Excerpts from Louis Kaplow, The Theory of Taxation and Public Economics (2008).

3

The Social Objective

Optimal income tax analysis employs the standard welfare economic approach to policy assessment. Because much analysis of tax policy and related subjects in public economics does not employ this method, this chapter begins by explaining the need for explicit attention to the social objective function. Many of the policies to be examined pose tradeoffs, not only in choosing the overall extent of redistribution but also in making many specific design decisions, and coherent formulation of these tradeoffs is often impossible without direct reference to the social objective. Such explicitness is also helpful in articulating research agendas because it is necessary to identify which factors and effects need to be analyzed in the first place, and for many tax policies the pertinent list is difficult to discern unless the social objective is specified.

This chapter then elaborates the welfare economic framework in the form ordinarily used in optimal income tax analysis. In so doing, attention is given to the difference between a social preference for equality due to the formulation of the social welfare function (SWF) and a preference for equality arising from the concavity of individuals' utility functions. Comments are offered on the relationship between a variety of theories of distributive justice and particular SWFs and on the redistributiveness of some standard SWFs. Finally, the chapter explains the impact, or lack thereof, of the particular choice of SWF (from within a standard class) on the analysis in the rest of the book.

A. Motivation

This section begins by offering a variety of examples that illustrate the range of demands regarding social evaluation of tax policy. Then it draws out their implications.

1. Examples

Most obviously, it is necessary to be able to measure the benefit of greater equality quantitatively in such a manner that one can answer questions like whether a given increase in equality is worth a specified reduction in average income.[1] One also needs to be able to assess reforms whose effects on inequality are more complex. For example, replacing a graduated income tax with a flat tax—one that raises the same revenue and has a moderately higher exemption—may well improve the lot of both the near-poor and the rich at the expense of the middle class.[2] Whether such a redistribution raises or lowers social welfare is not immediately obvious.

Tax design routinely raises even more subtle distributive questions, such as those involving whether to substitute administratively simpler but less accurate forms of presumptive taxation, how much to spend on reducing assessment errors, and the optimal degree of randomization in enforcement. In each case, cheaper solutions may result in some individuals paying more and others less tax than would be ideal, thus posing another sort of tradeoff between efficiency and distribution.

Additionally, there are important elements of heterogeneity in the population that raise distributive issues. For example, it is generally supposed that unhealthy or disabled individuals should pay less (or receive more), ceteris paribus, but it is not obvious to what extent this should be done or how significant the welfare loss would be if it were not. Even greater systematic heterogeneity involves differences in family units, concerning both the number of and relationship among adults in the same household and the number and ages of children or other dependents. It is necessary to determine the appropriate relative tax burdens (or subsidy entitlements) across family types, a matter of substantial dispute that has been resolved differently over time and across jurisdictions.

Other specific tax policy problems involve further complications. For example, in assessing estate and gift taxation, it is insufficient merely

[1] This familiar requirement has important although often unrecognized implications for the measurement of inequality, poverty, progressivity, and redistribution, as discussed in section 15.A.

[2] For further details, see note 9 in chapter 2.

to consider such matters as a possible distortion of savings. Transfers affect the utility of both donors and donees—thus involving an externality of sorts—and transfer taxation will likewise affect both groups. Relatedly, such private transfers involve voluntary redistribution.

Even if one has all of the pertinent facts, there still exists a need to make a balancing judgment in each of these instances. Knowing certain elasticities will help, but it is not immediately obvious, especially in the latter sets of examples, which elasticities are most relevant, what other information is required, and how all the inputs should be combined to form a social decision. As a corollary, setting research agendas in many of these areas is hardly straightforward.

2. *Implications*

The welfare economic approach to social assessment is designed to address the issues raised by the examples in subsection 1. This method of evaluation has two aspects: specification of a common denominator and adoption of a method of aggregation. First, one determines the effects of any policy under consideration on each individual's utility—also referred to as an individual's well-being or welfare. Thus, whether considering matters involving accuracy or assessment errors, treatment of different family units, or estate and gift taxation, the necessary (and sufficient) positive analysis entails identifying policies' consequences for each individual. Second, to form a social assessment, the information on everyone's utility is aggregated using an SWF, in particular an individualistic SWF, indicating that social welfare is a function (only) of individuals' utilities. The chosen functional form implicitly indicates the weight given to equality, whether, for example, the benefits to the near-poor and rich under the postulated flat tax reform outweigh the added burdens on the middle class, and how one should evaluate the effects of estate and gift taxation on donors, donees, and other individuals.

Although this method of social assessment, which is elaborated in section B, is widely accepted by economists in principle (though less so by others), it is not usually employed explicitly in studying many features of tax policy. Often analysts, following a range of prominent public finance economists from Musgrave and Musgrave (1973) to Stiglitz (2000, pp. 456–481), refer to the multiple objectives of tax policy, offering lists that typically include efficiency, fairness or equity (itself usually stated to be multidimensional, including horizontal and vertical equity, among other principles), revenue adequacy, simplicity, and administrability. Such formulations obviously suffer from a lack of a common denominator and of a principle of aggregation. Many of the objectives (for example, simplicity and administrability) relate to efficiency even though they are separately listed, and they may also involve distributive concerns and thus be related to fairness or equity.

As will be explored further in chapter 15, these fairness or equity notions are particularly problematic. Many, such as the familiar "ability to pay" principle, are highly indeterminate. Additionally, there exist multiple, often conflicting, notions of tax equity that seem to be invoked selectively, in an ad hoc manner, guided largely by intuition. Some of these concepts are incomplete or even incoherent. For example, the insistence that taxes used to finance public goods adhere to one of the so-called sacrifice theories or the benefit principle is meaningful only if there is no other taxation: If redistributive taxation is also present—and by definition it need not adhere to such principles (how can redistributive taxes and transfers involve equal sacrifice under any of the sacrifice theories?)—there is no real constraint on the tax system as a whole. Moreover, some of these principles are arbitrary even when considering public goods in isolation. Notably, under the sacrifice theories, it is not true in general that tax burdens equal benefits from public goods, so redistribution is involved, and the magnitude of this redistribution is determined by the amount of public goods provided. Thus, the technological happenstance of which goods are "public" and which of those happen to be efficient to provide determines the extent of taxation and, accordingly, the extent of redistribution that results.

Both the lists of general criteria and particular notions of fairness and equity can best be understood as loose, intuitive proxies for social welfare, or at least for aspects thereof.[3] Generally, simplicity is a virtue (ceteris paribus, that is); likewise for administrability. Greater efficiency is better (again, ceteris paribus). And some notions of equity are related, even if indirectly, to a coherent notion of social welfare.

[3] See subsection 13.A.3.c and chapter 15.

This loose and incomplete relationship between many asserted objectives of tax policy and social welfare is not, however, sufficient in many settings. Ultimately, it is necessary to state with some precision one's actual objective function. Without doing so, many of the examples in subsection 1 cannot be analyzed meaningfully. For instance, without a specified SWF, how is one to trade off the greater simplicity and administrability of coarser rules against the resultant reduction in equity? Furthermore, partial analyses are sometimes taken (perhaps by policymakers unaware of their proxy character) as complete, which can be affirmatively misleading.

Regarding the latter concern, a problem of great relevance to the subject of this book involves analyses that examine only inefficiency, often captured by measures of deadweight loss. This approach is legitimate when there are no distributive effects. Thus many studies of inter-asset or inter-sector distortions in the taxation of capital may provide a reasonably complete indication of overall effects on social welfare because of the similar pattern of ownership of different physical and financial assets. Moreover, as suggested in chapter 2, it is often best—precisely because of the normative relevance of distributive effects—to undertake distribution-neutral policy comparisons, in which case only efficiency is at stake.

However, distribution-neutral analysis is not the norm in many areas of inquiry, yet sometimes distribution is nevertheless ignored.[4] The use of representative-agent models—wherein everyone is assumed to be identical, in which case there are no distributive effects—does not legitimize application of the results to the actual world of heterogeneous individuals. The problem is especially acute when the tax instruments under analysis, notably the income tax, are employed precisely on account of distributive concerns. For example, many analyses of environmental policies—comparisons of regulations, taxes, permit systems, and the like—report deadweight loss estimates as the measure of social welfare even though the deadweight loss arises significantly (sometimes predominantly) on account of differences in the redistributiveness of the policies

[4] This point is developed further in subsection 8.C.3 with regard to government expenditures on goods and services, and it is applied to regulation in section 8.G.

under consideration. Two policies may have similar effects on the environment, but one may involve greater use of redistributive income taxation. The more redistributive policy will cause greater distortion and accordingly be deemed inferior, but the greater distortion is due precisely to the fact that the policy increases redistribution. The presumed benefit of such greater redistribution, however, is ignored in the welfare assessment. To avoid such problems, even policies like certain environmental interventions that may appear to be unrelated to distributive issues must be analyzed explicitly in terms of a well-specified SWF—or, alternatively, one must in fact hold distribution constant or employ some other suitable technique if one is to justify analysis in terms of partial or proxy criteria.

As suggested by the examples in subsection 1, ignoring distribution when it may be relevant is but a part of the overall problem. In some settings, such as in deciding how to trade off inequality and efficiency or in assessing replacement of a graduated income tax with a flat tax, distributive concerns are at the fore. And in many other contexts, such as those involving accuracy and error, taxation of the family, and estate and gift taxation, the challenge concerning more subtle distributive effects may not be one of avoiding their accidental omission but rather one of determining how to incorporate them.

Explicit attention to the social objective offers the solution to these problems. In the study of tax policy, this is currently done mainly in the field of optimal income taxation. Indeed, one of the important (although underemphasized) contributions of Mirrlees (1971) was precisely his synthesis of positive and normative analysis. However, much work on taxation and other subjects in public economics is not seen primarily as posing the tradeoff between redistribution and efficiency, and it is not generally conducted using this inclusive framework. There is tremendous variation in the extent to which existing work is deficient because of the failure to undertake explicit social welfare analysis. As will be seen in later chapters, careful attention to the social objective will in many instances affirm the validity of existing understandings. In other settings, prevailing results will need to be modified. And sometimes it will be necessary to reorient thinking substantially. Furthermore, many issues—like taxation of the family—have proved largely intractable, yet significant progress is possible when analysis is related directly to an SWF. Proceeding in the spirit of Mirrlees, rather than invoking familiar lists of tax policy criteria and various notions of fairness and equity, has the potential to illuminate and guide analysis of a wide array of subjects in taxation and other areas of public economics. Important prior work demonstrates the value of such an approach, and this book seeks to apply it in additional settings.

B. Exposition

This section begins by presenting the standard welfare economic approach as it is applied in the assessment of income redistribution. Next, a range of possible SWFs is discussed. Finally, remarks are offered concerning the relevance of the choice of a particular SWF to the analysis in the rest of this book.

The task here is purely expositional. Matters of normative justification are deferred to chapters 13 and 14 because most economists and other policy analysts are at least roughly comfortable with the welfare economic approach and because a number of the most controversial philosophical issues do not in any event relate very directly to most of the analysis in intervening chapters.

1. Social Welfare Functions

A social welfare function $SW(x)$ indicates how any regime or social state x (taken as a complete description thereof) is evaluated. Here we are concerned with individualistic SWFs, wherein social welfare depends only on individuals' utility or well-being. The normative premise, referred to by Sen (1977, 1979) as "welfarism," is that the only relevant aspect of a regime is the manner in which it affects each individual's well-being. An implication is that notions of fairness or equity have no role unless they are concerned with the distribution of utility or they are in some respect a proxy for effects on utility. (For further elaboration, see section 13.A.)

In assessing redistributive taxation, it is common to use an additive social welfare function that assumes a continuous population.

$$SW(x) = \int W\big(u_i(x)\big) f(i)\, di, \tag{3.1}$$

where u is a utility function, subscripts index individuals' types, and $f(i)$ is the density of type i individuals in the population.[5] Because welfare is taken to be the integral of some transformation W of individuals' utilities, this functional form for the SWF is not necessarily utilitarian; the functional form of W on the right side of (3.1) incorporates a view of distributive justice. This can be seen from the following formulation used, for example, in Stern (1976).

$$\begin{aligned} SW(x) &= \int \frac{u_i(x)^{1-e}}{1-e} f(i)\, di, \text{ for } e \neq 1 \\ &= \int \ln u_i(x) f(i)\, di, \text{ for } e = 1, \end{aligned} \tag{3.2}$$

where $e \geq 0$ indicates the degree of aversion to inequality in the distribution of utility levels.[6] Thus, $e = 0$ indicates that social welfare is the sum of utilities—utilitarianism—and taking the limit as e approaches infinity yields the maximin formulation associated with Rawls (1971), under which all weight is placed on the utility of the least-well-off individual.[7]

It is useful to distinguish two different factors in expression (3.2) that may favor a more egalitarian distribution of disposable income. The magnitude of e has already been identified as one factor: The greater is e, the greater the increase in social welfare due to a given redistribution from an individual with a higher utility to one with lower utility, ceteris paribus. The second factor is the concavity of u itself, that is, the rate

[5] For a finite population of n individuals,

$$SW(x) = \sum W(u_i(x)),$$

where the summation is over i from 1 to n.

[6] To explain the latter version in (3.2), for the case in which $e = 1$, the numerator in the former may alternatively be written as $u_i(x)^{1-e} - 1$ (subtracting the constant having no effect on the ordering of social states). Then, taking the limit as e approaches 1 (using l'Hôpital's rule) yields the latter expression.

[7] For more on Rawls and maximin, see subsections 13.B.4 and 14.A.1.a.

at which individuals' marginal utility of consumption falls as consumption rises—equivalently, individuals' degree of risk aversion. To elaborate, consider the oft-used constant-relative-risk-aversion utility function,

$$\begin{aligned} u(c) &= \frac{c^{1-\rho}}{1-\rho}, \text{ for } \rho \neq 1 \\ &= \ln c, \text{ for } \rho = 1, \end{aligned} \tag{3.3}$$

where c denotes consumption (typically, income after taxes and transfers) and ρ is individuals' coefficient of relative risk aversion.[8] The case in which $\rho = 0$ is one of risk neutrality; higher levels of ρ indicate greater risk aversion, which likewise indicates a greater rate at which the marginal utility of consumption falls as consumption rises. Hence, ceteris paribus, a higher ρ also favors greater equality in the distribution of consumption. (Note that, for ease of exposition in this section, the formulation in expression (3.3) abstracts from the effect of labor effort on utility, which obviously will be important in the subsequent analysis of optimal income taxation in chapter 4.)

In some of the literature, including work on optimal income taxation and inequality measurement (see Atkinson 1970, 1973), analysis employs a reduced-form SWF (which, again, abstracts from the effect of labor effort on utility),

$$\begin{aligned} SW(x) &= \int \frac{c_i^{1-\gamma}}{1-\gamma} f(i)di, \text{ for } \gamma \neq 1 \\ &= \int \ln c_i f(i)di, \text{ for } \gamma = 1, \end{aligned} \tag{3.4}$$

where γ indicates the degree of aversion to inequality in the distribution of consumption. That is, in formal analysis and simulations, analysts may consider the overall social aversion to inequality in consumption, without regard to how much of that aversion is attributable to individuals' decreasing marginal utility of consumption (ρ in expression (3.3))

[8] Regarding the case in which $\rho = 1$, see note 6.

and how much is attributable to the concavity of the SWF in utility levels (e in expression (3.2)).

It is important to distinguish these two components because the former is a matter of empirical fact about individuals whereas the latter involves a normative judgment, external to the individuals in question, that must be grounded in a theory of distributive justice.[9] It is worth observing, moreover, that the more concave are individuals' utility functions (the greater is ρ in (3.3)), the less relevant will be the degree of concavity in social welfare as a function of individuals' utility levels (3.2). The reason for this tendency is that when utility is more concave, there may be little relative difference in utility levels even when there are significant relative differences in individuals' marginal utilities of consumption.[10] (A further technical complication that will not be explored here is that, despite appearances, there are difficulties in interpreting the function in expression (3.4) as a simple composite of those in expressions (3.2) and (3.3).[11])

2. Comments on a Range of Social Welfare Functions

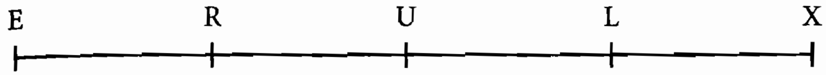
Although, as the next subsection explains, the overall redistributiveness of the SWF (which may be captured in the formulation of expression (3.4) by the composite parameter γ) does not qualitatively affect most of the analysis in this book, it provides a useful orientation to consider a

[9] Subsequent authors interpreted the parameter in Mirrlees (1971) corresponding to e as entailing a normative judgment about the SWF, whereas Mirrlees (1982, p. 77 n. 21) indicates that he adheres to a utilitarian norm, with the parameter e indicating possible degrees of concavity of individuals' utility functions.

[10] See Kaplow (2003a). As explained there, this result is more pronounced as consumption rises. Therefore, concavity in the welfare function is likely to matter most at the bottom end of the income distribution (for example, to the design of transfer programs that differ in their treatment of the near-poor and very poor) and less so at the top (for example, to choices of tax rate graduation that affect the distribution between upper-middle-income individuals and the rich).

[11] If one directly combines (3.2) and (3.3), one would need to multiply by the constant $(1-\rho)^e$ on the right side of (3.4), which itself would not affect the optimization, and one would have $\gamma = 1-(1-\rho)(1-e)$. However, if one considers the case in which $\rho > 1$ or $e > 1$, this formulation is problematic. For further exploration, see Kaplow (2003a).

variety of possible SWFs and the degree of redistributiveness that they entail. It is helpful to situate such a discussion by considering a spectrum of distributive theories, which might crudely be presented as follows:



At the left end is pure egalitarianism, E, favoring complete equality at all costs. Next is a Rawlsian view, R, indicating that one should continue redistribution toward the least-well-off individual until the point at which further redistribution actually worsens that individual's situation. (Note that this is the most egalitarian view that is not obviously inconsistent with the Pareto principle.) Utilitarianism, U, favors the degree of redistribution that maximizes the sum of individuals' utilities. Libertarianism, L, opposes all redistribution. At the right end of the spectrum, denoted by X, one might consider a view under which as much wealth as is feasible should be concentrated in the hands of one or a select few (perhaps a governing elite that deems itself so entitled).

The formulation in subsection 1, entailed by the SWF in expression (3.2), encompasses the range from R to U and thus is restrictive. Positions to the left of R might be ruled out by invoking the Pareto principle. (Indeed, even position R is understood to be quite extreme, for it would favor a regime that reduces everyone to complete misery as long as it promotes the well-being of the most miserable individual by even an infinitesimal amount.[12]) If one adopts position L, no redistribution is warranted, rendering rather uninteresting the inquiry into how best to redistribute income. (Of course, this still leaves potentially relevant territory between U and L uncovered.) And position X, although it unfortunately seems to approximately describe many actual regimes in history, commands little normative support.

Much argument about the proper SWF proceeds by simple intuition. At one time, utilitarianism was seen as quite radical. Today, however,

[12] See, for example, Arrow (1973, p. 251) and also the further discussion in subsection 14.A.1.a.

many progressively minded economists and philosophers view utilitarianism as insufficiently concerned with equality. Such expressions are more aesthetic hunches than arguments, as explored in section 14.A. Moreover, they suffer from a number of specific drawbacks: Such views often fail to appreciate how much redistribution is implied by various SWFs; intuitions may well be based on equality of income (which is directly observable and thus more tangible) rather than on that of utility; and the optimal degree of redistribution entailed by any given SWF is itself a subtle matter that depends on a range of parameters concerning individuals' utility functions (the concavity parameter ρ and also parameters relating to the elasticity of substitution between consumption and leisure), the shape of the distribution of ability, the available technology of taxation, and a variety of other potentially important factors. Accordingly, any view about how much redistribution is appropriate that eschews normative argument and technical analysis is difficult to maintain.

Another form of argument about the choice of an SWF proceeds from the concept of equality, but equality has diverse and conflicting meanings. In a formal sense, the SWF described in expressions (3.2) and (3.4) is egalitarian in that each individual's utility is treated symmetrically and thus given equal weight. A libertarian view that opposes any redistribution also entails a formal sense of equality. Some favor equality of opportunity, but that too can be interpreted to favor just about anything from a libertarian view to a purely egalitarian one. If one looks at the specific method of aggregation, a utilitarian view gives equal weight to contributions to each individual's utility—that is, each individual's *marginal* utility is treated identically. More egalitarian SWFs place greater weight on equality of utility *levels,* but in the process do not treat contributions to each individual's utility equally. Thus, a belief in equality in the abstract is hardly a sufficient basis for choosing an SWF, even if one posits that equality per se should be the sole criterion for decision.

Finally, consider briefly the degree of redistribution implied by various SWFs. Attention here will be confined to a utilitarian SWF, but as the discussion in the preceding subsection suggests, one can to some extent translate these results to more concave SWFs by reinterpretation of the various parameters—specifically, by considering a value of γ in expression (3.4) that exceeds ρ in (3.3). Initially, suppose that there are no incentive effects associated with redistribution. Then, in the simple case in which all individuals have the same utility function and there is some degree of diminishing marginal utility of consumption ($\rho > 0$ in expression (3.3)), a utilitarian SWF favors perfect equality, as suggested by Edgeworth (1897). (Lerner (1944) extends this result to the case in which individuals' utility functions may differ and the government cannot observe utility functions.) Of course, the result would be the same if the SWF were more concave.[13]

Consider instead the case in which redistribution distorts labor effort. This considerably more complex problem is the focus of chapter 4 and much that follows. But insight can quickly be gleaned by undertaking partial analysis for some simple cases.[14] Suppose first that individuals' utility functions are given by ln c, that is, $\rho = 1$ in expression (3.3). Marginal utility equals $1/c$. For instance, the marginal utility of a poor person with consumption of \$10,000 is ten times that of an upper-middle-class individual with consumption of \$100,000 and is one hundred times that of a rich individual with consumption of \$1,000,000. Thus, even if distortion was so great that for every \$10 taken from the upper-middle-class person (or every \$100 from the millionaire) only \$1.01 reached the poor person, a utilitarian SWF would favor the redistribution. If one instead takes the case of $\rho = 2$, marginal utility is inversely proportional to the square of consumption, so the factors of ten and one hundred in the foregoing example would become one hundred and ten thousand. Accordingly, the extent of deadweight loss from redistribution that is tolerated by a utilitarian SWF is extremely high.[15] Furthermore, it

[13] Sen (1973b) in turn extends Lerner's result to any concave SWF. An interesting feature noted by Mirrlees (1971, p. 201) is that, if labor effort were involved but incentive problems could be avoided, utilitarian redistribution would overshoot the point of equal utility because highly able individuals would optimally work very hard—due to their higher productivity—and on account of the disutility of labor would be worse off than the less able. Thus, a utilitarian SWF would in a sense be more redistributive than more concave SWFs.

[14] The illustrations continue to abstract from the effect of labor effort on utility. Alternatively, one could allow labor to affect utility in an additively separable manner.

[15] These illustrations offer a perspective on studies reporting high efficiency costs of redistribution. For example, Feldstein and Feenberg (1996) find that the 1993 tax rate increase

should be noted that utility function concavity parameters in the range of 1 to 2 are widely considered plausible, and some estimates (notably, from the finance literature) are substantially higher.[16] Hence, SWFs in the range from U to R, which are commonly examined in the optimal income tax literature, entail substantially redistributive social preferences.

3. Relevance of the Choice of a Particular Social Welfare Function

As it turns out, the degree of overall weight on equality in the SWF—the level of γ in (3.4), whether attributable to concavity in *u*, as indicated by ρ, or concavity in *W*, as indicated by *e*—often has little or no effect on the qualitative nature of many results in this book. One reason for this is that, as section 2.C indicates, much of the analysis will be undertaken using distribution-neutral comparisons, in which case the social weight on equality does not matter at all. Just as cost-effectiveness analysis is useful in determining, for example, how to save the most lives for a given expenditure on safety—without having to specify a value of life—so too the determination of how best to achieve a given degree of redistribution is essentially independent of how much social value is placed on that redistribution.

Some analysis, particularly that in the next chapter on the optimal extent of redistribution, will depend on the SWF. One can, however, explore results for a range of parameters concerning the overall preference for equality (γ), allowing such results to be interpreted for various

in the United States, which was concentrated on high-income individuals, involved so much distortion that for each dollar raised, high-income taxpayers were worse off by two dollars. Clearly, the marginal social value of a dollar to those who benefited could well have been greater, especially given that some of the revenue was used to fund an increase in the Earned Income Tax Credit. See also Browning and Johnson (1984) (finding in their benchmark case that the income-equivalent loss to individuals in the top three quintiles is $3.49 per dollar of income-equivalent gain to those in the lowest quintile) and Browning (1993) (finding that the marginal efficiency loss, taking into account the value of additional leisure, is $3.23 from a supplemental proportional tax that finances an equal grant to all households).

[16] See, for example, Barsky et al. (1997), Campbell (1996), Choi and Menezes (1992), and Kocherlakota (1996).

combinations of empirical estimates of the concavity of utility functions (ρ) and views of the appropriate concavity of the SWF (e). Typically, the effect will be one of degree: The greater the social preference for equality, the more redistribution will be optimal.

Some problems, however, involve greater subtlety, notably, when a higher marginal utility of consumption is associated with a higher rather than a lower level of total utility. This conjunction does not ordinarily arise when examining income redistribution because, as one increases an individual's income, marginal utility falls as utility rises, and conversely when income is reduced. Accordingly, the posited situation of both utility levels and marginal utility being higher (or both being lower) is generally due to factors other than differences in consumption levels, such as when individuals have different constitutions (for example, disabilities), are in different family units (for example, when children are present), or are beneficiaries of amenities, public goods, or in-kind transfers that influence the marginal utility of consumption differently from how it is affected by changes in consumption levels.

In such cases, the functional form of W matters qualitatively. With a purely utilitarian SWF (that is, W is linear; equivalently, $e = 0$ in (3.2)), only individuals' marginal utilities matter for redistributive purposes; ceteris paribus, redistribution toward higher-marginal-utility individuals is always favored. But if W is strictly concave (equivalently, $e > 0$ in (3.2)), then redistribution toward lower-utility-level individuals is also favored. The more concave the SWF, as a function of utilities, the more utility levels matter and the less individuals' marginal utilities matter. In the limit, as e approaches infinity (maximin), only the utility level of the least-well-off individual counts, regardless of how low that person's marginal utility is or how high the marginal utilities of others are. Thus, the direction of optimal redistribution can in some settings depend on the shape of the SWF.

Foreshadowing chapter 12, an important illustration of this phenomenon arises in connection with issues concerning the proper unit of taxation. For example, adults who voluntarily choose to have children and succeed in having healthy children will presumably achieve a higher level of utility as a consequence (as implied by their choice to conceive children). They may also, however, have a higher marginal utility of consumption because their available resources now must be shared among a greater number of individuals. If the SWF is utilitarian or slightly concave, redistribution toward families with children will tend to be optimal, whereas if the SWF is strongly concave in utilities, redistribution would optimally be away from families with children. In such instances, the approach taken in this book will be to present the possibilities, explaining how the nature of the SWF may affect the character of optimal treatment.

The relevance of the choice of SWF to optimal tax policy varies greatly: often with no effect, sometimes a difference in magnitude, and occasionally a difference in direction. Nevertheless, it is usually the case that analysis can proceed without a commitment to a particular SWF. Accordingly, further exploration of the choice of an SWF will be deferred to part V. Likewise deferred are a range of other normative issues, such as the merits of welfarism, the nature of well-being, the possibility of interpersonal utility comparisons, and the question of whose utility should be included in the SWF.

13
Welfare

In the standard welfare economic framework sketched in chapter 3 and used thereafter, policies are assessed by reference to a social welfare function (SWF) that aggregates the effects of policies on each individual's well-being (utility). Chapter 14 will consider the process of aggregation itself. The present chapter focuses on what social welfare is taken to be a function of. Considered first is the doctrine referred to as welfarism, under which social welfare is taken to depend on individuals' levels of well-being and on nothing else. Then attention shifts to the meaning of the concept of well-being and an examination of various factors that some analysts believe require modifications of individuals' utilities if they are to serve as proper arguments of an SWF. These two subjects are of more than abstract academic interest: Some economists, including Sen (discussed in subsection B.4), advance views that appear to call for wholesale deviations from welfarism; certain prominent tax equity norms, including some of those examined in chapter 15, conflict with welfarism; and important proposed extensions to optimal tax analysis, for example, those designed to account for interdependent preferences (see section 5.D), are controversial because of disagreement about the proper normative stance regarding the notion of individuals' well-being.[1]

[1] Much of the content in part V draws extensively on the mid-1990s draft of the earlier version of this book mentioned in the preface, and most of the ideas in this chapter, especially those in section A on welfarism, were substantially developed in connection with the writing of Kaplow and Shavell (2002).

A. Welfarism

1. Definition

As stated in subsection 3.B.1, an individualistic SWF—so called because social welfare depends only on individuals' utility or well-being—is a real-valued function of individuals' utility levels. Thus, for the case of a finite population, we can write $SW(x) = W(u_1(x), \ldots, u_n(x))$, where x is the social state or regime. By contrast, a nonindividualistic SWF may be written as $SZ(x) = Z(u_1(x), \ldots, u_n(x), x)$, where Z depends nontrivially on its final argument. The difference between these two formulations is that SW depends on x only through the effect of x on each of the u_i's, whereas SZ at least sometimes depends on x independently of—or in addition to—its effect on the u_i's.

Suppose, for example, that adherence to a specific governmental decision-making process is held to be an intrinsic social good, which is to say that it is deemed to contribute to social welfare independently of the quality of the decisions produced or of how individuals' utilities may be affected by participation in the process itself. (It may well be that following the process tends to improve decisions or to please individuals directly, but these possible benefits are not all that is deemed to matter.) Then the value of SZ would be higher in a state in which the process was followed more often, ceteris paribus.

An equivalent way to state the difference between individualistic and nonindividualistic SWFs is in terms of their information requirements. Knowledge of how a regime x affects each individual's level of well-being $u_i(x)$ is always sufficient to form an assessment under SW but is not always sufficient under SZ. That is, if for two regimes x and x', $u_i(x) = u_i(x')$ for all i, it is necessarily true that $SW(x) = SW(x')$, but it is sometimes true that $SZ(x) \neq SZ(x')$.[2] Continuing with our example, suppose that the specified decision-making process is followed more often in x' than in x, but that other distinctive features of x' precisely offset

[2] If there were never such an inequality in assessments, then each profile of the u_i's would—regardless of the state x that produced it—be associated with a unique real number, so it would be possible to express SZ as an individualistic SWF. (Conversely, if it is sometimes true that, even though everyone's utility is the same in two states, the social assessment differs, then it is impossible to express the SWF in the individualistic form SW.)

the effect of this difference on each individual's utility. The individualistic SWF, *SW*, would judge the two states indifferent, whereas *SZ* would value x' more highly.

Welfarism is the doctrine that social judgments should be in accord with an individualistic SWF. Although this doctrine is largely accepted by economists, there are notable deviations, some of which are explored later in this chapter and others in chapter 15. For example, much work on inequality measurement and most on horizontal equity is inconsistent with welfarism. Furthermore, most twentieth-century moral philosophers and many principles of common morality oppose welfarism; more specifically, many critiques of utilitarianism involve attacks on welfarism (rather than objections to the use of an additive function for purposes of aggregation).[3] Accordingly, the following subsections consider the rationale for welfarism and offer some perspectives on the pertinent debates.

2. Basis for Welfarism

Controversy over welfarism primarily concerns the insistence that the SWF must depend *exclusively* on individuals' well-being. Although this requirement may seem unduly stringent, it is justified by two considerations. First is the lack of an affirmative rationale for giving weight to anything that is truly unrelated to well-being (especially given the encompassing definition of well-being, elaborated in subsection B.1).[4] Singer (1988, p. 152) asks, "But how can something *matter* if it does not matter *to anyone*, or to any group of beings?" Second is the cost of giving

[3] See the arguments of Williams (1973) in the well-known collection *Utilitarianism: For and Against* and also the essays in *Utilitarianism and Beyond* (Sen and Williams 1982). Kaplow and Shavell's (2002) defense of welfarism addresses most of the subjects in this chapter in greater depth as well as other critiques of welfarism. See also Ng (2000a) and, for a defense by a moral philosopher, Hare (1981).

A comment on terminology is in order. Most debates in moral philosophy distinguish between consequentialism—the doctrine that only consequences (whether for individuals' well-being or otherwise) of actions or policies matter—and nonconsequentialism (usually deontological views). Welfarism is a species of consequentialism, and utilitarianism is the most discussed form of welfarism. However, as noted, much debate about utilitarianism concerns welfarism (or often consequentialism) more broadly.

[4] It is difficult to elaborate the point concerning lack of justification in the abstract. Later discussion of particular nonwelfarist criteria will consider the matter in more detail.

weight to other aims: If additional objectives are credited, some tradeoff with, that is, reduction in, well-being is required. Smart (1973, p. 5) presents "a persuasive type of objection" to nonwelfarist principles, namely, the existence of cases in which they "prescribe actions which lead to avoidable human misery."

To develop these points, it is useful to consider a simple series of hypothetical societies. First, imagine a populace that consists of only a single individual. There it would seem difficult to resist welfarism, for what basis could be offered for sacrificing that individual's well-being? Next, suppose that there are n individuals, each identically situated. There may be any manner of interaction among them; it is merely assumed that, say, each individual spends the same amount of time in any particular role and is affected in precisely the same manner as is any other individual. In this case, principles regulating human intercourse, whether informally or through government or other institutions, are potentially applicable. Nevertheless, since each person is affected identically by any regime, it is literally true that what is good for one is good for all. Again, nonwelfarist principles seem hard to justify; giving them weight would entail making everyone worse off. Yet, if a nonwelfarist SWF was correct, it would be applicable to such a society and thus require making everyone worse off whenever it diverged from welfarism.

The notion that nonwelfarist assessment violates the Pareto principle is demonstrated formally by Kaplow and Shavell (2001).[5] The argument is as follows: Beginning with $SZ(x)$, a nonindividualistic SWF, we know from the previous discussion that there exists x and x' such that $u_i(x) = u_i(x')$ for all i but $SZ(x) \neq SZ(x')$. Assume (without loss of generality) that $SZ(x) > SZ(x')$. Assume further that SZ is continuous with respect to some good and that this good has the property that, if each person has more of it, then each person is better off.[6] Now construct x'' from x' by increasing everyone's amount of this good from the levels in

[5] Sen (1970) previously demonstrated a conflict between the Pareto principle and a particular nonwelfarist principle involving (an arguably contentious notion of) libertarian rights. For discussions that relate Sen's argument to the ideas discussed in subsection 3, see, for example, Hardin (1986) and Kaplow and Shavell (2002).

[6] Note that full continuity of SZ, such as with respect to the degree of satisfaction of various nonwelfarist principles, is not required.

x' by a small positive amount. If this increment is sufficiently small, by continuity we have $SZ(x) > SZ(x'')$. However, by construction of x'', it is also true that $u_i(x'') > u_i(x')$ for all i. Combined with the premise that $u_i(x) = u_i(x')$ for all i, this implies that $u_i(x'') > u_i(x)$ for all i. Thus, if the (weak) Pareto principle is satisfied, it must be that $SZ(x'') > SZ(x)$, a contradiction.

The intuition for this result is that a nonindividualistic SWF must give weight to some factor independent of its effect on individuals' well-being. Accordingly, we can compare a social state to another that is identical except in two respects: It is inferior with respect to the nonutility factor, and every individual is ever-so-slightly better off (due to having a bit more of some good). As long as the weight ascribed to the nonutility factor is positive, we can make each individual's utility benefit sufficiently small such that the nonindividualistic SWF favors the state in which everyone is worse off.

The affirmative warrant for nonwelfarist social assessment is thus difficult to comprehend. The puzzle is deepened by the fact that most moral theorists claim to ground their nonwelfarist views in concern for the individual, often under the rubric of freedom, autonomy, or fairness. Yet it is hard to understand to whom any nonwelfarist principle is being fair when all potential subjects of concern may be made worse off by application of that principle. Furthermore, the conflict with the Pareto principle means not only that nonwelfarist prescriptions would fail to command unanimous assent, but also that there exist situations in which there would be unanimous dissent; hence, nonwelfarist views seem inherently to be in tension with notions of freedom and autonomy as well.

3. *Perspectives on Welfarism*

As noted, many philosophers adhere to nonwelfarist views, and some economists advance nonwelfarist ideas, including in the assessment of taxation and redistribution. Additionally, various norms of common morality deviate from pure welfarism. For example, breaking promises is often regarded to be wrongful independently of whether doing so is socially detrimental. And many feel that rewards and punishments should reflect merit or desert aside from any incentive benefits from their doing so.

How can the appeal of nonwelfarist principles be reconciled with the contrary arguments and the lack of any explicit, affirmative rationale? Why do we have moral intuitions in particular contexts that seem to support rules that would, upon reflection, sometimes operate to everyone's detriment? A number of perspectives on these questions have been offered through the ages.[7]

a. Two-level moral theory. It is important to distinguish two different levels at which a moral theory might operate. At the fundamental level, the question addressed is: What is the ultimate criterion of the social good? Advancing individuals' well-being—welfarism—is one possible answer to this question. (It is a partial answer because the SWF must be specified further.) This is the level at which the foregoing analysis is conducted.

At the practical level, the question is: What constitution, regime, policies, or rules should be implemented, in light of how society actually operates, to best advance the fundamental notion of the social good? One particularly interesting aspect concerns the personal character traits and other dispositions that would be best to inculcate in individuals, given what we know about human nature.

This distinction between levels of moral analysis has a long lineage. Hare (1981) traces it to Aristotle, Plato, and the classical utilitarians. Mill's (1861) famous essay, *Utilitarianism*, devotes its longest chapter (a third of the total) to this subject—which is ironic given that subsequent critics often advance the very arguments that Mill addressed therein, without acknowledging his response. Prominent twentieth-century exponents of two-level theory include the economist Harrod (1936) (writing in a philosophy journal), Rawls (1955) (preceding his more well-known work), the psychologist Baron (1994), and, most extensively, Hare (1981).[8]

[7] For detailed discussion and references to pertinent literatures, see Kaplow and Shavell (2002).

[8] The two-level distinction is related to the well-known debate between act- and rule-utilitarians. In that setting, however, the distinction between levels is often confused. Ultimately, most agree that act-utilitarianism is correct regarding how an act should be assessed against the ultimate criterion of the social good. The problem in much of the debate is that the question under consideration (level of moral thought) is ambiguous. For example,

This distinction helps to explain a wide range of human institutions. Formal organizations, notably government (but also corporations and other entities), are often best governed by rules (some embodied in constitutional "rights," including the right to "due process") that limit actors from freely maximizing the true, fundamental, social objective for fear that if permitted to do so they would not, producing less good in the end. Bentham's (1822–1823) lengthy (and less-known) constitutional writing develops this idea in detail. In similar spirit, Mill ([1861] 1998, p. 93) observed, "We should be glad to see just conduct enforced and injustice repressed, even in the minutest details, if we were not, with reason, afraid of trusting the magistrate with so unlimited an amount of power over individuals."

Also significant is the use of rules in the realm of common morality, to regulate individuals' informal interactions in everyday life. The sanctity of promises is one example; others are norms of fair division, which reduce conflict, and the principle of retribution, which deters aggression. Such notions are usually more than mere rules of thumb designed to economize on information and costs of calculation. Rather, they carry their own behavior-influencing force, both internally, through emotions such as guilt, and externally, through social sanctions such as disapprobation. Thus, individuals may refrain from breaking promises even when doing so would otherwise be to their advantage. Such norm potency helps to modulate self-interested behavior, thereby promoting cooperation and restraining opportunism. The significance of these norms, emphasized by Hume (1751), Mill (1861), Darwin (1874), and Sidgwick (1907), has

if it is asked whether, in a highly unusual circumstance, a teacher should strike a student even though there is a rule—and a good rule, for the sorts of reasons discussed in the text to follow—against corporal punishment, two different answers may be given, both correct but each responsive to a different question. If striking the student would in fact (taking into account all subsequent, indirect effects) produce more good than harm, then, at the fundamental level, it would be "good" if the teacher struck the child. But if instead the question is something like "would we like our children to be taught by teachers who showed no compunction about striking the child" or "would we like teachers, instead of being presented with a hard-and-fast rule against striking their students, to be encouraged to exercise discretion on the matter, knowing that it often will be exercised in the heat of the moment"—both directed to the practical level—then the answer would be negative. There is no real inconsistency, but in some discourse on the matter these two types of questions—corresponding to two different levels of moral inquiry—are not clearly differentiated.

been increasingly emphasized by economists, such as Becker (1996), Frank (1988), and Hirshleifer (1987), and other social scientists and natural scientists, including Alexander (1987), Baron (1993), Campbell (1975), Daly and Wilson (1988), and E.O. Wilson (1980).

The distinction between the fundamental criterion of the good and the practical question of what rules and norms are best to guide human affairs helps to reconcile the tension between welfarism and competing principles. The suggestion is that welfarism, in its pure form, is indeed the appropriate ideal for assessing the social good, whereas competing notions that sometimes conflict with welfarism (and thus also with the Pareto principle) are practical, intermediate norms that may be related to common morality or the restraint of government power. Accordingly, the latter are not truly inconsistent with welfarism as long as their proper role is understood. This view is elaborated in the subsections that follow.

b. Moral intuitions. Well-socialized individuals, whether contemplating how to act toward others or engaging in philosophical reflection on proper conduct, will understandably have intuitions regarding what is morally correct. This is particularly likely on account of the aforementioned emotional and social consequences of various norm-related behaviors.

To the extent that intuitive normative principles—including notions of equity that seem applicable to tax policy and income redistribution—have their roots in common morality, a number of implications follow. First, because of the distinction between the two levels of moral theory, it is possible, indeed likely, that practical principles (even optimal ones, given constraints of the human condition) will deviate from the fundamental criterion of the good. This is simply the truism that the optimal, second-best rule in the presence of constraints generally will not implement the first best. Hence, as suggested, the apparent conflict between welfarism and competing criteria can be reconciled.

Second, within the practical level, different sorts of principles are often appropriate in different contexts. Norms regulating conduct within families, between friends, and among coworkers may differ from one another. Even greater is the difference between any of these sets of norms and those that should regulate relationships between different agencies of government (courts versus legislatures) and between government and the governed. For example, broken promises among friends, breached contracts between firms, and transgressed rules regulating interactions between police and citizens are subject to different forms of dispute resolution and different sanctions. Even more, the underlying rules themselves should and do differ. This contrast is important for understanding the appropriate role of notions of fairness and equity for tax policy. Norms of fair division or those dictating equality in human social interaction have functions and modes of enforcement that differ greatly from those appropriate to a tax authority. To be sure, there will be underlying similarities of purpose (just as there are among rules involving understandings between friends, between firms, and between citizens and police). But the differences across contexts are sufficiently great that we should anticipate that our internal sense of equity, developed in the realm of everyday social interaction, would at best be suggestive of how government fiscal affairs should be structured.[9]

Third, the weight of common morality suggests that its precepts may sway our judgment even when their use is inappropriate. Treating others unfairly or inequitably makes us feel guilty and raises, even if subconsciously, concerns about how others will react. And the prospect that individuals would treat us improperly arouses anger. These feelings are difficult to ignore as we contemplate subjects outside of everyday interaction, such as the formulation of tax policy, because of the existence of apparent similarities. Even when reasoned analysis suggests, for example, that violations of horizontal equity—the command that equals be treated equally—are not of concern in a particular realm, such violations may still feel wrong.

At least the first two of these implications are familiar from the increasing attention to the teachings of cognitive psychology.[10] Regarding

[9] For example, norms of fair division may have evolved to assist cooperation (dividing meat from the hunt) or to avoid conflict (focal points may provide useful boundaries). And norms respecting the "rights" of current possessors (of food or land) may similarly avert hostilities. Although related concerns may impose practical, political constraints on government, these norms are addressed to a substantially different question from the fundamental inquiry into the proper criterion of the social good, as applied to questions regarding the overall distribution of income.

[10] See, for example, Baron (2000), Hogarth and Reder (1987), Kahneman, Slovic, and Tversky (1982), Nisbett and Ross (1980), and Rabin (1998).

ideal theory versus practical rules of conduct, we are aware of the many respects in which human decision-making heuristics conflict with normative decision theory. It is acknowledged that this inconsistency does not imply the existence of any defects in the latter—that is, the discrepancy does not make decisions that fail to maximize our objective normatively compelling. Rather, it is recognized that our decision-making heuristics and biases are, at best, optimal adaptive responses for the environment in which they evolved.[11]

Furthermore, the practical decision rules we actually use are better suited to some contexts than others. A heuristic may lead to approximately optimal decisions in the realm in which it was developed but to very poor outcomes in some other settings presented, say, by recent technological advances. Nevertheless, despite these important context differences, we often err by overgeneralization and may continue to follow our habits even when their deficiencies have previously been brought to our attention.

The connection between decision heuristics and biases both in general and in the particular realm of moral decision-making has been emphasized, among others, by Baron (1993, 1998).[12] In addition to identifying the connection and exploring how some of the same errors are at work, particular infirmities in the context of moral reasoning have been identified. Moral intuitions, however, may differ with regard to the third implication concerning the emotional and social weight associated with deviation from norms of common morality. As noted, this feature suggests that we may find it even more difficult than in the case of, say, decision-making under uncertainty, to engage in the proper, ideal normative analysis (welfarism), unswayed by our everyday decision rules (common morality).

[11] See, for example, Gigerenzer and Selten (2001).

[12] "What is the causal connection between something being morally right and our intuition that it is right? It is easier to understand our intuitions as arising from overgeneralizations of learned principles, or from the emotions that evolution gave us. . . ." Baron (1998, p. 152). Philosophers have long emphasized the limitations of moral intuitions as sources of illumination of the ultimate good (see, for example, Mill 1861, Sidgwick 1907, Westermarck 1932, and Hare 1981).

c. Relevance of nonwelfarist principles under welfarism. To round out an account of the apparent conflict between nonwelfarist principles and welfarism, it is useful to state clearly the relevance of the former under the latter. First, as suggested by the discussion of two-level moral theory, it is clear that, even if the fundamental criterion of the good admits only consideration of individuals' well-being, various principles that on their face are nonwelfarist are likely to be appropriate at the practical level, for the regulation of human interaction and of government. For example, the norm of equal treatment of equals that underlies horizontal equity and notions of due process that are implicitly invoked in some critiques of utilitarianism (see subsection 14.A.1.b) are well understood as useful constraints on the behavior of public officials and institutions. Even in this context, however, such principles must be applied carefully. What process should be due obviously depends on the context (determining who is next in line at the post office compared to who should be subject to severe punishment). Insistence on equal treatment needs to be moderated (perfect equality is often impossible and near perfection may be excessively costly), and context is relevant (random enforcement, such as with tax audits, violates equal treatment, but if selection is truly random, the potential for abuse may be limited). In all cases, determination of optimal policies requires reference to the proper objective function, which, as suggested here, is purely welfarist.

Second, nonwelfarist principles may serve as proxy instruments for policy analysis. When various equity norms are violated, there may well be an underlying problem, something indicating that welfare is not being maximized. This is apparent with notions of merit and desert because rewards and punishments in accord with such concepts tend to encourage productive effort and deter misbehavior. Further examples involving horizontal equity will be discussed in section 15.B. However, understood as proxy instruments, nonwelfarist principles have their limits. They may motivate inquiry, but they are not part of the ultimate social objective that, in the end, is invoked to determine optimal policy. In discussing critiques of utilitarianism, Mirrlees (1982) offers the observation, so familiar in ordinary economic analysis, that counterintuitive findings do call for explanation but do not per se warrant rejection. After all, if our initial intuitions are never wrong, what is the point of analysis? This message is particularly apt regarding optimal taxation, where the analysis is complex and subtle and important results, such as those pertaining to the shape of the optimal income tax schedule, are often unexpected. Furthermore, as observed in subsection b, we know that our modes of decision-making are often mistaken, involving oversimplified heuristics that, in addition, may be erroneously extended to contexts in which their performance is poor. Hence, divergences between the results of careful tax policy analysis and our intuitions concerning ideal taxation should be reasonably frequent; if they are not, our analysis has probably been too casual. An important lesson is that, when considering familiar notions of tax equity—as with all intuitions, moral or otherwise—it is valuable to determine their underlying basis. Once their (usually intermediate, instrumental) relationship to welfare is identified, they can guide policy analysis more effectively, and it will be more apparent when their usefulness has been exhausted.

Third, because of our attachment to nonwelfarist principles—as explained, they are not mere rules of thumb but maxims that carry emotional freight—they may have weight under welfarist analysis because they are a component of welfare itself.[13] That is, we may have tastes regarding, say, how vicious criminals are punished or how income is distributed. Indeed, preferences regarding income redistribution were the subject of section 5.D, extending the analysis of optimal income taxation. Whether and to what extent nonwelfarist notions should be credited on this ground is an empirical question, quite distinct from whether the notions have any standing as fundamental criteria of the social good. (A separate question is whether, even if such tastes exist, they should be counted, a subject examined further in subsection B.3.)

[13] Understanding individuals' moral sense as akin to tastes dates at least to Hutcheson (1725–1755) and Hume (1751). Furthermore, Mill ([1861] 1998, p. 83) observed that individuals can develop tastes even for rules that begin as merely means to an end: "What was once desired as an instrument for the attainment of happiness, has come to be desired for its own sake. In being desired for its own sake it is, however, desired as *part* of happiness."

B. Well-Being

1. Definition

It is familiar to economists that well-being or utility (the terms are used interchangeably throughout) is a broad, subjective notion, not one limited to material pleasures, hedonistic enjoyment, or any other a priori class of pleasures and pains. Resources, often measured in monetary units, are means to obtain goods and services; these, in turn, are means to generating utility, which may be derived directly from goods or indirectly and intangibly, such as through fulfillment, sympathetic feelings for family and friends, aesthetic enjoyment of art or the environment, and so forth. What counts, and how much it counts, for each individual in society depends on that individual's mind, not on any analyst's view of what should constitute well-being.[14]

It is common for outsiders to welfare economics (particularly critics thereof) to suppose that utility is a narrower concept—paralleling attacks on utilitarianism—but this view reflects a misinterpretation.[15] Although minimal exposure to economics may leave the impression that only money or goods are thought to be important, their intermediate role is well understood within the profession, as indicated, for example, by Lancaster's (1966) framework in which goods are inputs that yield characteristics that in turn produce utility, Michael and Becker's (1973) household production function approach, and the wide range of objects

[14] Modern philosophical treatments, which vary in how closely they adhere to subjectivist accounts similar to those used in welfare economics, include Griffin (1986), Nussbaum and Sen (1993), and Sumner (1996). Among others, Mill (1861) is sometimes associated with the view that higher (intellectual) pleasures are superior to more basic (sensual) pleasures, although there is some dispute about the extent to which he believed this in principle rather than as an empirical proposition about human nature.

[15] Regarding criticism of classical utilitarians, Bentham ([1781] 1988, p. 33) explicitly includes among the "simple pleasures" the pleasures of skill, amity, a good name, piety, benevolence, imagination, and association. And Mill (1861, pp. 55–56) not only took a broad view himself but also lamented that followers of the Greek philosopher Epicurus were unfairly characterized as holding a narrow orientation toward pleasure.

of human concern that are addressed by economists. In any event, what should be deemed to constitute well-being for normative purposes is not a matter of stipulation or interpretation of economists' practices. Instead, well-being should be construed in a manner that connects to the ultimate motivation for welfarism, which deems individuals' well-being and only their well-being to be relevant to social welfare. As will be seen, this point illuminates the questions examined in the subsections that follow.

2. Limited Information and Other Decision-Making Infirmities

It has long been understood that individuals' decisions may not always be in their own best interests, a point that is receiving increased attention in work at the intersection of economics and psychology.[16] Individuals may lack pertinent information or suffer from cognitive biases, myopia, addiction, and so forth.[17] In such situations, what may be termed decisional utility—corresponding to expressed preferences as indicated by actual behavior—differs from experienced utility—the actual subjective states arising as a consequence of decisions.

The standard view, accepted here, is that at least in principle it is experienced utility that corresponds to subjective well-being, which is taken to be of normative significance. Thus, Sidgwick (1907), Harsanyi (1955), and Mirrlees (1982), among others, argue that social welfare should be assessed by reference to individuals' rational, fully informed preferences when they conflict with revealed preferences. What is desirable for individuals should be understood "supposing the desirer to possess a perfect forecast, emotional as well as intellectual, of the state of attainment or fruition," in the words of Sidgwick ([1907] 1981, p. 111).

[16] See, for example, the literature cited in note 10, and recent work on neuroeconomics, surveyed by Camerer, Loewenstein, and Prelec (2005).

[17] A particularly interesting case involves imperfect information about one's own future preferences and the possibility of changing one's preferences (see, for example, Cyert and De Groot 1975 and Weizsäcker 1971). Stigler and Becker (1977) suggest that one can analyze all such cases as ones involving stable tastes but possibly changed circumstances or limited information.

The basis for this view can be seen most clearly in cases of misinformation. Suppose that an individual desires to step forward in order to attain a better view from a mountaintop (and will do so unless restrained) but is unaware that the footing is insecure, so that the result would be a catastrophic fall. Utility is deemed to be lower, not higher, if the individual's stated preference is satisfied, for it is the actual consequence that is taken to matter.

The policy implications of this view are somewhat limited for familiar reasons emphasized by Bentham (1781) and Mill (1859), namely, that the government often lacks the information and may not have the right incentives to improve on individuals' imperfect situations. In certain areas, such as safety regulation, gains may be possible. Regarding tax and expenditure policy, in-kind provision of goods and services (public housing, free and compulsory public education) is sometimes offered in lieu of cash for these reasons, although, as explained in section 7.E, it may be that externalities and other explanations are the more weighty considerations. Another important application involves social insurance systems, often justified on grounds of individuals' myopia (see subsection 11.B.1).

Finally, mistakes may have a distributive incidence that affects the optimal extent of redistribution. In particular, deficient decision-making may be concentrated among the poor, in significant part because myopia and other infirmities may partially explain their poverty. One could model such infirmities analogously to subsection 12.A.2's model of economies of scale within the family (though the effect is opposite in direction). In the notation of that model, we can simply examine the case in which the lower is income, the lower is β, where $\beta = 1$ at the highest level of income. Then we can write individuals' utility functions (focusing, as there, only on consumption) as $u(\beta c)$. The marginal utility of consumption is $\beta u'(\beta c)$. The leading coefficient indicates that low-β individuals make less effective use of each dollar and hence have a lower marginal utility on this account. However, β also multiplies c as the argument of u', implying that low-β individuals' marginal utility is higher because they are at a lower level of effective consumption. Which effect is larger depends, as in subsection 12.A.2, on the curvature of u. If individuals' relative risk aversion is greater (less) than one, overall marginal utility of low-β individuals would be higher (lower) because the latter effect would be more (less) significant than the former. Additionally, there is the straightforward point that utility levels of low-β individuals will be lower, so if the SWF is strictly concave, greater redistribution would be optimal on this account.

3. Other-Regarding Preferences

Some economists and philosophers argue that certain sorts of preferences should not be credited in assessing social welfare. In analyzing redistribution, the most relevant preferences include both positive preferences toward others, such as altruism, and negative preferences, such as envy.[18] Before discussing these specific possible sources of utility, it is useful to consider the broader category of other-regarding (or external) preferences, which Dworkin (1981a), Harsanyi (1977, 1988), and Nozick (1974), among others, suggest should be ignored altogether.[19]

First, preferences should be distinguished from beliefs or viewpoints. For example, one who regards utilitarianism as the proper normative stance does not necessarily—in fact, is quite unlikely to—have personal preferences that weight all individuals in society equally, whether oneself, immediate family members, neighbors, or distant strangers. The arguments of a welfarist SWF are individuals' utilities—their subjective well-being, not their normative theories—so beliefs or viewpoints should indeed be disregarded when assessing the level of social welfare under a given regime.

Second, if actual preferences are involved, it is prima facie problematic to omit particular components of utility. The normative basis for welfarism depends on individuals' actual well-being and does not discriminate among its sources. It is not clear why any preferences should not count or how a list of objectionable preferences might be determined. Furthermore, because ignoring elements of well-being

[18] See section 5.D (preferences regarding redistribution), chapter 10 (taxation of transfers; especially section 10.B on transfer motives and subsection 10.C.3 on charitable giving), and chapter 12 (taxation of families).

[19] The topics in this subsection and pertinent literature are considered more fully in Kaplow and Shavell (2002).

constitutes a formal violation of welfarism, the demonstration in subsection A.2 indicates that all individuals may be made worse off as a result (which is demonstrated later for altruism and envy).

Third, it is unclear what it means to censor a preference. The implication is that the pertinent individuals would be treated as if they had a different utility function, so in principle it would be necessary to specify that alternative utility function. Yet an infinity of functions (not subject to censorship conditions) are possible. None of them is the actual utility function of the individuals in question, so it is hard to see the basis for making a selection. Perhaps more important, it is unclear what would be the normative force of maximizing such a utility function for a person who does not experience utility in accord with it.

Fourth, the suggestion that other-regarding preferences are a problematic class is, upon reflection, alarming. After all, much of what individuals value regards others, including the love of family members, companionship of friends, fulfillment from participation in teams or social groups, and appreciation of performing artists.[20] Perhaps one could distinguish between inward and outward concern for others, so that it would be acceptable to enjoy a friend's jokes or the play of one's children but not to get any utility from their well-being, but it is not clear that our feelings always draw such distinctions or that those distinctions should matter.

Turn now to the pertinent specific positive and negative other-regarding preferences that have been subject to challenge. Although crediting positive preferences, notably, altruism, in assessing matters pertaining to taxation and redistribution may seem unexceptional, Harsanyi (1988) and others object on the ground that doing so would violate equality because social welfare assessments would thereby give more weight to some individuals (subjects of others' altruism) than to others or that it would involve some sort of double counting.[21] The foregoing comments,

[20] Note that the case of performing artists includes instances of what may be viewed as negative other-regarding preferences since onlookers' pleasure may derive from performers' acts that involve their suffering pain.

[21] This objection was addressed briefly in note 10 in chapter 10 on taxation of transfers and was mentioned in chapter 12 on taxation of families because of the relevance of altruism (and other positive preference interdependencies) to those analyses.

of course, are fully applicable to the present case in which individuals derive positive utility from others' well-being. Note further that if such preferences receive no weight, then in principle one should prohibit many acts of caring and kindness within the family and toward others, a prohibition that would typically make both the altruist and the beneficiaries worse off. Suppose, for example, that all individuals were in pairs, one gaining altruistic utility from the other, that all started with equal resources, and that the altruists in each pair give gifts or otherwise bestow benefits on the others. From an initial point of equal resources, such transfers would reduce total utility if the altruistic utility is ignored (for the transfers are disequalizing and let us assume that otherwise the individuals have identical utility functions exhibiting diminishing marginal utility). Yet restricting gifts would reduce everyone's actual utility.

The particular objections to admitting altruism also are weak on their own terms. When altruism is credited, no one's utility or marginal utility in its own right counts more than anyone else's. To be sure, subjects of altruism may be lucky because they are likely to benefit from others' concern. But so are individuals whose talents happen to be highly in demand on account of others' preferences. (Performing artists' incomes vary wildly depending on what is in fashion.) To be sure, individuals who command more resources for whatever reason will thereby have a lower marginal utility of income (and a higher level of utility, relevant for a strictly concave SWF), which will be taken into account when maximizing the SWF. See, for example, subsection 10.A.3 on the effect of private voluntary transfers on the marginal social value of redistribution. The double-counting objection also is untenable. It suggests, in the case of a gift, that the donor's or the donee's utility should be ignored. But the donee does consume the resources and thus has a higher utility, and as an individual distinct from the donor, the donee should be counted in his or her own right. Likewise, the donor in fact receives distinctive utility—by revealed preference, more than if the resources had been used on own-consumption; to ignore it would be inconsistent with a concern for the donor's actual subjective well-being.

Basing social welfare judgments regarding distributive issues on negative other-regarding preferences, such as envy, seems more problematic than crediting altruism. First, satisfying such preferences usually reduces utility overall, for others' utility losses are ordinarily larger than any gain to the envious. Second, society may well be better off if individuals did not have such preferences (whereas altruistic preferences generally permit a higher level of social welfare to be achieved with given resources).[22] This latter argument, however, needs to be articulated more fully. To satisfy an individual's negative preference requires making others worse off. If the preference could be extinguished, the individual could be made as well off as before without making the other individuals worse off, making possible gains in well-being, ceteris paribus. Furthermore, it might be supposed that a social policy of discrediting preferences would help to alter them over time. For example, Becker (1996) and Frank (1985) are among the many who suggest that antidiscrimination laws may have influenced people's discriminatory attitudes. Note, however, that this sort of justification for ignoring certain negative preferences is instrumental: The preferences are ignored not because the individuals' actual well-being does not matter, but instead because the social act of ignoring them will lead to a superior state of affairs. (Note that this analysis exemplifies the sort of two-level reasoning and critical examination of moral intuitions that is discussed in subsection A.3.[23])

Perhaps these functional explanations underlie our reluctance to credit negative preferences, whether involving envy, racism, or sadism. We might believe that satisfying such preferences is rarely, if ever, desirable (even assuming that the preferences are permanent) and doubt that we could identify any exceptions, and we may also believe that closing

[22] Altruistic preferences are not entirely unproblematic because of difficulties like the Samaritan's dilemma (see, for example, Bernheim and Stark 1988). In addition, not all negative preferences lead to lower social welfare. For example, individuals' desire to see wrongdoers punished leads them to cooperate with authorities and to engage in social ostracism, the prospect of which deters misbehavior (although the retributive urge can also be excessive, leading to undesirable consequences).

[23] This explicit, welfarist account of why it may not be best for society to base policy on certain negative preferences offers grounds for determining which preferences are objectionable and what prescription follows from the judgment, unlike the typical objections to other-regarding preferences found in the pertinent literatures.

our eyes to such preferences or even denouncing them will contribute to the problems' dissolution. If instead negative preferences were strong, widely held, permanent, and could be satisfied at little direct cost, it is hardly clear that they should be ignored. Suppose that, beyond some point, additional income directly increased each individual's utility only slightly and actually contributed much more to everyone else's suffering on account of negative preferences. Then a policy that trimmed incomes would make everyone better off. Such a hypothetical example may not seem powerful because it is so fanciful—our imaginations cannot really find credible the premise that all actually would be better off if everyone's standard of living were reduced, so the conclusion seems less compelling. If, however, we transform the example, it seems plainly correct: Suppose that the negative preferences derive not from others simply having the added income, but from the fact that all use it to hold noisier parties that bother their neighbors. This "tangible" externality would be counted, yet such externalities are like envy in that they influence utility through complex neural activity that we are only beginning to understand.[24]

In sum, it seems difficult to articulate the actual meaning of preference censoring or to identify a convincing rationale for ignoring, as a matter of first principles, certain sorts of preferences. Nevertheless, it may be good social policy to set aside certain negative preferences, including envy. In any event, it is hardly clear that such preferences exist in sufficient strength to have a meaningful impact on the optimal income taxation problem. For example, in Boskin and Sheshinski's (1978, p. 599) analysis of how individuals' concern for relative status affects optimal redistribution, they cautioned the reader that empirical evidence for an extremely strong concern, which is often taken for granted, is "virtually nonexistent, let alone convincing."

[24] To take another, lighter example, suppose that a group of individuals greatly enjoy practical jokes, that the butt of any particular joke suffers only a little, and that each is at the suffering end equally often. Satisfying these negative preferences makes everyone better off, and it is hard to see why this source of utility should be ignored—that is, in the absence of additional effects such as an influence on character that would adversely affect other behaviors (in which case the preferences would not be irrelevant but instead would be outweighed).

4. *Capabilities, Primary Goods, and Well-Being*[25]

In considering the subject of redistribution, some economists and philosophers would not assess individuals' situations by reference to well-being. Instead, Sen (1985a, 1985b, 1997) examines capabilities and functionings, such as nourishment, shelter, physical mobility, and the ability to take part in the life of the community, and Rawls (1971, 1982) considers primary goods, including rights and liberties, opportunities and powers, and income and wealth.[26] According primacy to such means of fulfillment rather than to fulfillment itself, however, seems difficult to justify and (unsurprisingly, in light of the argument in subsection A.2) may well lead to everyone being worse off.

In many respects, these constructs are problematic on their face. Most primary goods and capabilities are patently instrumental rather than intrinsically valuable to individuals, just as the commodities in Michael and Becker's (1973) theory of household production and the characteristics in Lancaster's (1966) consumer theory are means of generating utility. Perhaps some primary goods and capabilities are ends, but in that case they constitute elements of well-being rather than its totality.

Additionally, many questions regarding these theories remain unanswered: Whether viewed as means, ends, or some combination thereof, how is the list of primary goods or capabilities to be determined? How can the same list be employed for heterogeneous individuals who face differing circumstances? Additionally, since multiple means are admitted, how are they to be aggregated to produce an overall assessment of individuals' situations?[27] Welfare economics answers such questions by reference to individuals' utility functions. Any resulting list, of course, may vary across individuals and states. If well-being is not the guide and if uniformity is imposed, then some substantial justificatory apparatus is required, but none has been offered.[28]

[25] This section draws on Kaplow (2007f).

[26] See also Dworkin (1981b), who focuses on equality of resources, which he contrasts with equality of welfare.

[27] Critics of Rawls's notion have suggested that this index problem is insurmountable (see Blair 1988, Gibbard 1979, and Plott 1978).

[28] Proponents of alternatives to well-being rely on the reader's intuition and seem to suggest that their theories, once revealed, are self-evident. They also sometimes make reference

Assuming these obstacles could be overcome, the result—whatever it may be—will conflict with the Pareto principle.[29] If all individuals were identical, all would be worse off if the social allocation of goods was that dictated by the alternative theory because (except by coincidence) individuals would receive more of some goods and less of others than the levels that would maximize their utilities. If individuals' utility functions varied, then the theory's allocation would be welfare-reducing for all (except possibly for a subset, by coincidence—although even then, if the theory's allocation were abandoned, a redistribution of benefits from the others would make possible a Pareto improvement). Moreover, in this case, individuals would wish to trade with each other in order to align their allocation of goods more closely with their actual preferences. To prevent the theory's dictates from being undermined, such Pareto-improving trade would have to be prohibited. As suggested in subsection A.2, such anti-Paretian results are troublesome for moral theories that purport to ground themselves in such notions as consent, freedom, and autonomy, which is the case for both Sen and Rawls.[30]

How, then, can one explain the attraction of such alternative theories? As with the other issues explored in this chapter, on reflection it appears that what are being treated as ends in themselves are better understood instrumentally. Well-being—especially in the presence of preference heterogeneity—is difficult to measure objectively and uncontroversially, so as a practical matter it may make sense to focus on specific, tangible indicators. Furthermore, one may be concerned about an excessive tendency

to a concern about individuals having expensive tastes (see Rawls 1982, pp. 168–169, and Sen 1997, pp. 197–198). As noted in Kaplow (2006a), however, they ignore that individuals have a disincentive (rather than the implicitly assumed affirmative incentive) to develop tastes that by definition reduce their well-being, ceteris paribus.

[29] For more formal statements, see Kaplow (2007f). Relatedly, Gibbard (1979) shows that Rawls's approach is incompatible with the Pareto principle when individuals are heterogeneous, on account of the price index problem (which arises even when the only primary good is income because prices of different underlying goods can vary across regimes).

[30] Sen (1985b) uses the terms "agency" and "freedom" in his title, and the concepts are central to his argument there and elsewhere in advancing his notions of capabilities and functionings. Rawls (1971), as is well known, imagines that individuals in an original position would unanimously consent to his framework (of which primary goods is a component), and his primacy of liberty makes clear that freedom is a central concept for him as well.

to focus on money, thereby omitting other important constituents of well-being. Interestingly, Sen's (1985a) book developing his capabilities-based approach features two appendixes, one showing how capability-based measures differ in practice from using per capita GNP and another showing how differences in the treatment of men and women are obscured unless one pays attention to capabilities.[31] Both discussions suggest that his work may be motivated by the need to provide a more nearly complete depiction of well-being, especially in considering developing countries, although the more conceptual body of the book and his other writings on the subject have more of an anti-welfarist flavor.

Primary goods or capabilities may best be understood as ways of gauging well-being rather than as substitute concepts, but in that event the answers to the questions posed at the outset of this subsection are empirical and contingent on the individuals involved and their circumstances. As a practical matter, taxation and income redistribution are largely undertaken in terms of money, and individuals are not generally distinguished on account of possible differences in their utility functions; to the extent that this approach is followed, Sen's and Rawls's approaches may not have implications that diverge sharply from welfarism's focus on well-being. However, as subsection 2 notes, some transfers are in-kind, and the government directly provides goods and services, often including public education. Furthermore, utility differences are pertinent to the taxation of private transfers and the taxation of families, discussed in chapters 10 and 12, respectively, and in the case of physical disabilities. In these realms, it may matter whether society cares about individuals' well-being or something else.[32]

[31] Rawls (1971, p. 95; 1982, p. 159) likewise offers practical arguments in support of his notion of primary goods.

[32] Regarding physical disabilities, for example, a capability deficit may, ceteris paribus, indicate a well-being deficit. It would not make sense to remedy the capability deficit directly, such as by providing expensive transportation services and other costly accommodations to individuals, if they could be made better off (and at lower cost) by being given money that could be spent otherwise. For example, some disabled individuals might prefer cheaper home entertainment to a toilsome journey to the opera, which would not be enjoyed on account of the exhaustion involved in getting there. Not all disabled individuals will have the goal of emulating the movements and other activities of those who are not disabled; treating them as if they do or should, regardless of their actual wishes, is avoided (at least in principle) when the social objective is denominated in terms of well-being.

14

Social Welfare Function

Chapter 13 presents the rationale for the SWF to be a function of individuals' utilities (and of nothing else). This chapter considers two remaining questions. First, how should individuals' utilities be aggregated? In subsection 3.B.1, it was noted that optimal income tax analysis often considers SWFs of an isoelastic form (see expression 3.2), where the parameter e indicates society's degree of aversion to inequality in utility levels: $e = 0$ corresponds to the utilitarian SWF and higher levels of e to strictly concave SWFs. (The limiting case, as e approaches infinity, yields the maximin formulation suggested by Rawls (1971), placing all weight on the least-well-off individual.) Even if this functional form were accepted, it would remain necessary to select a value of e.

Second, whose utilities are to be included as arguments in the SWF? A standard response is to count all citizens in the nation while excluding all others. Some of the fundamental normative issues regarding this view will be examined, although briefly and tentatively.

A. Aggregation

This section begins by examining two frameworks that have been suggested for choosing an SWF. Then some concerns are addressed.

1. Frameworks

a. Original position. Although most often associated with Rawls (1971), the notion that distributive judgments should be derived from individuals' choices in some sort of "original position" behind a "veil of ignorance" had previously been advanced by Vickrey (1945) and Harsanyi (1953), among others.[1] Specifically, they suggested that individuals consider what regime they would favor if they had an equal chance of being any member of society, in which case a utilitarian SWF would be employed.

Taking the case of a finite population of n individuals, each individual is postulated to have a $1/n$ probability of taking the role and thus experiencing the utility $u_i(x)$ of each individual i, where, as in chapters 3 and 13, x denotes the regime or social state. Hence, an individual would choose x to maximize expected utility, given by

$$\sum \frac{1}{n} u_i(x), \tag{14.1}$$

where summations are over i, from 1 to n. Obviously, maximizing an individual's expected utility, expression (14.1), is equivalent to maximizing

$$\frac{1}{n} \sum u_i(x), \tag{14.2}$$

that is, maximizing average utility over the population. Furthermore, for a given population (see subsection B.3), maximizing expression (14.2) is equivalent to maximizing

$$\sum u_i(x), \tag{14.3}$$

that is, maximizing total utility over the population. Therefore, as explored in section 3.B, the SWF would exhibit a (possibly significant) preference for equalizing resources, but with this utilitarian SWF, that preference would depend entirely on the concavity of individuals' utility

[1] See also Fleming (1952) and Strotz (1958). The motivation for this framework is that it embodies impartiality. In these respects, it is similar to the Golden Rule (commanding that individuals treat others as they would wish others to treat themselves), Kant's (1785) categorical imperative (insisting that individuals choose principles for themselves as if the principles would be generalized to the entire population), and Lewis's (1946) suggestion that individuals imagine that they would rotate through all positions in society.

functions—corresponding to the rate at which individuals' marginal utility declines, which is equivalent to their degree of risk aversion.

Vickrey and Harsanyi thereby presented a basic normative framework in which the problem of choosing an SWF was formally equivalent to the problem of individual choice under uncertainty. Rawls (1971) embraced this general approach but resisted the utilitarian implication. He insisted, in essence, that individuals in the original position would maximize the minimum value of utility, with the now familiar implication that society should reduce the entire population to misery if this would improve the lot of the least-well-off individual infinitesimally.

The reasons for Rawls's (1971) conclusion remain obscure and contested. He acknowledged Harsanyi's work and endorsed the view that individuals should be rational, as that concept is understood in decision theory. Nevertheless, he insisted that individuals' "risk aversion" in the original position and other considerations would lead them to choose a maximin SWF.[2] A useful clarification is offered in Arrow's (1973, p. 256) book review of Rawls (1971):[3]

> When I first wrote on this matter [Arrow 1951], I . . . denied the welfare relevance of the expected-utility theory. But the Vickrey-Harsanyi argument puts matters in a different perspective; if an individual assumes he may with equal probability be any member of society, then indeed he evaluates any policy by his expected utility, *where the utility function is specifically that defined by the von Neumann–Morgenstern theorem.* Rawls therefore errs when he argues that average utilitarianism assumes risk neutrality . . . ; on the contrary, the degree of risk aversion of the individuals is already incorporated in the utility function.[4]

[2] Rawls's approach differs in other respects not obviously pertinent to the concavity of the SWF, including his use of primary goods rather than utility (on which see subsection 13.B.4), his reference to the least-well-off group rather than individual, and his emphasis especially in later writings on political (as distinct from moral) theory.

[3] See also Hare's (1973) review.

[4] Sen (1979) suggests that Harsanyi does not establish that the SWF must be linear because, in essence, one could apply a nonlinear SWF to transformed von Neumann–Morgenstern (VNM) utility functions (see also Weymark 1991). For example, if one squares individuals' VNM utilities, the corresponding SWF would first take the square root of each

b. Social rationality. Another framework that can be used to narrow the range of admissible SWFs is one that imposes rationality requirements. This approach also was pioneered by Harsanyi (1955, 1977). He makes three sets of assumptions. First, individuals' preferences are assumed to conform to the standard rationality postulates of decision theory—namely, completeness, transitivity, continuity, and monotonicity—which implies that they can be represented by a von Neumann–Morgenstern utility function. Second, the SWF is assumed to respect the same rationality postulates. To this, Harsanyi adds a third set of explicitly value-laden assumptions, tantamount to welfarism, positive responsiveness, and equality (equivalently, symmetry or anonymity).

The first two sets of assumptions plus welfarism imply that the SWF is linear in individuals' von Neumann–Morgenstern utilities. The analysis parallels that for individual decision-making under uncertainty. The main difference is that the probabilities of individual decision analysis are replaced by weights, which in the abstract could be arbitrary. The linear SWF is further specified by positive responsiveness (that is, social welfare is increasing rather than decreasing in individuals' well-being) and equal-weighting, which brings us to the utilitarian SWF. It should be emphasized that the linearity result is produced not by the final set of value judgments but rather by the rationality postulates. Thus, to favor a standard, strictly concave SWF ($e > 0$), it is necessary to reject at least one of the rationality requirements.

In a brief, provocative comment, Diamond (1967, p. 766) offers an example that he suggests illustrates a shortcoming of Harsanyi (1955), namely, that in addition to final states "society is also interested in the process of choice." As explained in subsection 13.A.3, however, two-level moral theory allows welfarism (and, accordingly, utilitarianism) to

individual's utility function before summing them. This point, however, carries no implication: The result of transforming both the VNM utilities and the SWF in a precisely offsetting fashion is a nullity. (Note also that squared VNM utilities are not themselves proper VNM utilities for the individuals in question because the VNM utility function may only be subject to linear transformations.) Furthermore, regarding both the present derivation of utilitarianism and that presented in the next subsection, there is no doubt that Harsanyi and others refer to individuals' VNM utility functions, so the meaning of the statement that the SWF must be linear in individuals' utilities is unambiguous (see, for example, d'Aspremont and Gérard-Varet 1991).

accommodate this apparently competing consideration.[5] Indeed, as noted there, for both Bentham and Mill a concern for government process was a chief instance in which optimal pragmatic rules and institutions would likely deviate from a full, direct realization of normative first principles. To accept that the SWF should be utilitarian does not imply, for example, that one must give a tax collector wide discretion in the choice of audit targets. Furthermore, concern for abuse should motivate attention to governmental processes independently of whether the SWF that should guide design of the income tax schedule is linear or strictly concave.

Two further perspectives reinforce the suggestion that a rational SWF must be linear. The first draws on the Pareto principle, which might have been thought inapplicable to distributive questions. Suppose, following Kaplow (1995a), that under a proposed reform each individual's expected utility is higher, but the resulting distribution of utilities has a greater variance than in the status quo. The latter feature would reduce social welfare under a strictly concave SWF. Moreover, if the increase in each individual's expected utility is sufficiently small (holding the increase in variance constant), the inequality effect would dominate and the reform would be deemed to lower social welfare. Yet, by assumption, each individual's expected utility is higher under the reform. Thus, any strictly concave SWF sometimes violates the Pareto principle.

Relatedly, Hammond (1983), Myerson (1981), and Ng (1981) have noted that only a linear SWF satisfies a time consistency property relating ex ante and ex post welfare assessments. In the foregoing example, a strictly concave SWF offers a different assessment if, instead of assessing the ex post outcome, one inserts as arguments individuals' ex ante expected utilities. To illustrate the problem, elaborate the prior example by assuming specifically that nine individuals will gain and one will lose under the reform, and each individual has an equal chance of being the loser. An ex ante assessment favors implementation (because by hypothesis each individual's expected utility is higher under the reform). Once implemented, there are (with certainty) nine winners and one loser, and the strictly concave SWF judges the ex post situation worse than the preexisting

[5] For further discussion of Diamond's example, see Broome (1984), Harsanyi (1975), Mirrlees (1982), and Ng (1981).

status quo. Thus, if feasible, the reform should be repealed immediately. Once repealed, however, the SWF would favor re-implementation, that is, if ex ante (expected) utilities were the arguments. And so on.[6]

2. *Concerns*

a. Interpersonal comparisons of utility. The use of an individualistic SWF in optimal income tax analysis (or otherwise) implicitly involves interpersonal comparisons of utility. For concreteness, consider a utilitarian SWF. In choosing among the admissible representations of each individual's von Neumann–Morgenstern utility function (which is only unique up to a linear transformation), one is deciding how units of one individual's utility function compare to units of others' utility functions.[7]

During the mid-twentieth century and to an extent thereafter, interpersonal utility comparisons were eschewed in welfare economics, following the argument of Robbins.[8] It appears, however, that he was misinterpreted from the outset. As Robbins (1935, pp. vii–x) clarified

[6] To avoid such circularity, the SWF would have to use only ex post distributions of utilities—which, as noted previously, creates a conflict with the Pareto principle.

[7] This requires what is referred to as cardinal unit comparability (see, for example, Sen 1977). For a strictly concave SWF, one needs cardinal full comparability, which includes comparisons of utility levels as well. The extreme case of the maximin SWF requires only level comparability. It had been believed that this requirement is different from and, in some respects, less demanding than, cardinal unit comparability. However, cardinality derives from the von Neumann–Morgenstern axioms, which are independent of comparability. Furthermore, given cardinality, level comparability implies unit comparability, but not vice versa (see Ng 1984a). The intuition is that one can take any unit of one individual's utility—say the unit from 1 util to 2 utils—and, using level comparability, find the corresponding util measures on any other individual's utility function, at which point one knows how many utils of any other individual's utility correspond to one unit of the first individual's utility.

[8] This subject is related to the apparent paradox involving the use of the standard class of individualistic SWFs in spite of Arrow's (1951) famous impossibility theorem that seems to rule out all SWFs. The resolution is that individualistic SWFs are possible when one relaxes one of Arrow's assumptions, specifically, to allow the domain of social choice procedures to consist of individuals' utilities rather than just their orderings. (Note that orderings do not permit unit or level comparability; see note 7.)

in the second edition of *An Essay on the Nature and Significance of Economic Science* and in a subsequent essay (Robbins 1938), his argument was not that interpersonal comparisons should not be made—indeed, they were inevitable—but rather that they involve value judgments rather than scientifically verifiable statements.

Much modern welfare economics has pursued analysis of SWFs that depend on individuals' utilities and not just orderings because preference intensities matter and interpersonal comparisons of utility are required if distributive judgments are to be made.[9] A number of observations have been offered regarding the feasibility of interpersonal comparisons (even if such comparisons are contestable). First is the fact that comparisons are regularly made, whether in everyday interaction when deciding how to treat children or allocate burdens between spouses, in emergency rooms when conducting triage, or in the policy realm when setting priorities for assistance. We also make analogous intrapersonal comparisons, such as when deciding whether to change occupations, get married, or have children—comparisons that are based in part on our observations of others' analogous experiences. Interpersonal judgments may be grounded on the underlying similarity of members of the human species but also may take account of observable differences in individuals' constitutions, circumstances, and expressions. A further argument advanced by Binmore (1998) is that we have evolved to have some capacity to make interpersonal comparisons, for this ability is useful in social interactions. Additional support is provided by modern scientific research on humans' ability to register others' psychological states.[10] One might

[9] In very basic settings, little information is required. Lerner (1944) showed that allowing individuals' utility functions to differ in unobservable ways does not upset the conclusion that a utilitarian SWF is maximized by an equal allocation of resources (in a setting with lump-sum transfers and no labor supply). See also Sen (1973b), extending the result to all concave SWFs. However, when incentive tradeoffs are necessary, as in the optimal income taxation literature, further information about the distribution of utility functions is required, at which point the analysis may proceed allowing for such heterogeneity (see subsection 5.C.2).

[10] See, for example, Wicker et al. (2003), showing that one individual's facial expression of disgust activates a neural representation of the same experience in observers' brains, and Preston and de Waal (2002), exploring how empathy operates in the nervous system and discussing its adaptive features.

further speculate that, independently of any such internal mechanisms, advances in brain science may ultimately provide a scientific basis for measuring individuals' utility in an interpersonally comparable fashion.[11] Even well short of such knowledge, however, the need for and existence of at least some basis for interpersonal utility comparisons have led to a broader acceptance thereof.[12]

b. Weight on equality. Some analysts wonder whether the SWF derived in one of subsection 1's frameworks or in some other fashion gives appropriate weight to equality.[13] It is unclear, however, what implication, if any, follows from any deviation between our intuitions about the weight of equality or degree of redistribution that seems correct and the implications of whatever SWF emerges from analysis of compelling moral constructions.

First, as the analysis in subsection 13.A.3.b warns, our moral intuitions are likely to be unreliable, and counterintuitive results should be anticipated. This seems all the more likely given the nature of the redistribution problem. As chapters 4 and 5, among others, make clear, the optimal extent of redistribution—even for a given SWF—depends on a complex array of subtle factors, including various traits of utility functions, the distribution of skills, available tax instruments, general equilibrium effects, considerations of administration and enforcement, and so forth. The shape of the optimal income tax schedule is counterintuitive, and the optimal degree of redistribution is highly contingent,

[11] It seems no coincidence that the ordinalist revolution in economics roughly coincided with behavioralism in psychology, at a time when our limited knowledge rendered activity within the human brain almost entirely beyond our comprehension.

[12] See, for example, Brandt (1979), Hare (1981), Harsanyi (1975, 1982), Little (1957), Mirrlees (1982), Scitovsky (1951), and Sen (1973a).

[13] A prominent example is Sen's (1973a, p. 16) objection that utilitarianism should not be regarded as sufficiently egalitarian (or egalitarian at all) because it is unconcerned with the distribution of utility levels per se. Instead, Sen advances what he refers to as a "weak equity axiom," which requires that an individual (say, *A*) who receives lower utility for any given income level than another must receive a larger allocation of income. However, this axiom may be more extreme than is maximin: Consider the case in which the incentive effects of implementing such a scheme so reduce productivity as to make even individual *A* far worse off than under less radical schemes.

the concavity of the SWF being just one factor. Furthermore, all standard SWFs favor complete equality in the simple resource allocation problem (without labor supply considerations), and in most optimal income tax simulations, all SWFs, including a utilitarian one, favor substantial redistribution, reflecting the nontrivial concavity of utility as a function of income. It is unclear how we could have reliable intuitions that would indicate whether any particular result entailed too much or too little weight on equality.[14]

Consider also the various meanings of equality. Under perhaps the most basic understanding of the notion, that everyone should be treated in the same manner, all standard SWFs adhere perfectly to the concept, for every individual is treated symmetrically. (In Harsanyi's second derivation of utilitarianism, recall that the linearity of the SWF was due to the rationality postulates; a separate value judgment of equality was invoked to yield the result that each individual's utility is to be weighted equally in the summation.) It is a property of a utilitarian SWF that individuals' marginal utilities count equally, whereas changes in utility levels per se are irrelevant. (As suggested with regard to the derivation from the original position in note 1, equating contributions to marginal utility across individuals is entailed by maxims such as the Golden Rule, which commands us to treat others as we would wish others to treat ourselves.) Under maximin, utility levels count equally, but marginal utility is immaterial. For other strictly concave SWFs, the result is in between; neither individuals' marginal utilities nor their utility levels matter equally even though utility functions enter the SWF symmetrically.

B. Membership in Society

Under welfarism, social welfare is taken to be a function of individuals' utilities. Which individuals should be included, however, is not self-evident. Questions of membership in society are among the most

[14] An additional problem is that our basic intuitions about redistribution probably pertain to the distribution of resources (income), which is what we observe being redistributed, not utils, which are the unit of measure of the arguments of the SWF.

15

Other Normative Criteria

For purposes of assessing taxation and redistribution, chapters 13 and 14 present a complete framework. That is, once it is decided that social welfare should depend exclusively on individuals' utilities, what the functional form of the SWF is, and whose utilities count, it would appear that there is no room for further normative analysis, much less for additional and potentially conflicting normative criteria.

This approach to distributive justice and social welfare, however, became established in applied work only with the emergence of the literature on optimal income taxation in the 1970s. Before then, more informal, intuitive notions of tax equity had an important role. Moreover, as section 3.A indicates, many such notions continue to be invoked today, perhaps in part because optimal income tax analysis is complex and its application to a wide range of subsidiary questions is not immediately obvious. For some of these other normative concepts, there exists a substantial technical literature developing measurement indexes—for example, there are two handbooks on income inequality per se, Atkinson and Bourguignon (2000) and Silber (1999)—and they are often employed in applied policy analysis as a supplement to or substitute for an SWF.

All such normative criteria are prima facie problematic. If the welfare economic framework is accepted, either competing measures must be proxy indicators for aspects of social welfare or they register something else, the pursuit of which inevitably entails the sacrifice of social welfare. In fact, both elements are often present. The continuing appeal of the other normative criteria seems in part attributable to their resonance with intuitions about distributive justice. The analysis in subsection 13.A.3 explains, however, that such intuitions are unreliable and can lead us astray, especially because the context of application differs substantially from that in which the intuitions originate. As will be seen, the residual usefulness of these various tax equity criteria as proxy principles varies greatly and may require that they be used differently from the manner suggested in the pertinent literatures.

A. Inequality, Poverty, Progressivity, Redistribution

Economists have developed indexes of inequality and poverty in a society and of the progressivity and redistribution associated with part or all of the fiscal system.[1] These measures are often employed to offer a normative assessment of taxation, the standard implication being that regimes exhibiting less inequality and poverty and tax schemes involving more progressivity and redistribution are superior.[2] As shown in Kaplow (2005), however, the literature does not offer an affirmative basis for this approach, and it does not appear that any justification can be provided. For concreteness, the reasoning will be presented for the case of inequality indexes, where the literature is most developed, and then briefly applied to measures of other traits of policies or outcomes.

Many inequality indexes are incomplete and ungrounded. Unless one distribution of income dominates another (that is, individuals below any given percentile level have more income under one income distribution than under another), inequality measures generally conflict. Some may satisfy certain axioms, but those in turn need justification—in principle, grounded in an SWF. For many measures, there exist inequality-reducing reforms that raise welfare under some SWFs but lower welfare under others. Most measures do not indicate the weight to be placed on inequality, so it is unclear how much a society should

[1] See, for example, on inequality: Atkinson and Bourguignon (2000), Cowell (1995), Lambert (2001), and Silber (1999); on poverty: Atkinson (1987), Clark, Hemming, and Ulph (1981), Lambert (2001), Ravallion (1994), Ruggles (1990), Sen (1976), and Silber (1999); and on progressivity and redistribution: Jakobsson (1976), Kakwani (1977), Lambert (1999, 2001), Musgrave and Thin (1948), and Suits (1977).

[2] The indexes also have descriptive uses, which raise different issues (see Kaplow 2005).

sacrifice to increase equality by any specified amount. Furthermore, many indexes are based on the distribution of incomes, but the meaning of income differences depends on individuals' utility functions. (For example, resolution of the question of whether inequality should be deemed constant if all incomes increase by the same proportion would presumably depend on the curvature of individuals' utilities as a function of income.)

To address such problems, Dalton (1920) argued that inequality measures had to be related to economic welfare, in essence that they needed to be based explicitly on an SWF. The most prominent such measure—associated with Atkinson (1970), Kolm (1969), and Sen (1973a)—is constructed in four steps. First, an SWF must be chosen; typical is the isoelastic reduced form introduced in expression (3.4) in subsection 3.B.1:

$$SW(x) = \int \frac{y^{1-\gamma}}{1-\gamma} f(y)dy, \text{ for } \gamma \neq 1, \tag{15.1}$$

where for simplicity all that is taken to matter in a regime x is each individual's disposable income, here denoted y following convention, and γ indicates the degree of aversion to inequality (which, recall, in this formulation is a sort of composite of the curvature in individuals' utility functions and in the social welfare function).[3] Second, using information on the density of the income distribution, $f(y)$, social welfare is calculated using (15.1). Third, one computes what is referred to as the equally distributed equivalent level of income, denoted y_e, which has the property that, if everyone had income at that level, social welfare would be the same as that computed in step 2. For the stated SWF (15.1), this can readily be determined as follows:

$$\frac{y_e^{1-\gamma}}{1-\gamma} = SW(x) = \int \frac{y^{1-\gamma}}{1-\gamma} f(y)dy, \text{ or}$$

$$y_e = \left[\int y^{1-\gamma} f(y)dy\right]^{1/(1-\gamma)}. \tag{15.2}$$

[3] The analysis to follow could be replicated for the case in which $\gamma = 1$ using ln y.

Fourth and finally, we construct the index of inequality, *I*, as follows:

$$I = 1 - \frac{y_e}{y_\mu}, \tag{15.3}$$

where y_μ refers to mean income. The ratio of equivalently distributed income to mean income, y_e/y_μ, indicates the portion of actual income that would be necessary to achieve the existing level of social welfare if instead that income were distributed equally. Therefore, the inequality index *I* indicates how much income could in principle be destroyed while leaving social welfare unchanged if only the remaining income were distributed equally.

Reflection upon this derivation, however, reveals that there is little possible use for the resulting inequality measure *I*. The index, of course, tells only part of the story; it must be combined with additional information to produce a full assessment of any regime *x*. Yet the full assessment was already obtained in step 2, when social welfare was calculated. In other words, to produce the partial indicator *I*, one must first derive the complete measure of social welfare and then perform further operations to strip away relevant information. Furthermore, no shortcut is possible. In particular, there is no way to derive *I* without first specifying an SWF and determining the level of social welfare under the pertinent regime. Different SWFs will produce different measures of *I*; furthermore, it was the very need to provide a normative grounding for inequality measurement that motivated the SWF-based approach. Because this approach does not allow construction of an inequality index based upon partial information, it is difficult to identify how the resulting measure can be useful.

Nevertheless, as noted at the outset, inequality indexes are often used to grade policies and governments. It should be apparent that such assessments are inherently incomplete and potentially misleading. Ceteris paribus, what inequality indexes omit—mean income—is in the relevant range negatively related to what is measured. This is the familiar equity-efficiency tradeoff. Knowing that a policy increases (or reduces) inequality or that one country's inequality is higher (lower) than another's does not indicate that the policy is detrimental (desirable) or that the government's performance is subpar (or superior).

The primary justification for the normative use of inequality measures is as a corrective to comparisons that focus excessively on per capita GDP, ignoring the distribution of income. Assessing policies or governments based purely on GDP in settings in which distribution is not held constant is similarly incomplete and misleading. Nevertheless, rather than supplementing GDP data with an inequality measure, which cannot be obtained without choosing an SWF and measuring social welfare, it would be superior to provide the social welfare assessment that had to be computed in step 2 along the way to generating the inequality measure. One of this book's themes, developed initially in section 3.A, is that explicit reference to the social objective is necessary. Consideration of inequality measures reinforces this view.

Poverty indexes may be analyzed similarly. For a normative measure to be policy relevant, it must be based on an SWF. Indeed, some poverty indexes are essentially truncated normative inequality indexes. The main differences are that poverty measures are even more incomplete and also are sensitive to the arbitrary judgment involved in drawing a poverty line and to the manner of determining how to assess the circumstances of individuals who fall below it. Attention to poverty is understandable, for if there is sufficient concavity in either utility functions or the SWF, the situation of individuals at the bottom of the income distribution will significantly influence the optimal design of redistribution policies. This fact, however, does not make it necessary to derive separate poverty measures, for the assessment provided by the SWF itself will by definition have already taken the importance of low-income individuals' situations into account—and will also reflect the impact of policies on other individuals, including those just above the poverty line, whose situations are also likely to be relevant.[4]

Indexes of progressivity and redistribution have similar shortcomings. In some instances, the connection is especially close, such as when the degree of redistribution is defined by the difference between the level

[4] For purposes of designing some transfer programs, there may be an administrative need for a cutoff, which might be set at some threshold. Yet the ideal level of such a cutoff for any particular program is the outcome of the optimization process, not an input or a constraint.

of inequality before and after tax, measured by an inequality index.[5] And progressivity, in turn, may be measured by the extent of redistribution. (If not, the measure is even more arbitrary, such as when one looks at ratios of average tax rates but not their magnitude; for example, very low rates, even if applied only to the very rich, might be quite progressive by some measures yet still be of little import.) In any event, to be of normative relevance, the measures must be derived from an SWF, but once the SWF is specified and social welfare is measured, a complete assessment already exists.[6]

B. Horizontal Equity

Horizontal equity, one of the basic notions of tax fairness advanced by Musgrave (1959), is the seemingly uncontroversial command that equals be treated (taxed) equally. Economists have generated a variety of indexes of horizontal inequity.[7] Typically, these indexes measure the extent of inequality of treatment of individuals under a tax reform whose incomes were equal prior to the reform (or in the pre-tax distribution of income or some other benchmark setting). Such horizontal inequity is viewed negatively, suggesting the need to trade it off against social welfare, the latter presumably defined by reference to a standard SWF. As developed in Kaplow (1989, 1995a, 2001b), however, this normative recommendation lacks affirmative justification and accordingly results

[5] Also, the literature engaging in explicit normative measurement derives indexes from an SWF, often drawing directly on Atkinson (1970).

[6] Some advocates directly assess taxation (sometimes the entire system, sometimes a single tax instrument) by reference to whether it is progressive or proportional. For example, Blum and Kalven's (1952, 1953) well-known essay criticizes sacrifice theories as well as other principles and concludes that taxes should be proportional, but as Bankman and Griffith (1987) and Groves (1974), among others, have noted, their conclusion rests on little more than a presumption in favor of proportionality. (Furthermore, since Blum and Kalven would allow exemptions and do not restrict how tax proceeds may be spent, they implicitly allow progressivity in the standard economic sense of rising average rates and also substantial redistribution, which seems inconsistent with most of their critique.)

[7] See, for example, Aronson and Lambert (1994), Auerbach and Hassett (2002), Atkinson (1980), Feldstein (1976), King (1983), Musgrave (1990), and Plotnick (1981).

in a needless sacrifice in welfare, possibly everyone's welfare. The intuitive appeal of horizontal equity, moreover, may be understood by the manner in which it serves as a proxy indicator for other possible sources of welfare reduction, although standard indexes are not well suited for this surrogate function.

To begin, horizontal equity has familiar definitional problems. What if few individuals (or none) are precise equals in the benchmark distribution? What is the basis for seemingly ad hoc judgments about how pre-reform individuals are to be grouped and post-reform differences are to be weighted? Are rank reversals, which may or may not involve unequal treatment of pre-reform equals, similarly problematic? And finally, can indexes that address some of these and other challenges still be properly understood as measures of horizontal inequity? Some attempts to address these questions offer decompositions of standard SWFs, wherein one component is deemed to be a measure of horizontal inequity. However, the basis for such decompositions seems largely semantic, and the purpose for distinguishing certain influences on social welfare is unclear.

The shortcomings of horizontal inequity indexes are similar to those of the inequality and related indexes examined in section A. For example, if one needs to specify an SWF and have complete information on its determinants, there seems no good reason to undertake further computations (some of which require further information) to produce an index that represents merely an aspect of social welfare. More worrisome is the suggestion, often explicit but sometimes implicit, that one component of social welfare should be given more weight than another. Why should a one-util shortfall to individual *i* count more than some other one-util shortfall to the same individual *i* because the former is classified as one particular form of inequity? The literature does not attempt to answer this basic question or, relatedly, indicate why a distinct measure of horizontal inequity is required in the first instance.[8]

[8] Many measures of horizontal inequity have further defects. For example, most define the equals who should be treated equally by reference to a pre-reform status quo. However, once a reform that reduces horizontal equity is implemented, the resulting regime becomes the status quo; hence, repeal can only further reduce horizontal equity, and under most indexes it ordinarily would. One might add more broadly that reforms are motivated by the existence of defects in the status quo, which would seem to contraindicate placing a negative value on departures from the status quo per se. These problems can be avoided by specifying

Furthermore, it follows from the general discussion of nonwelfarist principles in subsection 13.A.2 that giving any weight to horizontal inequity conflicts with the Pareto principle.[9] Consider, for example, a society of two individuals, identical in all respects ex ante, and a reform that raises one individual's utility and lowers the other's by a smaller amount, with each individual having an equal chance of occupying each position. Ex ante expected utility rises with the reform, so both would favor it. If, however, any weight is put on the violation of horizontal equity (the equals will not be treated equally), there exist cases—where the expected utility increase is sufficiently small (one can reduce the mean but maintain the variance)—in which the reform will be opposed.[10]

Despite these deficiencies in horizontal equity indexes, the maxim that equals should be treated equally has intuitive appeal, and this attraction can readily be explained. The discussion in subsection 13.A.3 suggests that we often have moral intuitions that arise in contexts distinct from the setting of optimal tax design but that nevertheless seem compelling and may have a proxy role even though they are not independent moral principles to be pursued at the expense of individuals' well-being. Most directly, in ordinary settings, if it maximizes social welfare to subject individual *i* to treatment *X* rather than *Y*, and if individual *j* is identical in all relevant respects to *i*, then *j* also should receive treatment *X*. If we observe that *i* receives *X* and *j* receives *Y*, it is likely that a mistake has been made. In other words, optimization

some idealized distribution as the benchmark rather than the status quo, but why only deviations involving unequal treatment of equals should matter—or why they should matter differently from other deviations—is obscure. Why would we not evaluate any regime or reform by reference to the objective function (SWF) from which the idealized distribution was derived? These sorts of difficulties reinforce the suggestion that the literature has developed multiple, conflicting indexes without first specifying the purpose of measurement.

[9] There are a number of ways to see that measures of horizontal inequity are nonwelfarist. If not all utils are weighted in the same manner, then information on each individual's utility level obviously is insufficient to form a social judgment. Also, most indexes make reference to the pre-reform income distribution, the pre-tax distribution, or some other benchmark, indicating that information other than individuals' utilities in the actually prevailing state is relevant.

[10] This is essentially the same example (from Kaplow 1995a) used in subsection 14.A.1.b.

typically implies equal treatment, so unequal treatment typically signifies a failure to optimize.[11] In such instances, it is necessary to identify the mistake (is *j* being treated suboptimally, or *i*, or possibly both?) and correct it. The value of such correction is already indicated by the SWF. Moreover, correction of mistreatment is valuable even if there exists no other individual who was already being treated correctly.

Unequal treatment of equals sometimes indicates particular sorts of welfare reductions. It may suggest the existence of greater income inequality (for example, a reform may not seem to affect the income distribution when data consists of decile averages, but dispersion within deciles might have increased).[12] Similarly, as suggested by the foregoing example involving a violation of the Pareto principle, horizontal inequity may correlate with exposure to risk, which is welfare reducing, ceteris paribus. Another sort of concern involves the possibility of corruption and various other abuses of power, which in important instances involve unequal treatment (racial discrimination, government favors to political allies).[13] Indeed, this problem probably best explains the intuitive force of the equal treatment principle.

The question remains how best to analyze these issues. Income inequality and risk are already fully assessed under a standard SWF, applied to the resulting distribution of income. Abuse of power, by contrast, might be detected by searching for deviant decisions or discrimination across pertinent classifications (race, political affiliation). None of these and other concerns, however, seem well captured by standard indexes of horizontal inequity, which compare certain components of dispersion between, say, the income distribution before and after imposition of a tax reform that applies generally to a large, otherwise anonymous population.

[11] There are subtle exceptions, such as in the case of nonconvexities, see Stiglitz (1982b), or when randomization is optimal (for example, with tax audits). Additionally, it is sometimes the case that otherwise sensible policies will result in incidental unequal treatment of individuals on account of their preference heterogeneity. See subsection 6.C.4 and also the analysis in subsection 5.C.2 of the implications of different preferences for consumption and leisure for the optimal income tax problem.

[12] See Atkinson (1980).

[13] See, for example, Mirrlees (1982) and Stiglitz (1982b).

Consideration of horizontal equity may seem orthogonal to most of the issues considered in this book, but this is not always the case. Appeals to horizontal equity are made, for example, when tax base concepts are elevated to evaluative norms (see section F) in order to resolve myriad issues such as tax base design (income versus consumption taxation), when uniform commodity taxation is favored independently of the applicability of economic arguments, including those that may justify deviations, and when it is argued that certain family types should be treated in the same manner as others. Horizontal equity is also invoked in addressing more detailed design questions, such as whether a deduction should be allowed for medical expenses, and in matters of administration and enforcement. Again, as a proxy principle, there may be some value (generally speaking, differences in treatment do need justification). Ultimately, however, direct use of the welfare economic approach is both necessary and sufficient.

Consider, for example, the problem of the taxation of families (chapter 12), where horizontal equity is often employed in attempts to resolve disputes about proper relative taxation. By definition, horizontal equity cannot tell us which groupings of individuals are the "equals" who are supposed to be treated equally, for that begs the very question at issue.[14] Confining attention to matters of distribution, the optimal treatment of various family configurations was seen to depend on a number of subtle factors that could be analyzed only by making explicit reference to an SWF (or a class of SWFs). Additionally, whatever would be optimal on distributive grounds must be adjusted in light of incentive considerations that may differ by family type for reasons largely unrelated to distributive considerations. The concept of horizontal equity does not contribute to these inquiries.

Furthermore, the apparently clear guidance that horizontal equity seems to offer in other settings may turn out to be misleading, if not mistaken. For example, it is fairly widely accepted that consideration of horizontal equity favors a tax deduction for medical expenses and casualty losses on the ground that individuals suffering from such bad luck should be compared to individuals not suffering such misfortune but

[14] This point about the notion of equal treatment is quite general, as emphasized by Westen (1990).

whose incomes are otherwise lower by the extent of such losses. This conclusion, however, ignores ex ante incentives, including the incentive to obtain insurance, which may well make a tax deduction counterproductive.[15] In sum, even the proxy role for horizontal equity should be viewed cautiously.

C. Sacrifice Theories

For centuries, a prominent view has been that taxpayers should make equal sacrifices in contributing to the cost of government activity.[16] There is disagreement about whether there should be equal absolute sacrifice (the rich and the poor should suffer the same absolute decline in utility), equal proportional sacrifice, or equal marginal sacrifice. For the most part, this view expresses an intuition rather than an implication of a developed theory of distributive justice. The exception is that the equal marginal sacrifice principle, advanced by Edgeworth (1897) and Pigou (1928), is a corollary of utilitarianism for the familiar reason that maximizing total utility implies equating individuals' marginal utilities (in the first best).

The intuition favoring equal sacrifice, like other moral intuitions examined in subsection 13.A.3.b, can be understood as being useful in the informal settings in which it probably arose—such as by serving as a focal point in organizing contributions to a group activity—while serving at best as a proxy indicator for an aspect of social welfare maximization in the context of the design of government policy. Regarding the latter, sacrifice theories purport to be applicable to how public goods are financed and are silent regarding other taxation, notably for redistributive purposes. Yet as long as redistributive taxation is unconstrained—suppose that the income tax may be freely adjusted to maximize a given SWF—it is immaterial how tax rates are provisionally set to meet a revenue requirement that covers the cost of public goods.[17]

[15] See, for example, Kaplow (1992b).

[16] On the history of thought relating to this principle, see Musgrave (1959).

[17] Nor can this difficulty be avoided by deeming redistribution—for example, expenditures on transfer programs—to be a public good, for it makes no sense to talk of equal sacrifice by both the rich and the poor when net payments involve a flow from the rich to the poor.

If instead the entire tax system must adhere to a sacrifice principle—that is, if (perhaps following a libertarian view) the only permissible use of taxation is to finance public goods—then the implications of sacrifice theories for redistribution have a substantially arbitrary character. The reason is that the sacrifice theories specify the distributive properties of taxation independently of the distributive incidence of the public goods being financed.[18] To take a simple example, suppose that there is some project that will provide a uniform benefit of $100 per capita. Further, imagine that the market could provide this benefit at a cost (and would charge a price) of $90 per capita. The government, by contrast, can provide the benefit at a cost of $89 per capita (charging no price). Finally, assume that under the prevailing equal sacrifice norm, the required tax would charge the rich $300 and the poor $30 in taxes if the government were to undertake the project.[19] Obviously this equal sacrifice tax redistributes income relative to a world in which the good is privately provided. The rich pay $300 for a benefit worth $100, for a net loss of $200, whereas the poor pay $30 for a benefit worth $100, for a net gain of $70. With private provision, each individual would have a net gain of $10. To reinforce this point, suppose that the technology changes so that market provision becomes more efficient—say, the market cost drops to $88 per capita—and it is therefore decided that the government should cease providing the benefit. Now, each individual pays $88 for a benefit of $100, a net gain of $12 per capita.

This example illustrates that the norms of equal sacrifice may call for substantial redistribution or none (one could also construct examples of negative redistribution, that is, net transfers to the rich) depending on the distributive incidence of public goods and, for example, on small differences in technology that have nothing to do with any plausible normative principle governing redistribution. Put another way, once one accounts for the distributive incidence of expenditures, which

[18] This is the mirror image of the situation examined in section 8.A in which the distributive incidence of a public good was taken as given and different income tax adjustments were considered (see figures 8.2A–8.2C).

[19] For any of the equal sacrifice norms except equal marginal sacrifice (which would require virtual equality), one could state utility functions and initial levels of income such that the tax payments postulated in the text would satisfy the pertinent norm.

chapter 8 emphasizes is critical, the norms do not have the implications one would expect. After all, how can one say that the rich and the poor make equal sacrifices when, as in the initial example, the rich have a net loss and the poor a net gain? In sum, the notion of equal sacrifice, despite its intuitive appeal in certain settings, is not a helpful guide in addressing the problem of taxation and redistribution. Accordingly, it is not surprising that it no longer receives significant attention.

D. Benefit Principle

The benefit principle (already examined in section 8.F on benefit taxation) is like the equal sacrifice norms in that it is limited to the question of how public goods should be financed and hence does not seem to address the question of redistributive taxation. It differs from the equal sacrifice norms in that it dictates that the incidence of taxes used to finance public goods should match the distributive incidence of the public goods being financed. Therefore, if all taxation had to be in accord with the benefit principle, no redistribution would occur. However, if redistributive taxation is permitted, then, like the sacrifice theories, it is essentially moot.

As explored previously, the benefit theory of taxation nevertheless receives some attention in modern public economics writing. There is controversy over the correct notion of benefit taxation although the purpose of such a canonical formulation is not apparent. Moreover, there is interest in whether benefit taxation is progressive, a concern that seems immaterial since, whatever is the incidence of benefit taxation, one might presume that it tends to offset the incidence of the public goods being financed. Finally, sections 8.E and 8.F raised the baseline problem: Are the benefits of public goods to be measured relative to a hypothetical state of anarchy or, if not, against what other benchmark? As explained there, no such baseline is required under the welfare economic framework either to decide which public goods should be provided or to determine how redistributive taxation should optimally be configured.

Under certain libertarian views, by contrast, redistribution is impermissible and only benefit taxation would be allowed. Accordingly, it would be necessary to state a baseline, for the distribution prevailing in that state is the one deemed to be normatively required under a just regime of taxation. A baseline of anarchy is problematic; even if it could be defined, it would preserve the advantages of the strong who successfully prey on the weak, the core injustice designed to be prevented by the libertarian minimal state. Also, there remains the question of how to allocate the surplus (which may constitute nearly all existing value) in moving from anarchy to a just regime. A contrary baseline with rightful behavior yet no public sector is problematic because, if individuals are entitled to their allotment in that state, no one would have to pay taxes. Finally, a baseline with rightful behavior and the minimal state in place begs the question of what tax regime, which is part of the minimal state, is normatively required. Some might insist that taxes be proportional—or satisfy one of the equal sacrifice theories—but any such stipulation appears arbitrary.

E. Ability to Pay

The notion that tax burdens should reflect individuals' "ability to pay" is commonly used in arguments about the ideal tax base and rate structure.[20] Just as with the norms of equal sacrifice and the benefit principle, however, the concept of ability to pay seems literally to be addressed only to the question of how to raise revenue to finance government programs, ignoring the question of redistribution to which it is often applied. After all, why would we ask what the poor are able to pay when we intend for them to receive?[21] Setting that basic problem aside, the concept nevertheless remains elusive. Anyone is *able* to pay any amount that he possesses (although if he starves as a result, he will not be able to pay anything in the next tax period), but this says nothing about what individuals *should* pay. Those suggesting that taxes be in accord with

[20] Musgrave (1959) traces this principle to the sixteenth century and identifies among its prominent supporters Rousseau and John Stuart Mill.

[21] And, as with the equal sacrifice norms, even regarding the finance of public goods, the distributive incidence of the public goods being financed is ignored, even though this incidence importantly influences the net distributive effect.

ability to pay obviously have in mind that the rich should pay more than the poor. But this uncontroversial dictate does not tell us how much more, what this difference may depend upon (rate of diminution in the marginal utility of income? elasticity of labor supply?), or why this is so. In sum, the underlying justification is unstated and the prescription is so imprecise as to render the principle of little use—or, reflecting actual practice, to leave it so open-ended as to mean whatever a proponent wishes it to mean (see Vickrey 1947, pp. 3–4, 374–375).

F. Definitions as Norms

Particularly in addressing questions about the appropriate tax base, various definitions are often treated as if they constitute normative principles of tax equity. For example, the familiar Haig-Simons definition of income—the sum of an individual's consumption plus change in wealth—has been taken by many as the test of an ideal tax base. A notable example is Simons (1938) himself, but one must include much subsequent advocacy of comprehensive income taxation and analogous arguments in favor of a comprehensive consumption tax. Also related are debates about the merits of the tax expenditure concept, which is often taken to suggest that any deviations from the idealized tax base are presumptively inappropriate.[22]

Definitions of income and the like are necessary for communication and sometimes are clarifying (for example, standard definitions make apparent that costs of producing income need to be subtracted from gross receipts). Nevertheless, it is apparent that such definitions are not in themselves normative principles. Moreover, as with many of the other normative criteria examined in this chapter, they are incomplete. For example, they do not indicate how much value is lost on account of various departures from the supposedly ideal base, information that is necessary whenever tax administration and enforcement are not costless.

[22] This idea is most associated with Surrey (1973). For a range of views, see, for example, Bittker (1969), Griffith (1989), Shaviro (2004), Surrey and McDaniel (1985), and Weisbach and Nussim (2004).

Furthermore, many particular debates—such as concerning transfe taxation and taxation of the family—raise ambiguities or are not addressed by such definitions. Of course, even when the definitions ar clear, one can always argue over which of various definitions to adopt a a guide in determining the tax base. Accordingly, it is not surprising tha the once-widespread practice of using definitions as norms, althoug still of some significance, is waning.