could use the fact that both $\hat{D}^s_A(x)$ and $\hat{D}^s_B(x)$ are jointly normally distributed, and hence also $d^s(x) = \hat{D}^s_A(x) - \hat{D}^s_B(x)$. While tests for inequality could be straightforwardly derived (as χ^2-tests), things are more complicated for one-sided tests. Still, utilizing joint normality should make it possible to design tests with more statistical power than those used in this paper.

4.2 Empirical comparison of SWFs

For each SWF examined, we summarize the resulting (potentially incomplete) ordering in a Hasse diagram—a graph of directed links. A directed link $A \to B$ denotes $A \succeq B$ (weak preference) at the 5-percent significance level, where the null hypothesis is specified as indifference. More precisely: if $B \succ A$ does *not* hold at the 5-percent level, we conclude $A \succeq B$. Unless specifically stated otherwise, all rankings are transitive—i.e., arrows $A \to B$ and $B \to C$ implies that $C \succ A$ has been found not to hold at the 5-percent significance level, so we deduce $A \succeq C$. The absence of a link (direct or inferred by transitivity) represents non-comparability. Thus, each Hasse diagram summarizes $\binom{6}{2} = 15$ pairwise comparisons.[9]

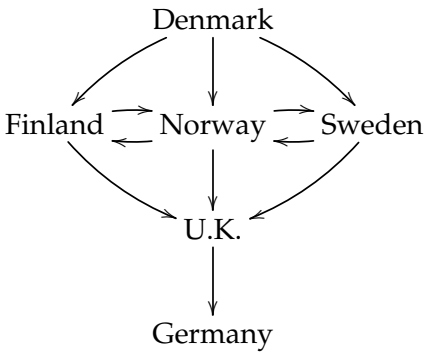
We start with a minimal measurement assumption, where well-being is considered to be ordinally measured, but with interpersonally comparable levels (OFC). We first consider the simplest imaginable SWF: calculating the population share above a specific satisfaction threshold. Figure 2 shows the rankings resulting from two such headcount ratio measures, at thresholds 5 and 2.

Panel (a) uses the threshold at the midpoint of the well-being interval, which can be interpreted as being "just satisfied". This is the measure that is proposed by Sechel (2014). According to this ranking, Finland, Norway and Sweden are equivalent, all three are dominated by Denmark and dominate U.K., which in turn dominates Germany. However, if we instead choose a threshold of 2, as in panel (b), the situation changes somewhat. Now, Denmark, Finland and Norway are equivalent, but only the latter two dominate Sweden, which in turn dominates both of Germany and U.K. Of course, it is an open question

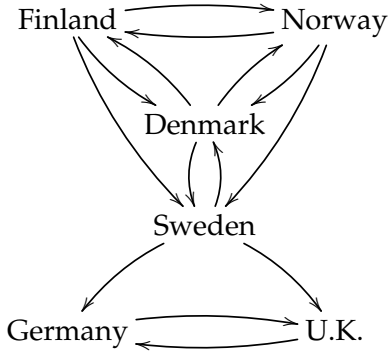
[9] Furthermore, for the non-SWF social welfare orderings (stochastic dominance and leximin), each comparison relies on a statistical test for each discrete point in the range of the satisfaction distributions, as described above, thus resulting in 150 statistical tests per figure. The full estimation results would clearly take up too much space here, but can be obtained from the author by request.

Figure 2: Ordinal SWFs: headcount ratio measures

(a) Headcount ratio at $stf = 5$



(b) Headcount ratio at $stf = 2$



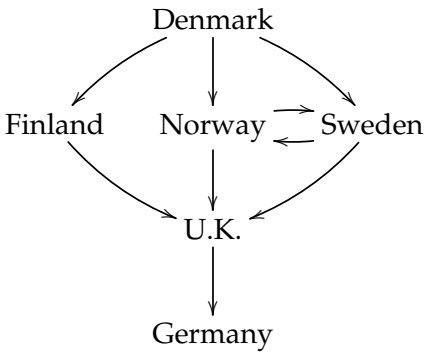
whether a larger sample size (more statistical power) would push Denmark closer to Finland and Norway (and dominate Sweden) or closer to Sweden (and be dominated by Finland and Norway).

Now, is Sechel right in choosing ranking (a) rather than ranking (b)? This question does not seem to have an obvious answer. Choosing a low threshold is clearly more in line with the Rawlsian egalitarian ideal, but what moral doctrine is ranking (a) consistent with? A utilitarian would accept neither, of course, since both disregards vast parts of the welfare information available.

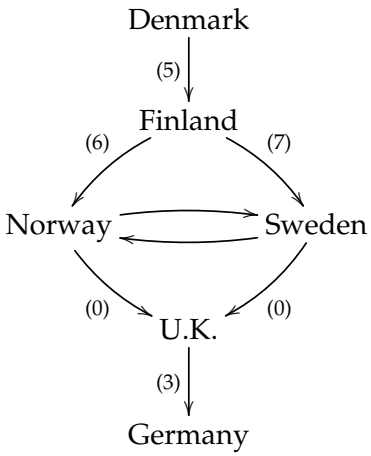
Headcount measures are arbitrary in that they rely on choosing a single point in the satisfaction distribution and disregard all information above or below. We consider in figure 3 two alternatives that refrain from considering only a single point, yet respect the assumption of only ordinal measurability. In figure 3b, we consider the leximin ranking, which can be interpreted as a generalized version of the maximin rule. This ranking puts maximum weight on the unhappiest share of the population, and is thus an egalitarian extreme. The number in parentheses indicates at what level of satisfaction a significant difference is found between the two ranked distributions. The leximin-equivalence of

Figure 3: Ordinal SWFs: dominance measures

(a) 1st order stochastic dominance



(b) Leximin (first sign. level difference in par.)



Norway and Sweden implies that their distribution functions are not significantly different at *any* satisfaction level. Note that the statistical test statistics are here modified by a Holm-Bonferroni scaling, so that we take into account of the fact that we are making multiple comparisons. This explains why Norway and Sweden cannot be ranked by leximin, despite the fact that Norway dominates Sweden for a head count ratio at $stf = 2$ (using unadjusted p-values). Again, Denmark dominates all other countries. While Finland dominates both Norway and Sweden, the cumulative distributions are not significantly different until we reach $stf = 7$. In contrast, Nordic dominance over U.K. and Germany can be established already at the bottom level of satisfaction. Note that leximin need not be transitive in finite samples, although it did result in a transitive ranking in this case.

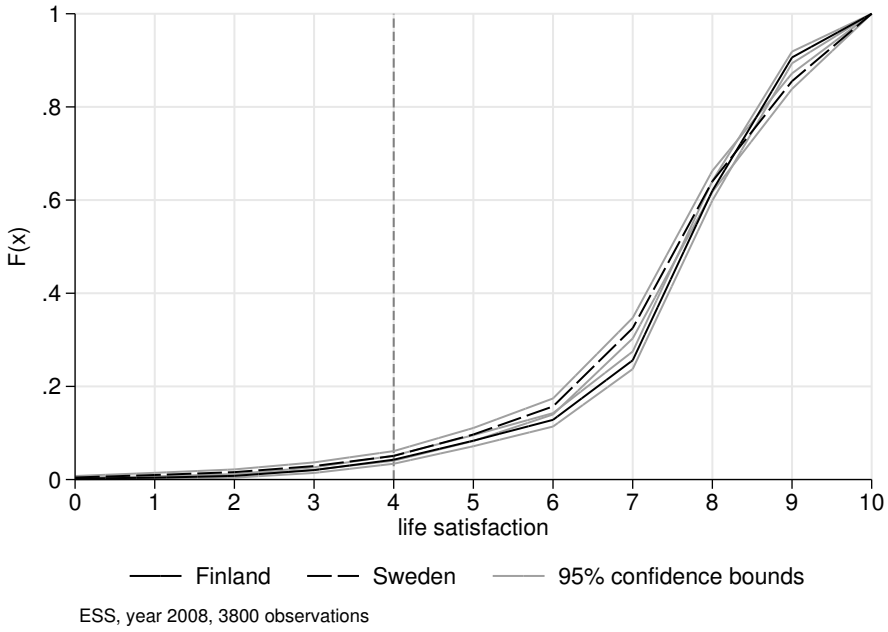
We finally turn first-order stochastic dominance. As discussed above, this is probably the most general ranking we can hope for, since this ranking only assumes anonymity (impartiality) and Pareto (and, if comparing populations of different sizes, the principle of population). The potential problem with stochastic dominance is that it may lack in

power. I.e., since it makes so weak ethical assumptions on comparability, in addition to its weak measurability assumptions (ordinal interpersonal comparability), can we really say anything? Surprisingly, it turns out in figure 3a that we can say a great deal using only the minimal assumptions of stochastic dominance. Denmark dominates all other countries under study, while all Nordic countries dominate both U.K. and Germany. This means that under *any* SWF that is increasing, satisfies anonymity, and the principle of population—including all SWFs considered in this paper and in the entire standard literature—adheres to this ranking. Note the difference between the pair Norway-Sweden and Finland-Norway: the former is comparable under stochastic dominance and found equivalent (their distribution functions are not significantly different at any point), while the latter pair is *not* comparable under stochastic dominance, since their distribution curves intersect (statistically significantly). Also Finland and Sweden are non-comparable under first-order stochastic dominance.

One can also consider a modification of first-order stochastic dominance, studying the cumulative distribution up to a certain cutoff, say $stf = 5$. This would then be similar to a head count ratio, but also considering the distribution under the cutoff point. For the sample here, stochastic dominance capped at $stf = 5$ turns out to induce the same ranking as in figure 3a, except that now Fin $\sim$ Nor and Fin $\sim$ Swe (whereas these pairs are non-comparable when considering the entire distribution).

As a way of rounding off the discussion of ordinal social welfare functions, let us consider the case of Finland vs Sweden. Figure 4 shows their distribution functions, along with 95 percent confidence intervals. We see that for the head count ratio at $stf = 5$, marked as a dashed line at $F(4) = 1 - P(stf \geq 5)$, the CDFs are not significantly different, in line with figure 2a (at $F(1)$ the difference $D_1(x) - D_2(x)$ is significant when using unadjusted confidence intervals, consistent with figure 2b). Using Holm-Bonferroni adjusted p-values,

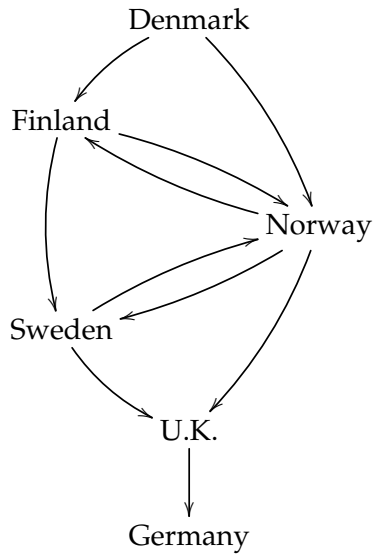
Figure 4: Finland vs. Sweden (confidence intervals not adjusted for multiple comparisons)



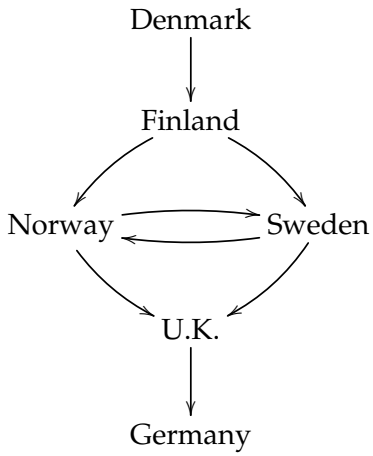
	$D_1(x) - D_2(x)$	s.e.	t-value	p	adj. p
0	−0.002	0.002	−1.173	0.241	0.482
1	−0.006	0.003	−2.095	0.036	0.217
2	−0.008	0.004	−2.242	0.025	0.175
3	−0.009	0.005	−1.703	0.089	0.443
4	−0.008	0.007	−1.172	0.241	0.241
5	−0.013	0.009	−1.424	0.155	0.619
6	−0.029	0.011	−2.512	0.012	0.096
7	−0.069	0.015	−4.664	0.000	0.000
8	−0.020	0.016	−1.285	0.199	0.596
9	0.051	0.011	4.799	0.000	0.000

Figure 5: Cardinal SWFs

(a) Mean life satisfaction



(b) 2nd order stochastic dominance



the first significant difference between the CDFs is found at $stf = 7$, resulting in Finland leximin-dominates Sweden. However, since Sweden has a significantly higher share of the population in the highest category, i.e., $F_{Swe}(9) < F_{Fin}(9)$, no country first-order stochastically dominates the other. A prioritarian would surely prefer the more egalitarian distribution of Finland, but in the end, what distribution one prefers must ultimately depend on some underlying moral premise. We will return to this example later on.

Having considered ordinal social welfare functions, the question arise whether it is possible to arrive at more complete rankings if we add the assumption that the underlying SWB measure is cardinally measurable. We remind the reader that this is a considerably stronger assumption than mere ordinal comparability, as discussed in section 3 (table 1). We start out with the conventional measure—the population mean of well-being—and the resulting ranking is depicted in figure 5a. Interestingly, the ranking obtained is similar to first-order stochastic dominance, except that Finland has significantly higher satisfaction

mean than Sweden. However, this gain—being able to rank Finland and Sweden—comes at a high cost: mean-ranking as a social welfare criterion is only valid if one accepts the moral doctrine of utilitarianism.

The final ranking we consider is that of second-order stochastic dominance. This again results in similar ranking as previous aggregation rules. It is illustrative to compare this ranking with that implied by the leximin criterion; while these two SWOs result in exactly the same ordering (in this application), they rely on very different assumptions. Regarding measurement, leximin is the more conservative SWO, assuming only ordinal measurability of the well-being measure. In contrast, second-order stochastic dominance require full cardinal measurement, with both intra- and interpersonal comparability. When it comes to ethical assumptions, however, the leximin is an implementation of a specific moral theory, while second-order stochastic dominance is considerably more general, with $A \underset{2.s.d.}{\geq} B$ implying $A \geq B$ for *any* (mean) utilitarian or prioritarian SWO that satisfies Pareto and anonymity. Similar to first-order dominance, we can restrict the range so as to compare second-order stochastic dominance only up to some threshold. For example, with an upper threshold of $stf = 5$, Finland no longer dominates Norway and Sweden, but otherwise the ranking stays the same.

In summary, it turns out that for our sample of countries, different aggregation methods result in surprisingly similar rankings. In all cases, Denmark dominates all others, and Germany-U.K. are dominated by the Nordic countries.

The assumption of cardinality does give some leeway in relaxing the ethical assumptions and still be able to further rank the Nordic countries, but the gain is not large. Above all, there seems to be no reason at all to use mean-ranking as a way of comparing countries: mean-ranking is dependent on *both* the measurement assumption of cardinal full comparability and the ethical assumption of mean-welfarism. *If* cardinality is to be

Table 3: Comparison of social welfare rankings across years

	year 2002		year 2008		year 2014	
	#ties	#non-comp.	#ties	#non-comp.	#ties	#non-comp.
Ratio satisfied	3	0	3	0	7	0
1st order SD	2	0	3	2	2	1
1st order SD ≤ 5	3	0	3	0	5	0
Leximin	2	0	1	0	1	0
Mean	3	0	2	0	4	0
2nd order SD	2	0	1	0	1	0

assumed, second-order stochastic dominance is, it turns out, just as powerful, and additionally makes less ethical presumptions. But the main takeway from this section is surely that first-order stochastic dominance works surprisingly often, and, since it is very light on both ethical and measurement assumptions, there is no reason to use any other SWO for ranking pairs of countries. Table 3 shows the result of performing the same analysis as above also for the years 2002 and 2014. In neither case does there appear to be large gains from imposing stronger assumptions with respect to measurement or ethics.

However, if we instead want to be able to assign *metrics* to different well-being distributions, we need to restrict our choice set to social welfare *functions*; hence, stochastic dominance is not an option. In this case, rather than selecting an *ad hoc* overall metric, I would suggest using a collection of head count ratio measures. This is equivalent to sample the cumulative distribution functions at a number of predetermined points. In this way, the researcher simply provides a description of the well-being distributions, and the ethical considerations are left to the discretion of the reader. Table 4 collects head count ratios for all countries in our sample at three thresholds, together with the cardinal mean for comparison.

Table 4: Comparison of aggregation metrics, 95% confidence intervals in brackets

	Ordinal			Cardinal
	$\mathbb{E}[stf \geq 2]$	$\mathbb{E}[stf \geq 5]$	$\mathbb{E}[stf \geq 8]$	$\mathbb{E}[stf]$
DK	1.00 [0.99–1.00]	0.97 [0.96–0.98]	0.82 [0.80–0.85]	8.47 [8.38–8.57]
FI	1.00 [0.99–1.00]	0.96 [0.95–0.97]	0.74 [0.72–0.76]	7.93 [7.86–8.00]
NO	1.00 [0.99–1.00]	0.95 [0.94–0.96]	0.70 [0.68–0.73]	7.88 [7.78–7.98]
SE	0.99 [0.99–1.00]	0.95 [0.94–0.96]	0.68 [0.66–0.70]	7.85 [7.76–7.93]
DE	0.98 [0.97–0.98]	0.86 [0.84–0.87]	0.49 [0.47–0.51]	6.91 [6.81–7.00]
GB	0.98 [0.98–0.99]	0.88 [0.87–0.90]	0.52 [0.50–0.54]	7.09 [6.99–7.18]

Finally, it should be noted that comparability with respect to stochastic dominance is dependent on sampling error; with larger sample sizes, we would in all likelihood have found significant differences at more points of the distributions, perhaps making at least some pairs non-comparable. That being said, the rankings do show a surprising degree of consistency across SWOs, so it remains an open question whether larger sample sizes really would result in a different picture.

5 Conclusion

Any method of constructing an aggregate metric out of individual well-being is associated not only with measurement assumptions on the underlying SWB measure, but also with normative assumptions, embodied in the specific functional form of the aggregation rule.

By comparing a number of aggregation metrics for a number of countries, we find

that we can, in general, get away with surprisingly weak assumptions, with respect to both measurement and ethics. When comparing two countries, first-order stochastic dominance can rank most country pairs in our sample. In contrast, using the mean of individual SWB for ranking countries has little going for it. If we feel comfortable in accepting the assumption of cardinal full comparability for our SWB measure, second-order stochastic dominance has similar power, yet does not depend on a specific moral philosophical doctrine. It seems that only if one is already a staunch utilitarian, it is motivated to use the mean.

If we instead need metrics to capture aggregate well-being, a tuple of headcount ratios is a possibility. For cardinally measured SWB, one can imagine calculating $\int_{-\infty}^{x} F(y)\mathrm{d}y$ for different values of x (similar to how headcount ratio measures calculate different points in $F(x)$).

References

Berlin, M. and N. Kaunitz (2015). Beyond Income: The Importance for Life Satisfaction of Having Access to a Cash Margin. *Journal of Happiness Studies 16*(6), 1557–1573.

Bossert, W. and J. A. Weymark (2004). Utility in social choice. In S. Barbera, P. Hammond, and C. Seidl (Eds.), *Handbook of Utility Theory, vol. 2: Extensions*, pp. 1099–1177. Boston: Kluwer Academic Publishers.

Davidson, R. and J.-Y. Duclos (2000). Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica 68*(6), 1435–1464.

Di Tella, R. and R. MacCulloch (2006). Some Uses of Happiness Data in Economics. *Journal of Economic Perspectives 20*(1), 25–46.

Diener, E. (2000). Subjective Well-Being: The Science of Happiness and a Proposal for a National Index. *American Psychologist 5*(1), 34–43.

Diener, E., E. Suh, R. Lucas, and H. Smith (1999). Subjective Well-Being: Three Decades of Progress. *Psychological Bulletin 125*(2), 276–302.

Donaldson, D. and J. A. Weymark (1980). A single-parameter generalization of the Gini indices of inequality. *Journal of Economic Theory 22*(1), 67–86.

Fleurbaey, M. (2009). Beyond GDP: The Quest for a Measure of Social Welfare. *Journal of Economic Literature 47*(4), 1029–75.

Frey, B. and A. Stutzer (2002). What Can Economists Learn from Happiness Research? *Journal of Economic Literature 40*(2), 402–435.

Helliwell, J., R. Layard, and J. Sachs (2015). World Happiness Report 2015. LSE Research Online Documents on Economics, London School of Economics and Political Science, LSE Library.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*(2), 65–70.

Kahneman, D., A. B. Krueger, D. Schkade, N. Schwarz, and A. Stone (2004). Toward National Well-Being Accounts. *American Economic Review 94*(2), 429–434.

Layard, R. (2005). *Happiness: Lessons from a new science*. London: Allen Lane.

Lerner, A. P. (1944). *The economics of control: principles of welfare economics*. New York: Macmillan.

OECD (2016). Society at a Glance 2016: OECD Social Indicators. Technical report, OECD Publishing, Paris.

Parfit, D. (1984). *Reasons and persons*. Oxford: Clarendon Press.

Roberts, F. (1979). *Measurement theory, with applications to Decision Making, Utility and the Social Sciences*. Boston: Addison-Wesley.

Schokkaert, E. (2007). Capabilities and Satisfaction with Life. *Journal of Human Development and Capabilities 8*(3), 415–430.

Sechel, C. (2014). Subjective Well-Being across Countries: A Headcount Aggregate. Technical report.

Sen, A. K. (1979). Utilitarianism and Welfarism. *Journal of Philosophy 76*(9), 463–489.

Sen, A. K. (1985). *Commodities and capabilities*. Amsterdam: North-Holland.

Stiglitz, J. E., A. Sen, and J.-P. Fitoussi (2010). *Mismeasuring our lives: why GDP doesn't add up. . . .* New York: New Press.

Tännsjö, T. (1998). *Hedonistic utilitarianism*. Edinburgh: Edinburgh University Press.