CESifo WORKING PAPERS

7910
2019
October 2019

Markets, Queues, and Taxes

Erwin Ooghe

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7910*

Markets, Queues, and Taxes

Abstract

In his thought-provoking book What Money Can't Buy. The Moral Limits of Markets, Sandel (2012) claims that some nonmarket ways of allocating goods, such as the ethics of the queue (first come, first served), are gradually being displaced by the ethics of the market. He highlights inequality as one of two reasons why we should care about this tendency: “In a society where everything is for sale, life is harder for those of modest means” (Sandel, 2012, p. 8). I investigate whether queuing can improve redistribution in a second-best setting where also commodity and earnings taxes are available. I specify first a set of bench-mark assumptions - reminiscent of the Atkinson-Stiglitz model - and show that it is never optimal to introduce queuing. It suggests, contrary to Sandel, that introducing more market and less queuing improves the life of ‘those of modest means.’ Afterwards, I also relax some of the bench-mark assumptions. Two cases pro queuing seem promising: differentiated queuing and paternalism.

JEL-Codes: D470, D630, H210.

Keywords: market allocation, queuing, earnings taxes, commodity taxes.

Erwin Ooghe
Department of Economics
KU Leuven
Belgium – 3000 Leuven
erwin.ooghe@kuleuven.be

October 16, 2019

1 Motivation

Sandel's (2012) thought-provoking book *What Money Can't Buy. The Moral Limits of Markets* has received considerable attention, also from economists.[1] He argues that the United States, and probably also other countries, are drifting from "having a market economy" to "being a market society."

> We live at a time when almost everything can be bought and sold. Over the past three decades, markets—and market values—have come to govern our lives as never before. [...] The reach of markets, and market-oriented thinking, into aspects of life traditionally governed by nonmarket norms is one of the most significant developments of our time. (Sandel, 2012, pp. 5–7)

But why should we worry about this tendency? Sandel highlights inequality as one of two possible reasons.[2]

> In a society where everything is for sale, life is harder for those of modest means. The more money can buy, the more affluence (or the lack of it) matters. [...] Where all good things are bought and sold, having money makes all the difference in the world. (Sandel, 2012, p. 8)

According to Sandel (2012, p. 28), one of the nonmarket ways of allocating goods that is being displaced by the ethics of the market is the ethics of the queue, i.e., first come, first served. While he praises the ethics of the queue for its egalitarian appeal, he also mentions its potential downside.

> To an economist, long lines for goods and services are wasteful and inefficient, a sign that the price system has failed to align supply and demand. Letting people pay for faster service at airports, at amusement parks, and on highways improves economic efficiency. (Sandel, 2012, p. 21)

While it is true that economists often point to the inefficiency of using queues to allocate goods and services, many economists also stress that inefficiency is

[1] See, e.g., McCloskey (2012), Besley (2013), Bruni (2013), and Calel (2013) for book reviews.

[2] This paper focuses only on inequality. The other reason according to Sandel (2012, p. 9) is "about the corrosive tendency of markets. Putting a price on the good things in life can corrupt them. [...] Economists often assume that markets [...] do not affect the goods they exchange. But this untrue. Sometimes, market values crowd out nonmarket values worth caring about."

one side only of the equity-efficiency trade-off. Nichols, Smolensky, and Tideman (1971), for example, discuss the pros and cons of queuing as follows: "Since time is more equally distributed than money, this rationing device [queuing] may be thought to be desirable because of equity considerations even though it is known to be economically inefficient" (Nichols, Smolensky, and Tideman, 1971, p. 313).

This potential of queuing to improve equity has been stressed several times in the mainstream economic literature.[3] Nichols, Smolensky, and Tideman (1971) observe that queuing is less costly for individuals with lower wages and conclude therefore that queues can be an efficient device to redistribute towards lower wage individuals. Barzel (1974) shows that the poor do not necessarily benefit from the provision of free goods on a first-come first-served basis, as it may imply that the poor end up with less of this good. While Barzel looks at the distribution of the good itself, other authors focus on the distribution of utility and arrive at the opposite conclusion. Sah (1987) shows that, for the poor, queuing outperforms the market (in utility terms), while the opposite is true for the rich. Polterovich (1993) looks at the transition from rationing by queuing towards a market system and demonstrates that, in line with Sah, the transition hurts the poor and benefits the rich. Weitzman (1977) introduces needs differences and compares the market and queuing in terms of their ability to deliver a commodity to those who need it most. Queuing turns out to be more effective if the distribution of needs is more uniform or if the distribution of income is less uniform.

Besides the potential to enhance equity, some economists also point to the use of differentiated queuing schemes as an efficiency-enhancing screening device. Nichols, Smolensky, and Tideman (1971) write:

> In some cases, it may be desirable to charge many different money prices for the identical publicly subsidized commodity. Queues of different lengths will form with the shortest queues occurring at the facilities with the highest money prices. Individuals will then have a choice of paying for a commodity with various combinations of money and time, each choosing that combination which is cheapest for him. There may be substantial efficiency gains to be had from such differentiation. (Nichols, Smolensky, and Tideman, 1971, p. 322).

Similarly, Clarke and Kim (2007) propose a differentiated "pay or wait"-scheme for the public provision of a private good such that its final allocation does not

[3]There also exists a more specialised literature on the axiomatic characterization of allocation mechanisms in queueing models. Besides efficiency, fairness can be a key concern in the proposed mechanisms too; see, e.g., Maniquet (2003), Chun (2006), Kayı and Ramaekers (2010), and Chun, Mitra, and Mutuswami (2014).

depend on income, but increases in strength of preference/need.

While queuing can be used as an efficiency-enhancing screening device, the appreciation of queuing in the economics literature is mainly equity-based. But the immediate next question is: what can queuing add in comparison to other more standard redistribution mechanisms? Unfortunately, standard redistribution mechanisms—say, commodity and earnings taxes—are often underdeveloped or completely neglected in this literature. A few exceptions exist; see Bucovetsky (1984), O'Shaughnessy (2000), and Alexeev and Leitzel (2001).[4] Closest to the spirit of this paper is Bucovetsky (1984). He introduces linear commodity taxation and a lump-sum grant and shows that, even if we can use these tax instruments optimally, queuing, especially for luxuries, will further enhance social welfare.

While the economic literature typically uses abstract reasoning to uncover the properties of market versus nonmarket allocation mechanisms, Sandel (2012) objects this approach: "Whether, in any given case, markets or queues do this job [allocating goods and services] better is an empirical question, not a matter that can be resolved in advance by abstract economic reasoning" (Sandel, 2012, p. 32). I agree that abstract reasoning cannot decide *in advance* whether one system is superior compared to another. Yet, I do think that abstract economic reasoning offers a logically coherent way to uncover both the normative and the positive-empirical assumptions under which one system can be expected to outperform another. In this paper, I therefore investigate the following question: can queuing be desirable from a societal point of view and, if so, under what conditions?

In section 2, I introduce a fairly general model based on four assumptions that define the nature of the goods, the choice behavior of individuals, the normative goal of society, and the available policy instruments.[5] I show that, under the assumptions made, it is never optimal to introduce queuing. Contrary to Sandel's claim, this result suggests that introducing more market and less queuing has the potential to improve the lifes of 'those of modest means.' In section 3, I look at the different assumptions underlying the bench-mark model. Because the number of possible relaxations is large, the aim is not to provide a detailed analysis for each possible case, but rather to discuss a number of relaxations that seem promising to justify queuing. My discussion points to two promising routes. The first route—introduce queuing only for low earners as a deterrence mechanism for high earners—is known to economists, but may feel uncomfortable

[4]It is fair to say that also Sah (1987) looks at very specific forms of commodity taxation.

[5]The spirit of this model is close to the bench-mark model introduced by Atkinson and Stiglitz (1976) and further refined in the public economics literature to analyse the optimal tax mix between commodity and earnings taxes. I use a simpler and more pedagogical two-type version; see., e.g., Boadway (2012, chapter 3).

to some philosophers. The second route—paternalistically introducing non-utility information into social welfare—has been recommended by philosophers in this field, but could be unacceptable to some economists. A final section 4 concludes.

2 A case against queuing

In this section I introduce four bench-mark assumptions that describe the nature of the goods, the choice behavior of individuals, the normative goal of society, and the available policy instruments. Taken together, these assumptions provide a case against queuing. While I introduce all models and results in an informal way—with an occasional footnote to provide some formal details for the interested reader—the appendix contains a formal statement of all models, results, and proofs.

First, all goods are private (i.e., rival and excludable) without blatant externalities. The production technology of each good exhibits constant returns to scale such that they are bought and sold at fixed prices in a competitive market environment. For simplicity, I consider two goods only: the first good can be thought of as a composite good (say, the expenditures for all goods except one) and the second good is the good that is targeted for queuing. The (so-called) queuing good can be obtained by a mixture of queuing and paying. Special cases include a pure price mechanism (the price is sufficiently high such that queuing is not required) and a pure queuing mechanism (the price is zero and rationing is completely based on queuing). More will be said about prices and queues where I discuss the available policy instruments. Finally, it can be ensured (at low costs) that the queuing good cannot be bought elsewhere and cannot be resold afterwards.[6] All in all, a regular doctor's visit, cleaning services, and a music concert are possible examples of goods that can be targeted for queuing.

Assumption 1 (the goods). There exist two private goods with no externalities and constant returns to scale. The first good is a composite good. The second good is a queuing good: besides a price to be paid, also some queuing is required to obtain the good. The queuing good cannot be bought elsewhere and cannot be resold.

Second, individuals have preferences over goods (the composite and the queuing good) and activities (working and queuing). Similar to working, queuing can be interpreted as the time spent in a queue, but also effort to obtain some of

[6] Both the production and provision of the queuing good can be organised publicly or privately. If private, strict regulation and control of the private sector is obviously required.

the good.[7] All individuals like the two goods, but dislike working (to generate earnings) and queuing (to specifically obtain the queuing good), ceteris paribus. Preferences can be represented by a (twice) differentiable utility function that is additively separable between the goods and the activities. Additive separability implies that the willingness to pay for one of the goods does not depend on the activity level.[8] The two activities, working and queuing, are perfect substitutes.[9] While queuing is not productive, working generates gross (pre-tax) earnings at an individual-specific, but constant rate, called ability. Individuals differ in ability and, for ease of exposition, ability can be either low or high. Besides abilities, individuals are equal in all relevant aspects (e.g., preferences and needs). Moreover, individuals are rational: they choose a bundle (a combination of goods and activities) that they like most (according to their preferences) among the feasible bundles (given the prices of the goods, the required queuing, and the ability of the individual).

Assumption 2 (the individuals). Individuals rationally choose the best bundle (given their preferences) among the feasible ones (given their opportunities). They have the same preferences that can be represented by a twice differentiable utility function that is additively separable between the goods (the composite and the queuing good) and the activities (working and queuing). Goods are desirable, ceteris paribus, while activities are not desirable. Working and queuing are perfect substitutes. While queuing is not productive, working is productive at a constant rate (called ability). Individuals differ only in ability (and thus not in preferences or needs) and ability can only be low or high.[10]

Third, the goal of society is to redistribute from high- to low-ability individuals, because low-ability individuals have worse opportunities (for which they are not held fully accountable). We leave open how much redistribution is desired, but we

[7] While I usually stick to the term queuing, as opposed to queuing time or queuing effort, I will sometimes explicitly use queuing time in my examples for simplicity.

[8] So, from the three examples introduced before, cleaning services (probably with a higher willingness to pay at higher levels of activity) and concerts (probably with a higher willingness to pay at lower levels of activity) are excluded for now.

[9] This seems fine if the interpretation of queuing is time or effort, similar to labour time or effort, but is not compatible with a broader interpretation of queuing as waiting time that is not spent in a queue (think, e.g., of waiting time to get non-urgent surgery) and whose disutility is probably independent of labour time.

[10] Formally, individuals (rationally) choose the composite and the queuing good (x_1, x_2) and the amount of labour ℓ to maximize utility $U(x_1, x_2, q, \ell) = u(x_1, x_2) - v(\eta q + \ell)$ subject to a budget constraint (to be discussed later on). The utility-of-consumption function u is strictly increasing and concave, the disutility-of-activity function v is strictly increasing and convex, and the scalar η is strictly positive. Note that the amount of queuing follows from the chosen amount of the queuing good (to be discussed later on). Gross (pre-tax) earnings are $y = \theta\ell$, with $\theta > 0$ the ability (constant marginal productivity), which can be low or high ($\theta_H > \theta_L > 0$).

do impose that redistribution must be done in Pareto efficient, i.e., in an efficient and non-paternalistic way. Efficiency requires that it is not possible to further improve the well-being of the low-ability individuals without harming the high-ability individuals (and vice-versa). Non-paternalism requires that the evaluation of individual well-being is consistent with individual preferences: if an individual prefers one bundle over another, then individual well-being must be higher in the former bundle.

Assumption 3 (the societal goal). Society wants to redistribute from high-ability to low-ability individuals in a Pareto efficient, i.e., an efficient and non-paternalistic way.[11]

Fourth, because ability is not observable, non-distortive (first-best) lump-sum tax instruments are not feasible and redistribution is limited to distortive (second-best) instruments. Non-linear taxes on gross earnings as well as linear per-unit taxes on the goods are possible instruments.[12] In addition to earnings and commodity taxation, queuing can also be used as a policy instrument. The type of queuing is per-unit: each unit of the queuing good requires a fixed number of queuing units (say, time or effort). This so-called queuing rate will be defined in equilibrium such that the resulting demand for the queuing good is equal to the supply, which is set by the government.[13] Because the equilibrium queuing rate and the choice of the supplied quantity are one-to-one, the queuing rate will be the instrument and supply adjusts automatically to be equal to demand.

Assumption 4 (the policy instruments). Society can optimally set (linear) commodity taxes and (non-linear) earnings taxes. In addition, society can optimally set a constant queuing rate per unit of the queuing good and supply automatically adjusts to equate demand.[14]

The core question is: can it be useful to introduce queuing? Because queuing is unproductive, it creates an efficiency loss. Yet, queuing also gives rise to an efficiency gain in the current second-best setting. This gain follows from the

[11] Pareto-efficient redistribution boils down to choosing the commodity tax rate, the queuing rate, and (net and gross) earnings (for each type) to maximize the utility of the low ability type subject to three constraints: a minimal level of utility for each high ability type, a self-selection constraint (such that the high ability type does not mimick the low ability type), and a budget constraint; see Boadway (2012 : 62-63). Appendix A provides further details.

[12] As usual, negative taxes correspond with subsidies.

[13] If supply exceeds demand, then the queuing rate is zero in equilibrium.

[14] The budget constraint of an individual can now be formally defined as $x_1 + (p_2 + t_2)x_2 \leq c = y - T(y)$, with the composite good chosen as the numéraire, t_2 the per-unit linear tax rate on the queuing good, and T the non-linear tax scheme as a function of gross earnings. The total amount of queuing q is equal to $q_2 x_2$, with q_2 the (chosen) per-unit queuing rate of good 2.

fact that queuing is especially costly for high-ability types as they have a higher opportunity cost. Because second-best earnings taxation is constrained by the fact that high-ability types may mimick low-ability types, queuing allows to reduce mimicking and results therefore in an efficiency gain.

Whether queuing is, in the end, useful depends on the magnitude of both efficiency effects. Under the assumptions made, it turns out that the efficiency loss is always larger than the efficiency gain, and hence, it is never optimal to introduce queuing.

Theorem 1. Under assumptions 1–4, it is never optimal to introduce queuing.

Under these assumptions, it is well-known that also (differentiated) commodity taxes are not useful; see, e.g., Atkinson and Stiglitz (1976). So, one may conclude that the market, corrected by earnings taxes only, is the best institution under the assumptions made. While Sandel (2012) states that we should worry about replacing queuing by markets as it makes life harder for 'those of modest means,' our first result suggests the opposite. If the assumptions are reasonable, then more market and less queuing will improve the utility of 'those of modest means,' being, the low-ability types. The key question is how important the different assumptions are, our next topic.

3 Cases in favor of queuing?

Cases in favor of queuing may arise if one or more assumptions are relaxed. As it is impossible to discuss all possible combinations of relaxations, I will focus on some promising cases.[15] Note from the beginning that we often obtain the conclusion that queuing *can be* optimal. It does not follow that queuing *is* necessarily optimal. If queuing can be optimal, it means that I cannot prove that queuing is not optimal and I also cannot prove that it is optimal. So, these cases hint at possibilities requiring further analysis.

First, theorem 1 only says that *introducing* queuing is not optimal, which is, in the end, a local result only. In addition, it also assumes that society can optimally design its tax instruments. As a first extension, proposition 1 shows that, given a strictly positive, but sufficiently small queuing rate and an arbitrary commodity and earnings tax scheme, it is possible to construct a sequence of feasible and

[15]The following cases are not discussed (with my a priori, if any, between brackets): (i) the queuing good is a public good or has externalities, (ii) the queuing good can be bought elsewhere or the queuing good can be resold (this is likely to make the case for queuing weaker), (iii) there are more than two ability types (does not matter), and (iv) society wants to redistribute form low to high ability types (not interesting, but, in any case, proposition 1 is an efficiency result that does not depend on the direction of redistribution).

Pareto-improving reforms to end up with a policy without queuing and without (differentiated) commodity taxation.[16]

Proposition 1. Suppose assumptions 1-4 hold, except that the different instruments are not necessarily designed in an optimal way. Suppose that the queuing rate is strictly positive, but not too high (such that the consumption of the queuing good is largest for the high ability type). Under these modified assumptions, it is possible to construct a sequence of feasible and Pareto-improving reforms such that, in the end, both the queuing rate and the (commodity) tax rate are equal to zero.

One might be tempted to rephrase Theorem 1 and Proposition 1 in a positive way: introducing queuing can be optimal if taxes can neither be optimized nor changed. Yet, this does not seem to be very plausible from a political economy perspective: why would society be willing to introduce queuing if, after its introduction, everybody in society can be made better off by reforming the system, including abolishing queuing?

Second, constant returns to scale imply that supply is perfectly elastic. This does not fit with the description of some goods, like land or kidneys, that are sometimes discussed in the queuing literature and whose supply is closer to being fixed in advance (and perfectly inelastic). As supply cannot be chosen anymore, queuing and supply are no longer one-to-one in equilibrium. Given a fixed supply, the tax rate and the queuing rate are still one-to-one. So, the planner sets the queuing rate and afterwards the tax rate (being the consumer price if the production cost is zero) is used to equate demand and supply. Proposition 2 tells us that it remains suboptimal to introduce queuing.

Proposition 2. Suppose assumptions 1-4 hold, except that the queuing good is in fixed supply (rather than produced via a CRS-technology). Under these modified assumptions, it is not optimal to introduce queuing.

Third, the two activities, queuing and labour, were assumed to be perfectly substitutable up to now, excluding a broader interpretation of queuing as waiting time (for, e.g., non-urgent surgery, whose disutility is probably unrelated to the amount of labour). The next proposition 3 says that as long as queuing and labour are not complementary (which does not exclude the possibility of queuing to be independent of labour as in its interpretation of waiting), it remains suboptimal to introduce queuing.

[16] The question remains why it would be more plausible that a society can do sequences of local policy changes, but not globally optimize its instruments. Therefore, the importance of proposition 1 is, in my view, that it relaxes the fact that Theorem 1 is a local result only about introducing queuing.

Proposition 3. Suppose assumptions 1-4 hold, except that queuing and labour are perfect substitutes. If queuing and labour are not complementary, i.e., it is not true that more queuing is less unpleasant at higher levels of labour, then it is not optimal to introduce queuing.

It is possible to give a positive twist to proposition 3: if queuing and labour are complementary (i.e., more queuing is less unpleasant at higher levels of labour), it can be optimal to introduce queuing. Although ultimately an empirical question, complementarity is not very plausible in my view.[17]

Fourth, for the goods, services, and events that are potential queuing goods (e.g., a regular doctor's visit, cleaning services, and a music concert), it seems not very difficult or costly to also record the actual consumption levels. This opens new possibilities as society could also use this information to design the tax and the queuing scheme. If we also allow for fully flexible preferences—relaxing the additive separability of preferences and the the perfect substitutability of the activities—then proposition 4 provides weaker conditions under which it is not optimal to introduce queuing.[18]

Proposition 4. Suppose assumptions 1-4 hold, except that the consumed quantity of the queuing good is observable to the planner such that taxes and queuing can each be made contingent on this additional information. Suppose, in addition, that preferences are fully flexible (in particular, they are not necessarily separable and the two activities are not necessarily perfect substitutes). Under these modified assumptions, it is never optimal to introduce queuing for the high-ability individuals. For the low-ability individuals, it is also not optimal to introduce queuing if the following two conditions hold: (1) for at least one good, more of it is not more pleasant at higher levels of labour *and* (2) more queuing is more unpleasant at higher levels of labour.[19]

The fact that the high ability type remains undistorted is a classical result in optimal taxation; see Stiglitz (1982). Proposition 4 provides sufficient conditions under which it is not optimal to introduce queuing for the low ability types. So, for queuing to be optimal, at least one of the sufficient conditions must be violated. A violation of the first condition requires that for all goods it is true that more of the

[17] If one interprets queuing as waiting (for non-urgent surgery, with a small risk caused by anaesthesia), then this waiting could be less unpleasant if you work (as you do not have time to worry about such negative consequences).

[18] While proposition 3 allows preferences to be fully flexible, it also allows for non-linear commodity taxation. Thus, a case not discussed is what happens if preferences were fully flexible (including non-separability) and quantities cannot be observed. This is an open case.

[19] Note indeed that both conditions are satisfied under (additive) separability and perfect substitutability.

good is more pleasant at higher levels of labour. Generally speaking, this does not seem to be true for some goods, like sporting equipment, as people probably derive more utility from such goods if they have less labour and thus more leisure time, ceteris paribus. A violation of the second condition requires that more queuing is less unpleasant at higher levels of labour. This case has been discussed before and seems a priori rather unlikely: an extra amount of queuing seems to be harder to swallow after a full day of work compared to not having worked that day. To sum up, while proposition 4 provides an opening, it seems a priori unlikely—but not impossible—that these conditions will be violated. And even if violated, such a violation is necessary only, so, it is not per se true that introducing queuing will in the end be optimal.

Fifth, up to now we focused on ability differences, but also needs and tastes differences may occur. While adding needs and taste differences on top of ability differences is complex, it is worthwhile to think of replacing ability differences by either taste or needs differences. As with ability differences, taste or needs differences are assumed to be unobservable and thus first-best solutions remain out of reach.[20] Suppose individuals only differ in tastes (over commodities, not activities). In this case, everyone has the same earnings and redistribution is possible only through commodity taxation. But why should we want to introduce different commodity taxes in this case? One reason could be that we assign a higher weight to individuals with specific tastes, e.g., individuals with strong tastes for opera relative to other goods. While possible, many economists would object this possibility. Saez (2002 : 222), for example, writes: "one might think that, in a liberal society, the government should not set judgment values on the citizens based on their consumption." Even if one accepts this normative proposal—that specific tastes should get higher welfare weights—the question remains whether queuing is useful as an additional device, because also subsidizing opera goes some way in redistributing from high to low taste types. I postpone this discussion for the moment because the same question pops up if we look at differences in needs. If individuals only differ in needs, then assigning a higher weight to individuals with higher needs seems acceptable. Yet, the final question is the same as before: what role can queuing play in addition to, e.g., subsidising the good that is needed more by some? Proposition 5 shows that it is not only suboptimal to introduce queuing, but also to have queuing (it shows essentially that decreasing the subsidy and simultaneously decreasing the queuing hurts no one, is implementable, and increases revenues). While the proposition focuses on needs, the same model also

[20] See Boadway and Pestieau (2006) for the optimal treatment of (observable and unobservable) needs differences in optimal taxation.

applies to tastes (if one would like to redistribute from low to high tastes for the second good).[21]

Proposition 5. Suppose assumptions 1-4 hold, except that individuals do not differ in ability, but in (unobservable) needs. Suppose society wants to redistribute from individuals with low needs to individuals with high needs. Under these modified assumptions, it is not optimal to queue.

Sixth, maybe it is not optimal to let everyone queue, but rather to introduce differentiated queuing schemes, e.g., so-called “pay or wait”-schemes where individuals can choose either to pay a price for the good without waiting or to wait for the good without paying.[22] There are essentially two ways to do differentiate. Either we offer joint bundles (containing the queuing rate, the tax rate, and the net and gross earnings), one for each type (so two bundles in total), and subject to self-selection constraints. Or, we offer separate bundles, one containing the queuing and the tax rate and one containing the net and gross earnings, one for each type (so four bundles in total), and subject to self-selection constraints. The last case is not only more complex (more self-selection constraints), but it is also the weakest case from the planner’s point of view because the set of self-selection constraints is larger (and the possibilities of the planner are more restricted). I therefore focus on the easy, but strongest possible case in favor of queuing. Proposition 6 shows that it can be optimal to introduce queuing, if the difference in ability is sufficiently large. If true, then the high ability types must buy the queuing good at full cost and without queuing (i.e., they remain undistorted again), while the low types must not only queue for the good, but also pay more than the full cost.

Proposition 6. Suppose assumptions 1-4 hold, except that we can differentiate the queuing and the tax rate (jointly with earnings taxes) over the different types. Under these modified assumptions, it is always optimal to leave the high ability types undistorted (i.e., they are not required to queue and pay no commodity taxes such that their consumer price is equal to the unit cost of production). In addition, if the ability difference between high and low ability types is sufficiently high, then introducing both queuing and commodity taxes for the low ability types *is* optimal.

Proposition 6 tells us that it is not “pay or wait,” but “pay and wait,” which recall, can be implemented only because low ability individuals cannot choose their

[21]Note that tastes and needs are modelled in the same way by letting the type play the role of a minimal required amount of consumption as in Cremer, Pestieau, and Rochet (2001).

[22]Clark and Kim (2007) analyze such extreme “pay or wait”-schemes in a context without redistribution.

queuing and tax rate separate from their (net and gross) earnings. The reason that "pay and wait" is optimal is that queues and taxes reinforce each other in the deterrence of mimicking. Once queuing is introduced, the mimicker will consume more of the queuing good because, having a high ability level, he must exert less effort to match earnings, which reduces the (utility) cost of consuming extra units of the queuing good. But, following the classical Atkinson-Stiglitz logic, a higher consumption implies that the queuing good must be taxed to further deter mimicking. Yet, although "pay and wait" seems detrimental for the low ability types at first sight, recall that this mechanism is optimal in a global sense. Because it allows to redistribute in a more efficient way, the low ability types will be better off in the end as they can and will be more than compensated for paying and waiting through earnings taxation.

Seventh, assumption 3 consists of two parts, efficiency and non-paternalism, and each part can be criticized. One could criticize efficiency because some efficiency improvements might not be desirable, e.g., if the richest person in society improves his well-being, ceteris paribus. A much weaker condition is to impose efficiency only for distributions exhibiting complete equality in well-being. In other words, if there is no well-being inequality in each of two states, then the state with the higher well-being is more desirable. Yet, proposition 1 remains true. The proof of proposition 1 shows essentially that the second-best frontier shrinks everywhere (i.e., also in the non-efficient parts of the frontier) if queuing is introduced. As a consequence, any welfare function that is utility-based and that (minimally) accepts efficiency for equal distributions, must accept proposition 1 and shy away from queuing policies.

One could also criticize non-paternalism, i.e., the assumption that well-being and utility are ordinally equivalent. In the queuing literature, such arguments have been put forward. Nichols et al (1971) introduce merit good considerations and Weitzman (1977) and Clark and Kim (2007) focus on specific egalitarianism.[23] The presence of merit goods—i.e., goods that are relatively undervalued, opera maybe, and therefore underconsumed—implies that well-being and utility are not equivalent. Similarly, specific egalitarianism—i.e., inequality in the consumption of specific goods or services, like regular doctor visits, matters besides inequality in utility—also implies that the current welfarist set-up breaks down. The question is to see whether queuing can be useful if the societal goal is not based on utilities only. Proposition 7 focuses on merit goods and tells us that an opening may occur.[24]

[23] A related argument could be made if individuals are not fully rational.

[24] I focus on merit goods because the arguments in favor of specific egalitarianism for certain

Proposition 7. Suppose assumptions 1-4 hold, except that the queuing good is a good (like opera) with specific merits that are not (sufficiently) captured by utility.[25] Under these modified assumptions, it *can be* optimal to introduce queuing.

4 Conclusion

In this paper we start from the claim that market mechanisms, compared to nonmarket mechanisms like queuing, are worrysome for 'those of modest means.' While there is a literature in economics that has investigated this claim, this literature typically eschews other redistributive mechanisms, such as commodity and earnings taxes.

I introduced a bench-mark model (in the spirit of the Atkinson-Stiglitz model) and showed that is never optimal to introduce queuing. Contrary to Sandel's claim, this means that introducing more market and less queuing will improve redistribution towards 'those of modest means,' at least if these assumptions apply. So, I also look at some promising relaxations of the different assumptions. Two routes turn out to be promising in my view.

The first route, differentiated queuing, is known to economists, but is also likely to be unpopular with philosophers. I show that if the difference in ability between the high and low type is sufficiently large, then it can be optimal to introduce a "pay *and* wait"-scheme for the low ability type and leave the high ability type undistorted. Although this seems devastating for the welfare of the poor at first sight, they will be more than compensated through the earnings tax scheme such that, in the end, they will be better off.

The second route, paternalism, is popular among philosophers, but likely to be unpopular with economists. There could indeed be reasons to distrust utility as an indicator of well-being, e.g., because individuals systematically underestimate the value of certain goods (merit goods) or because there could be serious concerns about inequality in specific goods (specific egalitarianism). In such cases, queuing *can be* optimal. More work needs to be done here to also uncover cases where queuing *is* optimal.

---

goods seem to be based on special merits of these goods.

[25] I assume that well-being is $V_i + \delta x_{2i}$ for a type i-individual, with $\delta > 0$. In contrast to the evaluation, choice is driven by V_i.

References

Alexeev M., Leitzel, J., 2001, Income distribution and price controls: targeting a social safety net during economic transition, *European Economic Review* 45, 1647-1663.

Atkinson, A., Stiglitz, J, 1976, The Design of Tax Structure: Direct versus Indirect Taxation, *Journal of Public Economics* 6 (1–2), 55-75.

Barzel, Y., 1974, A theory of rationing by waiting, *The Journal of Law and Economics* 17(1), 73-95.

Besley, T., 2013, What's the good of the market? An essay on Michael Sandel's "What money can't buy," *The Journal of Economic Literature* 51(2), 478-495.

Bruni, L., 2013, A book review of "What money can't buy: the moral limits of markets," *International Review of Economics* 60, 101-106.

Bucovetsky, S., 1984, On the use of distributional weights, *The Canadian Journal of Economics* 17(4), 699-717.

Calel, R., 2013, A book review of "What money can't buy: the moral limits of markets," *Economics and Philosophy* 29(2), 277-283.

Chun, Y., 2006, No-envy in queueing problems, *Economic Theory* 29, 151-162.

Chun, Y., Mitra, M., Mutuswami, S., 2014, Egalitarian equivalence and strategyproofness in the queueing problem, *Economic Theory* 56(2), 425-442.

Clarke, J., Kim, B., 2007, Paying vs. waiting in the pursuit of specific egalitarianism, *Oxford Economic Papers* 59(3), 486-512.

Kayı, Ç., and Ramaekers, E., 2010, Characterizations of Pareto-efficient, fair, and strategy-proof allocation rules in queueing problems, *Games and Economic Behavior* 68, 220-232.

Maniquet, F., 2003, A characterization of the Shapley value in queueing problems, *Journal of Economic Theory* 109, 90-103.

McCloskey, D., 2012, The poverty of communitarianism. A review of "What money can't buy: the moral limits of markets," *The Claremont Review of Books* XII(4), 57-59.

Nichols, D., Smolensky, E., Tideman, T.N., 1971, Discrimination by waiting time in merit goods, *The American Economic Review* 61(3), 312-323.

O'Shaughnessy, T., 2000, The taxing issues of queues, unpublished manuscript.

Polterovich, V., 1993, Rationing, queues, and black markets, *Econometrica* 61(1), 1-28.

Saez, E., 2002, The desirability of commodity taxation under non-linear income taxation and heterogeneous tastes, *Journal of Public Economics* 83, 217-230.

Sah, R.K., 1987, Queues, rations, and market: comparisons of outcomes for the poor and the rich, *The American Economic Review* 77(1), 69-77.

Sandel, M., 2012, *What Money Can't Buy: The Moral Limits of Markets*, New York: Farrar, Straus and Giroux.

Weitzman, M., 1977, Is the price system or rationing more effective in getting a commodity to those who need it most? *The Bell Journal of Economics* 8(2), 517-524.

A The model

I first study the problem of an individual, look at some comparative statics, derive several properties (Roy and Slutsky) for later use, and finish with the problem of society.

The problem of the individual. Consider two goods, good 1 (the composite good) and good 2 (the good targeted for queuing). Let p_2+t_2 denote the consumer price for the queing good, with $p_2 > 0$ the producer price and t_2 the per-unit tax rate (or subsidy, if negative). Good 1 is chosen to be the numéraire and its price $p_1 + t_1$ can be normalized to one without loss of generality. Working and queuing are perfect substitutes, so, activity can be defined as $a = \eta q + \ell$, with $\eta > 0$ a preference parameter, q the amount of queuing, and ℓ the amount of labour. The amount of queuing is equal to $q_2 x_2$, with q_2 the per-unit queuing rate for good 2. Preferences can be represented by a differentiable utility function defined over consumption and activity, say, $U(x_1, x_2, a)$.[26] Because individuals like consumption and dislike activity, the (partial) derivatives satisfy $\frac{\partial U}{\partial x_1} > 0$, $\frac{\partial U}{\partial x_2} > 0$, and $\frac{\partial U}{\partial a} < 0$; in addition, we assume that the marginal utility of consumption and activity decrease, i.e., $\frac{\partial^2 U}{\partial (x_1)^2} < 0$, $\frac{\partial^2 U}{\partial (x_2)^2} < 0$, and $\frac{\partial^2 U}{\partial a^2} < 0$.[27] Finally, net (post-tax) earnings are denoted by c and gross (pre-tax) earnings are denoted by y and are equal to $\theta\ell$, with $\theta > 0$ the ability level of an individual; the difference $y - c$ is the earnings tax (or subsidy if negative).

Conditional on (net and gross) earnings c and y and conditional on ability level θ, a rational individual chooses consumption to solve

$$\max_{x_1, x_2} U(x_1, x_2, \eta q_2 x_2 + \frac{y}{\theta})$$

subject to his or her budget constraint

$$x_1 + (p_2 + t_2)x_2 \leq c.$$

[26]Because utility is additively separable between consumption and activity, there must exist strictly increasing differentiable functions u and v such that $U(x_1, x_2, a) = u(x_1, x_2) - v(a)$ holds everywhere. Because some of the results in this section hold without this assumption, we use additive separability (and u and v) only if needed.

[27]In the case of additive separability, these conditions reduce to $\frac{\partial u}{\partial x_1} > 0$, $\frac{\partial u}{\partial x_2} > 0$, $\frac{\partial^2 u}{\partial (x_1)^2} < 0$, and $\frac{\partial^2 u}{\partial (x_2)^2} < 0$ for the utility-of-consumption function u and $\frac{dv}{da} > 0$ and $\frac{d^2 v}{da^2} > 0$ for the utility-of-activity function v.

Assuming an interior solution, the first-order conditions are

$$\frac{\partial U}{\partial x_1}(x_1, x_2, \eta q_2 x_2 + \frac{y}{\theta}) = \lambda, \tag{1}$$

$$\frac{\partial U}{\partial x_2}(x_1, x_2, \eta q_2 x_2 + \frac{y}{\theta}) + \frac{\partial U}{\partial a}(x_1, x_2, \eta q_2 x_2 + \frac{y}{\theta})\eta q_2 = \lambda(p_2 + t_2), \tag{2}$$

$$x_1 + (p_2 + t_2)x_2 = c, \tag{3}$$

with $\lambda > 0$ the Lagrange multiplier. Let $x_1(t_2, q_2, c, y, \theta)$ and $x_2(t_2, q_2, c, y, \theta)$ denote the solution.

For later use, we look at the comparative statics and derive some equations known as Roy equations and Slutsky equations.

Comparative statics. For later use, I am interested in the behaviour of $x_2(\cdot)$ with respect to its components. Combining the first-order conditions, we can eliminate λ and $x_1(\cdot)$ to obtain

$$-\frac{\partial U}{\partial x_1}(c - (p_2 + t_2)x_2(\cdot), x_2(\cdot), \eta q_2 x_2(\cdot) + \frac{y}{\theta})(p_2 + t_2)+$$
$$\frac{\partial U}{\partial x_2}(c - (p_2 + t_2)x_2(\cdot), x_2(\cdot), \eta q_2 x_2(\cdot) + \frac{y}{\theta})+$$
$$\frac{\partial U}{\partial a}(c - (p_2 + t_2)x_2(\cdot), x_2(\cdot), \eta q_2 x_2(\cdot) + \frac{y}{\theta})\eta q_2 = 0.$$

First, I look at the tax rate. Differentiating both sides with respect to t_2, we obtain

$$-SOC_2(\cdot)\frac{\partial x_2}{\partial t_2}(\cdot) = \frac{\partial^2 U}{\partial (x_1)^2}(\cdot)x_2(\cdot)(p_2+t_2) - \frac{\partial U}{\partial x_1}(\cdot) - \frac{\partial^2 U}{\partial x_1 \partial x_2}(\cdot)x_2(\cdot) - \frac{\partial^2 U}{\partial x_1 \partial a}(\cdot)\eta q_2 x_2(\cdot),$$

where $SOC_2(\cdot)$ denotes the second-order condition with respect to good 2 (and which is thus negative).[28] If utility is additively separable—i.e., $U(x_1, x_2, \eta q_2 x_2 + \frac{y}{\theta}) = u(x_1, x_2) - v(\eta q_2 x_2 + \frac{y}{\theta})$ everywhere, with $\frac{dv}{da} > 0$ and $\frac{d^2 v}{da^2} > 0$—then we simply get

$$\frac{\partial x_2}{\partial t_2}(\cdot) = \frac{\frac{\partial^2 u}{\partial (x_1)^2}(\cdot)x_2(\cdot)(p_2 + t_2) - \frac{\partial u}{\partial x_1}(\cdot) - \frac{\partial^2 u}{\partial x_1 \partial x_2}(\cdot)x_2(\cdot)}{-SOC_2(\cdot)}$$

whose sign is not defined.

Second, I look at the queuing rate. Differentiating both sides with respect to

[28] More precisely, it is the second-order condition of the maximization problem after substitution (maximize $U(c - (p_2 + t_2)x_2, x_2, \eta q_2 x_2 + \frac{y}{\theta})$ with respect to x_2).

q_2, we obtain

$$-SOC_2(\cdot)\frac{\partial x_2}{\partial q_2}(\cdot) = -\frac{\partial^2 U}{\partial x_1 \partial a}(\cdot)\eta x_2(\cdot)(p_2+t_2)+\frac{\partial^2 U}{\partial x_2 \partial a}(\cdot)\eta x_2(\cdot)+\frac{\partial^2 U}{\partial a^2}(\cdot)\eta^2 q_2 x_2(\cdot)+\frac{\partial U}{\partial a}(\cdot)\eta.$$

If utility is additively separable—i.e., $U(x_1, x_2, \eta q_2 x_2+\frac{y}{\theta}) = u(x_1, x_2)-v(\eta q_2 x_2+\frac{y}{\theta})$ everywhere, with $\frac{dv}{da} > 0$ and $\frac{d^2 v}{da^2} > 0$—then we simply get

$$\frac{\partial x_2}{\partial q_2}(\cdot) = -\frac{\frac{d^2 v}{da^2}(\cdot)\eta^2 q_2 x_2(\cdot) + \frac{dv}{da}(\cdot)\eta}{-SOC_2(\cdot)} < 0. \tag{4}$$

Third, I look at net earnings. Differentiating both sides now with respect to c, we obtain

$$-SOC_2(\cdot)\frac{\partial x_2}{\partial c}(\cdot) = -\frac{\partial^2 U}{\partial (x_1)^2}(\cdot)(p_2 + t_2) + \frac{\partial^2 U}{\partial x_1 \partial x_2}(\cdot) + \frac{\partial^2 U}{\partial x_1 \partial a}(\cdot)\eta q_2,$$

which, using additive separability, reduces to

$$\frac{\partial x_2}{\partial c}(\cdot) = \frac{-\frac{\partial^2 u}{\partial (x_1)^2}(\cdot)(p_2 + t_2) + \frac{\partial^2 u}{\partial x_1 \partial x_2}(\cdot)}{-SOC_2(\cdot)}.$$

Fourth, I focus on gross earnings. Differentiating both sides now with respect to y, we obtain

$$-SOC_2(\cdot)\frac{\partial x_2}{\partial y}(\cdot) = -\frac{\partial^2 U}{\partial x_1 \partial a}(\cdot)\frac{1}{\theta}(p_2 + t_2) + \frac{\partial^2 U}{\partial x_2 \partial a}(\cdot)\frac{1}{\theta} + \frac{\partial^2 U}{\partial a^2}(\cdot)\frac{1}{\theta}\eta q_2,$$

If utility is additively separable, then we simply get

$$\frac{\partial x_2}{\partial y}(\cdot) = -\frac{\frac{d^2 v}{da^2}(\cdot)\frac{1}{\theta}\eta q_2}{-SOC_2(\cdot)} < 0. \tag{5}$$

Finally, I look at ability. The reason is that a mimicker compared to a low type only differs in ability. If we differentiate both sides with respect to θ, then we obtain

$$-SOC_2(\cdot)\frac{\partial x_2}{\partial \theta}(\cdot) \quad = \quad -\frac{\partial^2 U}{\partial x_1 \partial a}(\cdot)(p_2 + t_2)(-\frac{y}{\theta^2}) + \frac{\partial^2 U}{\partial x_2 \partial a}(\cdot)(-\frac{y}{\theta^2}) + \frac{\partial^2 U}{\partial a^2}(\cdot)(-\frac{y}{\theta^2})\eta q_2.$$

If utility is additively separable, then we get

$$\frac{\partial x_2}{\partial \theta}(\cdot) \quad = \quad \frac{\frac{d^2 v}{da^2}(\cdot)\frac{y}{\theta^2}\eta q_2}{-SOC_2(\cdot)} > 0. \tag{6}$$

Roy. At the optimum, equation (3) reads

$$x_1(t_2, q_2, c, y, \theta) + (p_2 + t_2)x_2(t_2, q_2, c, y, \theta) = c.$$

Differentiating both sides (with respect to t_2, q_2, c and y) leads to

$$\begin{aligned}
\frac{\partial x_1}{\partial t_2}(t_2, q_2, c, y, \theta) + (p_2 + t_2)\frac{\partial x_2}{\partial t_2}(t_2, q_2, c, y, \theta) &= -x_2(t_2, q_2, c, y, \theta), &(7)\\
\frac{\partial x_1}{\partial q_2}(t_2, q_2, c, y, \theta) + (p_2 + t_2)\frac{\partial x_2}{\partial q_2}(t_2, q_2, c, y, \theta) &= 0, &(8)\\
\frac{\partial x_1}{\partial c}(t_2, q_2, c, y, \theta) + (p_2 + t_2)\frac{\partial x_2}{\partial c}(t_2, q_2, c, y, \theta) &= 1, &(9)\\
\frac{\partial x_1}{\partial y}(t_2, q_2, c, y, \theta) + (p_2 + t_2)\frac{\partial x_2}{\partial y}(t_2, q_2, c, y, \theta) &= 0. &(10)
\end{aligned}$$

The derivative of indirect utility, defined as

$$V(t_2, q_2, c, y, \theta) = U(x_1(t_2, q_2, c, y, \theta), x_2(t_2, q_2, c, y, \theta), \eta q_2 x_2(t_2, q_2, c, y, \theta) + \frac{y}{\theta}), \quad (11)$$

with respect to earnings (c and y) is equal to

$$\begin{aligned}
\frac{\partial V}{\partial c}(t_2, q_2, c, y, \theta) &\overset{(11),(1),(2)}{=} \lambda\frac{\partial x_1}{\partial c}(t_2, q_2, c, y, \theta) + \lambda(p_2 + t_2)\frac{\partial x_2}{\partial c}(t_2, q_2, c, y, \theta),\\
&\overset{(9)}{=} \lambda, &(12)\\
\frac{\partial V}{\partial y}(t_2, q_2, c, y, \theta) &\overset{(11),(1),(2)}{=} \lambda\frac{\partial x_1}{\partial y}(t_2, q_2, c, y, \theta) + \lambda(p_2 + t_2)\frac{\partial x_2}{\partial y}(t_2, q_2, c, y, \theta) +\\
&\qquad \frac{\partial U}{\partial a}(x_1, x_2, \eta q_2 x_2 + \frac{y}{\theta})\frac{1}{\theta},\\
&\overset{(10)}{=} \frac{\partial U}{\partial a}(x_1, x_2, \eta q_2 x_2 + \frac{y}{\theta})\frac{1}{\theta}, &(13)
\end{aligned}$$

where the numbers above the equality sign refer to the equations used. Similarly, the derivative of indirect utility with respect to the tax and the queuing rate is equal to

$$\begin{aligned}
\frac{\partial V}{\partial t_2}(t_2,q_2,c,y,\theta) &\overset{(11),(1),(2)}{=} \lambda\frac{\partial x_1}{\partial t_2}(t_2,q_2,c,y,\theta)+\lambda(p_2+t_2)\frac{\partial x_2}{\partial t_2}(t_2,q_2,c,y,\theta), \\
&\overset{(7)}{=} -\lambda x_2(t_2,q_2,c,y,\theta), \\
&\overset{(12)}{=} -\frac{\partial V}{\partial c}(t_2,q_2,c,y,\theta)x_2(t_2,q_2,c,y,\theta), \qquad (14) \\
\frac{\partial V}{\partial q_2}(t_2,q_2,c,y,\theta) &\overset{(11),(1),(2)}{=} \lambda\frac{\partial x_1}{\partial q_2}(t_2,q_2,c,y,\theta)+\lambda(p_2+t_2)\frac{\partial x_2}{\partial q_2}(t_2,q_2,c,y,\theta)+ \\
&\qquad \frac{\partial U}{\partial a}(x_1,x_2,\eta q_2x_2+\frac{y}{\theta})\eta x_2(t_2,q_2,c,y,\theta), \\
&\overset{(8)}{=} \frac{\partial U}{\partial a}(x_1,x_2,\eta q_2x_2+\frac{y}{\theta})\eta x_2(t_2,q_2,c,y,\theta), \\
&\overset{(13)}{=} \frac{\partial V}{\partial y}(t_2,q_2,c,y,\theta)\theta\eta x_2(t_2,q_2,c,y,\theta). \qquad (15)
\end{aligned}$$

Equation (14) is the usual Roy equation for the tax rate, while equation (15) is a Roy equation for the queuing rate.

Slutsky. The dual problem of the individual is to minimize expenditures, i.e.,

$$\min_{x_1,x_2} x_1+(p_2+t_2)x_2$$

subject to reaching a minimal utility level

$$U(x_1,x_2,\eta q_2x_2+\frac{y}{\theta})\geq u.$$

Let $\tilde{x}_1(t_2,q_2,u,y,\theta)$ and $\tilde{x}_2(t_2,q_2,u,y,\theta)$ be the solution. Duality implies that x_2 and $\tilde{x}_2$ coincide if we choose the minimal utility level u to be equal to the indirect utility level, i.e., we have

$$x_2(t_2,q_2,c,y,\theta)=\tilde{x}_2(t_2,q_2,V(t_2,q_2,c,y,\theta),y,\theta). \qquad (16)$$

Differentiating both sides with respect to net and gross earnings, one gets

$$\begin{aligned}
\frac{\partial x_2}{\partial c}(t_2,q_2,c,y,\theta) &\overset{(16)}{=} \frac{\partial \tilde{x}_2}{\partial u}(t_2,q_2,V(t_2,q_2,c,y,\theta),y,\theta)\frac{\partial V}{\partial c}(t_2,q_2,c,y,\theta), \qquad (17) \\
\frac{\partial x_2}{\partial y}(t_2,q_2,c,y,\theta) &\overset{(16)}{=} \frac{\partial \tilde{x}_2}{\partial u}(t_2,q_2,V(t_2,q_2,c,y,\theta),y,\theta)\frac{\partial V}{\partial y}(t_2,q_2,c,y,\theta)+ \\
&\qquad \frac{\partial \tilde{x}_2}{\partial y}(t_2,q_2,V(t_2,q_2,c,y,\theta),y,\theta). \qquad (18)
\end{aligned}$$

Differentiating both sides with respect to the tax rate and the queing rate leads to

$$\begin{aligned}\frac{\partial x_2}{\partial t_2}(t_2,q_2,c,y,\theta) &\overset{(16)}{=} \frac{\partial \tilde{x}_2}{\partial t_2}(t_2,q_2,V(t_2,q_2,c,y,\theta),y,\theta) \\ &\quad +\frac{\partial \tilde{x}_2}{\partial u}(t_2,q_2,V(t_2,q_2,c,y,\theta),y,\theta)\frac{\partial V}{\partial t_2}(t_2,q_2,c,y,\theta), \\ &\overset{(14),(17)}{=} \frac{\partial \tilde{x}_2}{\partial t_2}(t_2,q_2,V(t_2,q_2,c,y,\theta),y,\theta) - \\ &\quad \frac{\partial x_2}{\partial c}(t_2,q_2,c,y,\theta)x_2(t_2,q_2,c,y,\theta), \qquad (19)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial x_2}{\partial q_2}(t_2,q_2,c,y,\theta) &\overset{(16)}{=} \frac{\partial \tilde{x}_2}{\partial q_2}(t_2,q_2,V(t_2,q_2,c,y,\theta),y,\theta) \\ &\quad +\frac{\partial \tilde{x}_2}{\partial u}(t_2,q_2,V(t_2,q_2,c,y,\theta),y,\theta)\frac{\partial V}{\partial q_2}(t_2,q_2,c,y,\theta), \\ &\overset{(15),(18)}{=} \frac{\partial \tilde{x}_2}{\partial q_2}(t_2,q_2,V(t_2,q_2,c,y,\theta),y,\theta) \\ &\quad -\frac{\partial \tilde{x}_2}{\partial y}(t_2,q_2,V(t_2,q_2,c,y,\theta),y,\theta)\theta\eta x_2(t_2,q_2,c,y,\theta) \\ &\quad +\frac{\partial x_2}{\partial y}(t_2,q_2,c,y,\theta)\theta\eta x_2(t_2,q_2,c,y,\theta). \qquad (20)\end{aligned}$$

Equation (19) is the usual Slutsky equation for the tax rate, while equation (20) is a Slutksy equation for the queuing rate.

The problem of society. Society wants to redistribute from high to low ability types in a Pareto efficient way. Suppose that $n_L > 0$ individuals have ability level $\theta_L > 0$ and $n_H > 0$ individuals have ability level $\theta_H > \theta_L$. The societal problem is to choose the policy instruments to maximize the utility level of the low ability individuals

$$\max_{t_2,q_2,c_L,y_L,c_H,y_H} V(t_2,q_2,c_L,y_L,\theta_L)$$

subject to a minimal (exogenous) utility level $\underline{V}$ reserved for the high ability types

$$V(t_2,q_2,c_H,y_H,\theta_H) - \underline{V} \geq 0, \qquad (21)$$

subject to the self-selection constraint (to guarantee that high ability types do not mimick low ability types, which in turn guarantees that the earnings can be implemented via an earnings tax scheme),

$$V(t_2,q_2,c_H,y_H,\theta_H) - V(t_2,q_2,c_L,y_L,\theta_H) \geq 0, \qquad (22)$$

and subject to a budget constraint (to guarantee that the total tax revenues can finance some exogenous revenue requirement $\underline{R}$),

$$n_L(t_2x_2(t_2,q_2,c_L,y_L,\theta_L)+y_L-c_L)+n_H(t_2x_2(t_2,q_2,c_H,y_H,\theta_H)+y_H-c_H)-\underline{R}\geq 0. \tag{23}$$

For ease of exposition, I abbreviate the (good 2) consumption and (indirect) utility of low and high ability individuals as $x_{2L}(\cdot)=x_2(t_2,q_2,c_L,y_L,\theta_L)$, $x_{2H}(\cdot)=x_2(t_2,q_2,c_H,y_H,\theta_H)$, $V_L(\cdot)=V(t_2,q_2,c_L,y_L,\theta_L)$, and $V_H(\cdot)=V(t_2,q_2,c_H,y_H,\theta_H)$. Because high ability individuals can mimick low ability individuals, it is useful to introduce the consumption and indirect utility of a mimicker as $x_{2M}(\cdot)=x_2(t_2,q_2,c_L,y_L,\theta_H)$ and $V_M(\cdot)=V(t_2,q_2,c_L,y_L,\theta_H)$. Let $\alpha>0$, $\beta>0$, and $\gamma>0$ denote the Langrange multipliers for the minimal utility, self-selection, and budget constraint. The derivatives of the Lagrangian (denoted $\mathscr{L}$) with respect to the policy instruments are

$$\frac{\partial\mathscr{L}}{\partial c_L} = \frac{\partial V_L(\cdot)}{\partial c}-\beta\frac{\partial V_M(\cdot)}{\partial c}+\gamma n_L(t_2\frac{\partial x_{2L}(\cdot)}{\partial c}-1), \tag{24}$$

$$\frac{\partial\mathscr{L}}{\partial y_L} = \frac{\partial V_L(\cdot)}{\partial y}-\beta\frac{\partial V_M(\cdot)}{\partial y}+\gamma n_L(t_2\frac{\partial x_{2L}(\cdot)}{\partial y}+1), \tag{25}$$

$$\frac{\partial\mathscr{L}}{\partial c_H} = (\alpha+\beta)\frac{\partial V_H(\cdot)}{\partial c}+\gamma n_H(t_2\frac{\partial x_{2H}(\cdot)}{\partial c}-1), \tag{26}$$

$$\frac{\partial\mathscr{L}}{\partial y_H} = (\alpha+\beta)\frac{\partial V_H(\cdot)}{\partial y}+\gamma n_H(t_2\frac{\partial x_{2H}(\cdot)}{\partial y}+1), \tag{27}$$

$$\begin{aligned}\frac{\partial\mathscr{L}}{\partial t_2} &= \frac{\partial V_L(\cdot)}{\partial t_2}+(\alpha+\beta)\frac{\partial V_H(\cdot)}{\partial t_2}-\beta\frac{\partial V_M(\cdot)}{\partial t_2}\\ &\quad+\gamma(n_L(x_{2L}(\cdot)+t_2\frac{\partial x_{2L}(\cdot)}{\partial t_2})+n_H(x_{2H}(\cdot)+t_2\frac{\partial x_{2H}(\cdot)}{\partial t_2})),\end{aligned} \tag{28}$$

$$\begin{aligned}\frac{\partial\mathscr{L}}{\partial q_2} &= \frac{\partial V_L(\cdot)}{\partial q_2}+(\alpha+\beta)\frac{\partial V_H(\cdot)}{\partial q_2}-\beta\frac{\partial V_M(\cdot)}{\partial q_2}\\ &\quad+\gamma(n_Lt_2\frac{\partial x_{2L}(\cdot)}{\partial q_2}+n_Ht_2\frac{\partial x_{2H}(\cdot)}{\partial q_2}).\end{aligned} \tag{29}$$

To see that (introducing) queuing creates both an efficiency gain and an efficiency loss, we first show that, under separability and assuming no queuing, the optimal tax rate on good 2 is equal to zero. Equation (28) can be rewritten as

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial t_2} &\overset{(14)}{=} -\frac{\partial V_L(\cdot)}{\partial c}x_{2L}(\cdot) - (\alpha+\beta)\frac{\partial V_H(\cdot)}{\partial c}x_{2H}(\cdot) + \beta\frac{\partial V_M(\cdot)}{\partial c}x_{2M}(\cdot) + \\
&\quad \gamma(n_L(x_{2L}(\cdot) + t_2\frac{\partial x_{2L}(\cdot)}{\partial t_2}) + n_H(x_{2H}(\cdot) + t_2\frac{\partial x_{2H}(\cdot)}{\partial t_2})), \\
&\overset{(19)}{=} -\frac{\partial V_L(\cdot)}{\partial c}x_{2L}(\cdot) - (\alpha+\beta)\frac{\partial V_H(\cdot)}{\partial c}x_{2H}(\cdot) + \beta\frac{\partial V_M(\cdot)}{\partial c}x_{2M}(\cdot) + \\
&\quad \gamma(n_L(x_{2L}(\cdot) + t_2(\frac{\partial \tilde{x}_{2L}(\cdot)}{\partial t_2} - \frac{\partial \tilde{x}_{2L}(\cdot)}{\partial c}x_{2L}(\cdot))) + \\
&\quad \gamma(n_H(x_{2H}(\cdot) + t_2(\frac{\partial \tilde{x}_{2H}(\cdot)}{\partial t_2} - \frac{\partial \tilde{x}_{2H}(\cdot)}{\partial c}x_{2H}(\cdot))), \\
&\overset{(24)=0,(26)=0}{=} (-\beta\frac{\partial V_M(\cdot)}{\partial c} + \gamma n_L(t_2\frac{\partial x_{2L}(\cdot)}{\partial c} - 1))x_{2L}(\cdot) + \\
&\quad (\gamma n_H(t_2\frac{\partial x_{2H}(\cdot)}{\partial c} - 1))x_{2H}(\cdot) + \beta\frac{\partial V_M(\cdot)}{\partial c}x_{2M}(\cdot) + \\
&\quad \gamma(n_L(x_{2L}(\cdot) + t_2(\frac{\partial \tilde{x}_{2L}(\cdot)}{\partial t_2} - \frac{\partial x_{2L}(\cdot)}{\partial c}x_{2L}(\cdot))) + \\
&\quad \gamma(n_H(x_{2H}(\cdot) + t_2(\frac{\partial \tilde{x}_{2H}(\cdot)}{\partial t_2} - \frac{\partial x_{2H}(\cdot)}{\partial c}x_{2H}(\cdot))), \\
&= \beta\frac{\partial V_M(\cdot)}{\partial c}(x_{2M}(\cdot) - x_{2L}(\cdot)) + \\
&\quad \gamma t_2(n_L\frac{\partial \tilde{x}_{2L}(\cdot)}{\partial t_2} + n_H\frac{\partial \tilde{x}_{2H}(\cdot)}{\partial t_2}),
\end{aligned}$$

which must be equal to zero. The first part $(\beta\frac{\partial V_M(\cdot)}{\partial c}(x_{2M}(\cdot) - x_{2L}(\cdot)))$ is the efficiency gain (loss) of raising the tax rate for good 2 if the mimicker would consume more (less) of the good than the low type individual. The second part $(\gamma t_2(n_L\frac{\partial \tilde{x}_{2L}(\cdot)}{\partial t_2} + n_H\frac{\partial \tilde{x}_{2H}(\cdot)}{\partial t_2}))$ is the distortive effect of taxation, whose sign is inversely related to the sign of the tax rate (because the aggregate Slutsky effect $n_L\frac{\partial \tilde{x}_{2L}(\cdot)}{\partial t_2} + n_H\frac{\partial \tilde{x}_{2H}(\cdot)}{\partial t_2}$ is negative). Assuming separability and $q_2 = 0$, we have $x_{2M}(\cdot) = x_{2L}(\cdot)$ and the first part drops out. Being left with the distortion only, it is optimal not to tax good 2.[29]

Let us finally go back to the optimal queuing rate, governed by equation (29), and again evaluated at $q_2 = 0$. Plugging in the optimal zero tax rate for good 2,

[29] Because of the price normalization of good 1, it is actually better to write: it is optimal not to tax good 2 differently from good 1.

we get

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial q_2} &\overset{t_2^*=0}{=} \frac{\partial V_L(\cdot)}{\partial q_2} + (\alpha+\beta)\frac{\partial V_H(\cdot)}{\partial q_2} - \beta\frac{\partial V_M(\cdot)}{\partial q_2}, \qquad (30)\\
&\overset{(15)}{=} \frac{\partial V_L}{\partial y}(\cdot)\theta_L \eta x_{2L}(\cdot) + (\alpha+\beta)\frac{\partial V_H}{\partial y}(\cdot)\theta_H \eta x_{2H}(\cdot) - \beta\frac{\partial V_M}{\partial y}(\cdot)\theta_H \eta x_{2M}(\cdot),\\
&\overset{(25)=0,(27)=0,t_2^*=0}{=} (\beta\frac{\partial V_M(\cdot)}{\partial y} - \gamma n_L)\theta_L \eta x_{2L}(\cdot) - \gamma n_H \theta_H \eta x_{2H}(\cdot) - \beta\frac{\partial V_M}{\partial y}(\cdot)\theta_H \eta x_{2M}(\cdot),\\
&= -\beta\eta\frac{\partial V_M(\cdot)}{\partial y}(\theta_H x_{2M}(\cdot) - \theta_L x_{2L}(\cdot)) - \gamma\eta(n_L\theta_L x_{2L}(\cdot) + n_H\theta_H x_{2H}(\cdot)).
\end{aligned}$$

The first term between brackets $(-\beta\eta\frac{\partial V_M(\cdot)}{\partial y}(\theta_H x_{2M}(\cdot) - \theta_L x_{2L}(\cdot)))$ measures the efficiency gain of relaxing the self-selection constraint by introducing queuing. Because $-\beta\eta\frac{\partial V_M(\cdot)}{\partial y} > 0$, its sign depends on the sign of $\theta_H x_{2M}(\cdot) - \theta_L x_{2L}(\cdot)$. Evaluated at $q_2 = 0$ and given separability, we have $x_{2L}(\cdot) = x_{2M}(\cdot)$ and the term becomes strictly positive. The second term between brackets $-\gamma\eta(n_L\theta_L x_{2L}(\cdot) + n_H\theta_H x_{2H}(\cdot))$ measures the efficiency cost of queuing. For an individual with type i, an increase in the queuing rate corresponds with an increase of $\eta x_{2i}(\cdot)$ hours of activity which, expressed in monetary terms, correponds with a earnings loss of $\eta\theta_i x_{2i}(\cdot)$. The term $\eta(n_L\theta_L x_{2L}(\cdot) + n_H\theta_H x_{2H}(\cdot))$ measures the aggregate loss which, pre-multiplied with γ, gives an expression in welfare terms.

There is thus an efficiency gain and an efficiency loss. The proof of theorem 1, which I deal with in the next section, essentially shows that the gain is always smaller than the loss.

B Proof of theorem 1

We show that it is not optimal to introduce queuing, more precisely, we show that equation (30), evaluated at $q_2 = 0$, is negative. If earnings taxes can be optimally set, then equation (24) must be zero, which, given the optimal zero tax rate on good 2, implies

$$\frac{\partial V_L(\cdot)}{\partial c} - \beta\frac{\partial V_M(\cdot)}{\partial c} = \gamma n_L.$$

This equation can be rewritten as

$$(1-\beta)\frac{\partial V_M(\cdot)}{\partial c} = \left(\frac{\partial V_M(\cdot)}{\partial c} - \frac{\partial V_L(\cdot)}{\partial c}\right) + \gamma n_L. \qquad (31)$$

Given $t_2^* = 0$, $q_2 = 0$, and separability (between consumption and activity), we have $x_{2L}(\cdot) = x_{2M}(\cdot)$. Because the low type and the mimicker face the same budget set, also $x_{1L}(\cdot) = x_{1M}(\cdot)$ holds. But then, given equation (1), (12), and separability, the marginal (indirect) utility of net income is the same for the mimicker and the low type and the first term (between brackets) on the right-hand side of equation (31) drops out. Because the right-hand side is unambiguously strictly positive, we must have that $(1-\beta) > 0$ holds.

Equation (30), evaluated at $q_2 = 0$, can be rewritten as

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial q_2}|_{q_2=0} &= (1-\beta)\frac{\partial V_L(\cdot)}{\partial q_2} + (\alpha+\beta)\frac{\partial V_H(\cdot)}{\partial q_2} + \beta\left(\frac{\partial V_L(\cdot)}{\partial q_2} - \frac{\partial V_M(\cdot)}{\partial q_2}\right), \\ &= \underbrace{(1-\beta)\frac{\partial V_L(\cdot)}{\partial q_2}}_{<0} + \underbrace{(\alpha+\beta)\frac{\partial V_H(\cdot)}{\partial q_2}}_{<0} + \beta\eta\left(\frac{\partial V_L}{\partial y}(\cdot)\theta_L x_{2L}(\cdot) - \frac{\partial V_M}{\partial y}(\cdot)\theta_H x_{2M}(\cdot)\right).\end{aligned}$$

The first two terms are negative. The last term is negative too because (given $t_2^* = 0$, $q_2 = 0$, and separability) we have $x_{2L}(\cdot) = x_{2M}(\cdot)$ and, given additive separability—i.e., $U(x_1, x_2, a) = u(x_1, x_2) - v(a)$ everywhere with $\frac{dv}{da} > 0$ and $\frac{d^2v}{da^2} > 0$—we also have that (at $q_2 = 0$)

$$\frac{\partial V_L}{\partial y}(\cdot)\theta_L = -\frac{dv}{da}(\frac{y_L}{\theta_L}) < -\frac{dv}{da}(\frac{y_L}{\theta_H}) = \frac{\partial V_M}{\partial y}(\cdot)\theta_H,$$

as required.

C Proof of proposition 1

Suppose that the queuing rate q_2 is strictly positive and not too large such that $x_{2H}(\cdot) > x_{2M}(\cdot)$ at the current tax scheme. The proof proceeds in three steps. To explain the logic of the different steps, I provide a brief outline.

In the first step, I show that if $q_2 > 0$ and $t_2 < 0$ would hold, then we can introduce a Pareto-improving tax reform (no type is worse off and some type is strictly better off) that is feasible (that respects the self-selection and the budget constraint). This reform does not change the queuing rate q_2, but increases the tax rate t_2. As a consequence, while the combination $q_2 > 0$ and $t_2 < 0$ is not possible, $q_2 > 0$ and $t_2 \geq 0$ can still be optimal.

In the second step, I therefore start from $q_2 > 0$ and $t_2 \geq 0$ and show that it is again possible to introduce a feasible Pareto-improving reform that is now based on decreasing the queuing rate, but leaving the tax rate unchanged. So, both steps together, it must be that $q_2 = 0$ at the optimum. Now, given that there is no queuing at the optimum, Laroque () shows that also $t_2 = 0$ must hold. Let us now proof steps 1 and 2.

Suppose $q_2 > 0$ and $t_2 < 0$ hold (step 1). Consider the following tax reform:

a small increase in the tax rate $\Delta t_2 > 0$, together with a compensation for both types in terms of consumption, equal to $\Delta c_L = \epsilon + x_{2L}\Delta t_2 > 0$ (with $\epsilon > 0$) and $\Delta c_H = x_{2H}\Delta t_2 > 0$. This reform is a Pareto improvement because

$$\begin{aligned} \Delta V_L &\cong \frac{\partial V_L}{\partial t_2}\Delta t_2 + \frac{\partial V_L}{\partial c}\Delta c_L \overset{(14)}{=} \frac{\partial V_L}{\partial c}\epsilon > 0, \\ \Delta V_H &\cong \frac{\partial V_H}{\partial t_2}\Delta t_2 + \frac{\partial V_H}{\partial c}\Delta c_H \overset{(14)}{=} 0. \end{aligned}$$

Moreover, if we choose ϵ not too large ($\epsilon \leq (x_{2M} - x_{2L})\Delta t_2$),[30] then the reform relaxes the self-selection constraint (i.e., $\Delta V_H - \Delta V_M = -\Delta V_M \geq 0$), because

$$\Delta V_M \cong \frac{\partial V_M}{\partial t_2}\Delta t_2 + \frac{\partial V_M}{\partial c}\Delta c_L \overset{(14)}{=} -\frac{\partial V_M}{\partial c}((x_{2M} - x_{2L})\Delta t_2 - \epsilon) \leq 0.$$

Finally, the reform also relaxes the budget constraint, because

$$\begin{aligned} \Delta R &\cong \frac{\partial R}{\partial t_2}\Delta t_2 + \frac{\partial R}{\partial c_L}\Delta c_L + \frac{\partial R}{\partial c_H}\Delta c_H, \\ &\overset{(19),\Delta c_i}{\cong} t_2(n_L\frac{\partial \tilde{x}_{2L}}{\partial t_2} + n_H\frac{\partial \tilde{x}_{2H}}{\partial t_2})\Delta t_2 + n_L(t_2\frac{\partial x_{2L}}{\partial c} - 1)\epsilon. \end{aligned}$$

Because the first term is strictly positive (given $t_2 < 0$, $\frac{\partial \tilde{x}_{2i}}{\partial t_2} < 0$, and $\Delta t_2 > 0$), we can always choose ϵ within its bounds ($0 < \epsilon \leq (x_{2M} - x_{2L})\Delta t_2$) such that the total effect on revenues ΔR remains positive.

Suppose $q_2 > 0$ and $t_2 \geq 0$ hold (step 2). Consider the following reform: A decreasing in the queuing rate $\Delta q_2 < 0$ and an increase in gross earnings of the low ability type $\Delta y_L = -\theta_L \eta x_{2L}\Delta q_2 > 0$. This reform is again a Pareto improvement because

$$\begin{aligned} \Delta V_L &\cong \frac{\partial V_L}{\partial q_2}\Delta q_2 + \frac{\partial V_L}{\partial y}\Delta y_L \overset{(15)}{=} 0, \\ \Delta V_H &\cong \frac{\partial V_H}{\partial q_2}\Delta q_2 > 0. \end{aligned}$$

[30]Given $\epsilon > 0$, this choice is only possible if the right-hand side is strictly positive. This is true because $\Delta t_2 > 0$ (by assumption) and $x_{2M} - x_{2L} > 0$ follows from equation (6) (given separability and $q_2 > 0$).

The reform relaxes the self-selection constraint because

$$\begin{aligned}
\Delta V_H - \Delta V_M \quad &\cong \quad \frac{\partial V_H}{\partial q_2}\Delta q_2 - (\frac{\partial V_M}{\partial q_2}\Delta q_2 + \frac{\partial V_M}{\partial y}\Delta y_L), \\
&\overset{(15),\Delta y_L}{\cong} \quad \frac{\partial V_H}{\partial y}\theta_H \eta x_{2H}\Delta q_2 - (\frac{\partial V_M}{\partial y}\theta_H \eta x_{2M}\Delta q_2 - \frac{\partial V_M}{\partial y}\theta_L \eta x_{2L}\Delta q_2), \\
&\cong \quad \frac{\partial V_H}{\partial y}\Delta q_2 \eta \theta_H x_{2H} - \frac{\partial V_M}{\partial y}\Delta q_2 \eta(\theta_H x_{2M} - \theta_L x_{2L}).
\end{aligned}$$

To see that the sign of the right-hand side is positive, note first that

$$\frac{\partial V_H}{\partial y}\Delta q_2 \eta > \frac{\partial V_M}{\partial y}\Delta q_2 \eta > 0.$$

Given separability, this is true if

$$-\frac{\partial V_H}{\partial y} = \frac{\partial v}{\partial a}(\eta q_2 x_{2H} + \frac{y_H}{\theta_H})\frac{1}{\theta_H} > -\frac{\partial v}{\partial a}(\eta q_2 x_{2M} + \frac{y_L}{\theta_H})\frac{1}{\theta_H} = -\frac{\partial V_M}{\partial y}.$$

To see this, note that self-selection requires $y_H > y_L$ (and $c_H > c_L$), which, together with the assumption $x_{2H} > x_{2M}$ implies $\eta q_2 x_{2H} + \frac{y_H}{\theta_H} > \eta q_2 x_{2M} + \frac{y_L}{\theta_H}$. In addition, because $x_{2H} > x_{2M} > x_{2L}$ holds, we also have

$$\theta_H x_{2H} > \theta_H x_{2M} - \theta_L x_{2L} > 0.$$

Finally, the reform also relaxes the budget constraint. We indeed have

$$\begin{aligned}
\Delta R \quad &\cong \quad \frac{\partial R}{\partial q_2}\Delta q_2 + \frac{\partial R}{\partial y_L}\Delta y_L, \\
&\cong \quad t_2(n_L \frac{\partial x_{2L}}{\partial q_2} + n_H \frac{\partial x_{2H}}{\partial q_2})\Delta q_2 + n_L(t_2 \frac{\partial x_{2L}}{\partial y} + 1)\Delta y_L, \\
&\cong \quad t_2 n_L(\frac{\partial x_{2L}}{\partial q_2}\Delta q_2 + \frac{\partial x_{2L}}{\partial y}\Delta y_L) + t_2 n_H \frac{\partial x_{2H}}{\partial q_2}\Delta q_2 + n_L \Delta y_L.
\end{aligned}$$

As t_2 and Δy_L are positive, it suffices to show that the changed consumption of the queuing good, being,

$$\begin{aligned}
\Delta x_{2L} \quad &= \quad \frac{\partial x_{2L}}{\partial q_2}\Delta q_2 + \frac{\partial x_{2L}}{\partial y}\Delta y_L, \\
\Delta x_{2H} \quad &= \quad \frac{\partial x_{2H}}{\partial q_2}\Delta q_2,
\end{aligned}$$

are both positive. For the high type, it is obvious because both factors are negative (see equation (4) and Δq_2 by assumption). For the low type, use equations (4) and (5) together with the definition of Δy_L, to rewrite the change in consumption as

$$\begin{aligned}
\Delta x_{2L} &= -\frac{\frac{d^2v}{da^2}(\cdot)\eta^2 q_2 x_{2L}(\cdot) + \frac{dv}{da}(\cdot)\eta}{-SOC_2(\cdot)}\Delta q_2 + \frac{\frac{d^2v}{da^2}(\cdot)\frac{1}{\theta_L}\eta q_2}{-SOC_2(\cdot)}\theta_L \eta x_{2L}\Delta q_2, \\
& \frac{-\frac{dv}{da}(\cdot)\eta}{-SOC_2(\cdot)}\Delta q_2 > 0.
\end{aligned}$$

D Proof of proposition 2

Suppose that good 2 is in fixed supply, say, S. I set the unit production price equal to zero and I take the queuing rate q_2 as the policy instrument, and the tax rate $t_2(q_2)$ adjusts to make demand and supply equal, so it is implicitly defined as

$$n_L x_2(t_2(q_2), q_2, c_L, y_L, \theta_L) + n_H x_2(t_2(q_2), q_2, c_H, y_H, \theta_H) = S.$$

For later use, this implies that $t_2'(q_2) < 0$ because differentiation leads to

$$\left(n_L \frac{\partial x_{2L}(\cdot)}{\partial t_2} + n_H \frac{\partial x_{2H}(\cdot)}{\partial t_2}\right) t_2'(q_2) + n_L \frac{\partial x_{2L}(\cdot)}{\partial q_2} + n_H \frac{\partial x_{2H}(\cdot)}{\partial q_2} = 0.$$

The problem of the planner is

$$\max_{q_2, c_L, y_L, c_H, y_H} V(t_2(q_2), q_2, c_L, y_L, \theta_L)$$

subject to

$$V(t_2(q_2), q_2, c_H, y_H, \theta_H) - \underline{V} \geq 0, \tag{32}$$

$$V(t_2(q_2), q_2, c_H, y_H, \theta_H) - V(t_2(q_2), q_2, c_L, y_L, \theta_H) \geq 0, \tag{33}$$

and

$$n_L(y_L - c_L) + n_H(y_H - c_H) + t_2(q_2)S - \underline{R} \geq 0. \tag{34}$$

The first-order condition with respect to consumption require

$$\begin{aligned}
\frac{\partial V_L(\cdot)}{\partial c} &= \beta \frac{\partial V_M(\cdot)}{\partial c} + \gamma n_L, \\
(\alpha + \beta)\frac{\partial V_H(\cdot)}{\partial c} &= \gamma n_H.
\end{aligned}$$

The first-order condition with respect to the queuing rate is

$$\frac{\partial A}{\partial t_2}(\cdot) t_2'(q_2) + \frac{\partial B}{\partial q_2}(\cdot),$$

with

$$
\begin{aligned}
\frac{\partial A}{\partial t_2}(\cdot) &= \frac{\partial V_L}{\partial t_2}(\cdot) + (\alpha+\beta)\frac{\partial V_H}{\partial t_2}(\cdot) - \beta\frac{\partial V_M}{\partial t_2}(\cdot) + \gamma S, \\
&= -\frac{\partial V_L}{\partial c}(\cdot)x_{2L}(\cdot) - (\alpha+\beta)\frac{\partial V_H}{\partial c}(\cdot)x_{2H}(\cdot) + \beta\frac{\partial V_M}{\partial c}(\cdot)x_{2M}(\cdot) + \gamma S, \\
&= \beta\frac{\partial V_M}{\partial c}(\cdot)(x_{2M}(\cdot) - x_{2L}(\cdot)) \geq 0, \\
\frac{\partial B}{\partial q_2}(\cdot) &= \frac{\partial V_L}{\partial q_2}(\cdot) + (\alpha+\beta)\frac{\partial V_H}{\partial q_2}(\cdot) - \beta\frac{\partial V_M}{\partial q_2}(\cdot), \\
&= (1-\beta)\frac{\partial V_L}{\partial q_2}(\cdot) + (\alpha+\beta)\frac{\partial V_H}{\partial q_2}(\cdot) - \beta(\frac{\partial V_M}{\partial q_2}(\cdot) - \frac{\partial V_L}{\partial q_2}(\cdot)).
\end{aligned}
$$

Given separability and $q_2 = 0$ it is still true that $(1-\beta) > 0$ and $\frac{\partial V_M}{\partial q_2}(\cdot) > \frac{\partial V_L}{\partial q_2}(\cdot)$ hold, so $\frac{\partial B}{\partial q_2}(\cdot) < 0$. And given $q_2 = 0$ and separability, we also have $\frac{\partial A}{\partial t_2}(\cdot) = 0$ (or, even more generally, we have $\frac{\partial A}{\partial t_2}(\cdot)t_2'(q_2) \leq 0$ because $\frac{\partial A}{\partial t_2}(\cdot) \geq 0$ and $t_2'(q_2) < 0$). So, introducing queuing has a negative welfare impact, as required.

E Proof of proposition 3

While the utility function remains additively separable, assume now that queuing and labour are no longer perfect substitutes, i.e., utility is given by

$$
U(x_1, x_2, q_2x_2, y/\theta) = u(x_1, x_2) - v(q_2x_2, y/\theta).
$$

In this case, equation (15) becomes

$$
\frac{\partial V}{\partial q_2}(t_2, q_2, c, y, \theta) = -\frac{\partial v}{\partial q}(q_2x_2, \frac{y}{\theta})x_2(t_2, q_2, c, y, \theta), \tag{35}
$$

Additive separability still guarantees that, given $q_2 = 0$, it is optimal to neither tax nor subsidize the queuing good (so, $t_2^* = 0$). Given $t_2^* = 0$, $q_2 = 0$, and separability, also $(1-\beta) > 0$ remains to be true at the optimum (see proof of Theorem 1). Finally, equation (30), evaluated at $q_2 = 0$, can now be rewritten as

$$
\begin{aligned}
\frac{\partial \mathscr{L}}{\partial q_2}|_{q_2=0} &= (1-\beta)\frac{\partial V_L(\cdot)}{\partial q_2} + (\alpha+\beta)\frac{\partial V_H(\cdot)}{\partial q_2} + \beta\left(\frac{\partial V_L(\cdot)}{\partial q_2} - \frac{\partial V_M(\cdot)}{\partial q_2}\right), \\
&= \underbrace{(1-\beta)\frac{\partial V_L(\cdot)}{\partial q_2}}_{<0} + \underbrace{(\alpha+\beta)\frac{\partial V_H(\cdot)}{\partial q_2}}_{<0} + \beta\left(\frac{\partial v}{\partial q}(q_2x_{2L}(\cdot), \frac{y_L}{\theta_H}) - \frac{\partial v}{\partial q}(q_2x_{2L}(\cdot), \frac{y_L}{\theta_L})\right)x_{2L}(\cdot),
\end{aligned}
$$

using $x_{2L}(\cdot) = x_{2M}(\cdot)$ in the last step. For this expression to be negative, we must have $\frac{\partial v}{\partial q}(q_2 x_{2L}(\cdot), \frac{y_L}{\theta_H}) \leq \frac{\partial v}{\partial q}(q_2 x_{2L}(\cdot), \frac{y_L}{\theta_L})$. So, $\frac{\partial v^2}{\partial q \partial \ell} \geq 0$ suffices, or, in words, it suffices that queuing and labour are substitutes (but not necessarily perfect substitutes): extra queuing is equally or more unpleasant at higher levels of labour.

F Proof of proposition 4

Suppose that the consumption level of good 2 can be recorded such that the tax and the queuing scheme can both be a general function of earnings and the consumption level of the queuing good (being good 2; good 1 remains untaxed). In other words, the tax and queuing time/effort are $T(x_2, y)$ and $Q(x_2, y)$ with obvious notation. Suppose in addition that separability between consumption and activity does not hold and that labour and queuing are not necessarily perfect substitutes; in other words, utility is a general function of consumption, (total) queuing, and labour, denoted $U(x_1, x_2, q, \frac{y}{\theta})$.

It implies that we can make it impossible for the mimicker to mimick the gross earnings of the low type (i.e., y_L), but to choose a consumption level that is different from the consumption level of the low type (i.e., x_{2L}).[31] Society can therefore directly set the consumption level of both goods for both types as well as the (total) queuing time/effort ($q_L = Q(x_{2L}, y_L)$ and $q_H = Q(x_{2H}, y_H)$); thus, the societal program is

$$\max_{\{x_{1i}, x_{2i}, q_i, y_i\}_{i=L,H}} U(x_{1L}, x_{2L}, q_L, \frac{y_L}{\theta_L})$$

subject to a minimal (exogenous) utility level $\underline{U}$ reserved for the high ability types

$$U(x_{1H}, x_{2H}, q_H, \frac{y_H}{\theta_H}) - \underline{U} \geq 0, \tag{36}$$

subject to the self-selection constraint (to guarantee that high ability types do not mimick low ability types)

$$U(x_{1H}, x_{2H}, q_H, \frac{y_H}{\theta_H}) - U(x_{1L}, x_{2L}, q_L, \frac{y_L}{\theta_H}) \geq 0, \tag{37}$$

and subject to a budget constraint (to guarantee that the total tax revenues—the difference between gross earnings and expenditures—can finance some exogenous

[31] In other words, combinations (x_2, y) that differ from (x_{2L}, y_L) and (x_{2H}, y_H) are infinitely taxed and/or queued for forever such that they are never chosen.

revenue requirement $\underline{R}$)

$$n_L(y_L - x_{1L} - p_2 x_{2L}) + n_H(y_H - x_{1H} - p_2 x_{2H}) - \underline{R} \geq 0. \qquad (38)$$

Again, we use $\mathscr{L}$ to denote the Langrange function and α, β and γ as the Lagrange multipliers.

We focus on the optimal amount of queuing for both types. As usual, the high ability type remains undistorted. In the current setting it also implies that no queuing is required (i.e., the corner solution $q_H = 0$ follows from $(\alpha + \beta)\frac{\partial U}{\partial q}(x_{1H}, x_{2H}, q_H, \frac{y_H}{\theta_H}) < 0$). For the low ability type we derive the Lagrange function with respect to consumption and queuing:

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial x_1^L} &= \frac{\partial U_L(\cdot)}{\partial x_1} - \beta \frac{\partial U_M(\cdot)}{\partial x_1} - \gamma n_L, \\
\frac{\partial \mathscr{L}}{\partial x_2^L} &= \frac{\partial U_L(\cdot)}{\partial x_2} - \beta \frac{\partial U_M(\cdot)}{\partial x_2} - \gamma n_L p_2, \\
\frac{\partial \mathscr{L}}{\partial q_L} &= \frac{\partial U_L(\cdot)}{\partial q_L} - \beta \frac{\partial U_M(\cdot)}{\partial q_L}.
\end{aligned}$$

The first two derivatives must equal zero, leading to

$$\begin{aligned}
(1-\beta)\frac{\partial U_M(\cdot)}{\partial x_1} &= \gamma n_L + \left(\frac{\partial U_M(\cdot)}{\partial x_1} - \frac{\partial U_L(\cdot)}{\partial x_1} \right), \\
(1-\beta)\frac{\partial U_M(\cdot)}{\partial x_2} &= \gamma n_L p_2 + \left(\frac{\partial U_M(\cdot)}{\partial x_2} - \frac{\partial U_L(\cdot)}{\partial x_2} \right).
\end{aligned}$$

The signs of the terms between brackets (at the right-hand side of each equation) depend on the signs of $\frac{\partial^2 U}{\partial x_1 \partial \ell}$ and $\frac{\partial^2 U}{\partial x_2 \partial \ell}$. Indeed, $\frac{\partial U(x_{1L}, x_{2L}, q_L, \frac{y_L}{\theta_H})}{\partial x_i} - \frac{\partial U(x_{1L}, x_{2L}, q_L, \frac{y_L}{\theta_L})}{\partial x_i}$ will be positive (negative) if extra consumption of good i is less (more) pleasant if one works harder. If $\frac{\partial^2 U}{\partial x_i \partial \ell} \leq 0$ holds for one of the two goods, then $1 - \beta > 0$ follows. Let $MRS(x_1, x_2, q, \ell) = -\frac{\frac{\partial U(x_1, x_2, q, \ell)}{\partial x_1}}{\frac{\partial U(x_1, x_2, q, \ell)}{\partial x_2}}$ be the marginal rate of substitution of good 2 for good 1. The sign of the derivative of the marginal rate of substitution with respect to labour ($\frac{\partial MRS(x_1, x_2, q, \ell)}{\partial \ell}$) depends on the sign of

$$\frac{\partial^2 U(x_1, x_2, q, \ell)}{\partial x_1 \partial \ell} \frac{\partial U(x_1, x_2, q, \ell)}{\partial x_2} - \frac{\partial^2 U(x_1, x_2, q, \ell)}{\partial x_2 \partial \ell} \frac{\partial U(x_1, x_2, q, \ell)}{\partial x_1},$$

which implies that, for the sign of $\frac{\partial MRS(x_1, x_2, q, \ell)}{\partial \ell}$ to be positive everywhere (good 2 is complementary to labour, i.e., this good will be consumed more if someone works harder/more (for the same net income)) or negative everywhere (good 2 is complementary to leisure, i.e., this good will be consumed more if someone works less (hard) for the same net income), the signs of the cross-derivatives $\frac{\partial^2 U(x_1,x_2,q,\ell)}{\partial x_i \partial \ell}$ must be opposite. To sum up, if good 2 is complementary to labour or if good 2 is complementary to leisure, then $\frac{\partial^2 U}{\partial x_i \partial \ell} \leq 0$ must hold for one of the goods.

The last derivative can be rewritten as

$$\frac{\partial \mathscr{L}}{\partial q_L} = (1-\beta)\frac{\partial U_M(\cdot)}{\partial q_L} + (\frac{\partial U_L(\cdot)}{\partial q_L} - \frac{\partial U_M(\cdot)}{\partial q_L}),$$

and consists of two terms. The sign of the first term $((1-\beta)\frac{\partial U_M(\cdot)}{\partial q_L})$ simply depends on the sign of $1-\beta$, which we discussed before. The sign of the second term (being $\frac{\partial U(x_{1L},x_{2L},q_L,\frac{y_L}{\theta_L})}{\partial q_L} - \frac{\partial U(x_{1L},x_{2L},q_L,\frac{y_L}{\theta_H})}{\partial q_L}$ in full) depends on the sign of the cross-derivative $\frac{\partial^2 U}{\partial q \partial \ell}$: if $\frac{\partial^2 U}{\partial q \partial \ell} \leq 0$ holds then the marginal disutility of queuing is less negative for the mimicker (as $\frac{y_L}{\theta_H} < \frac{y_L}{\theta_L}$) and the last term is negative.

To summarize, if at least one good is more enjoyable if working less hard ($\frac{\partial^2 U}{\partial x_i \partial \ell} \leq 0$ holds for some good i) and if queing is less unpleasant if working less hard ($\frac{\partial^2 U}{\partial q \partial \ell} \leq 0$) then it is not optimal to introduce queuing for the low ability individuals.

G Proof of proposition 5

To analyze tastes or needs differences, we slightly change the model, following the lines of Cremer, Pestieau, and Rochet (2001). While the analysis is compatible with both interpretations, it is somewhat less controversial to assume that society wants to redistribute from low to high needs. I replace ability by a needs level, again denoted by θ, for the queuing good 2 where a high value stands for high needs for good 2. Contrary to before, redistribution now goes from low to high needs. Because individuals do not differ in ability and given additive separability, there is no reason to distort labour supply. So, we introduce a lump-sum tax T on income. A rational individual chooses consumption and earnings to solve

$$\max_{x_1,x_2,y} U(x_1, x_2-\theta, \eta q_2 x_2 + y)$$

subject to his or her budget constraint

$$x_1 + (p_2+t_2)x_2 \leq y - T.$$

Assuming an interior solution, the first-order conditions are

$$\frac{\partial U}{\partial x_1}(x_1, x_2 - \theta, \eta q_2 x_2 + y) = \lambda, \tag{39}$$

$$\frac{\partial U}{\partial x_2}(x_1, x_2 - \theta, \eta q_2 x_2 + y) + \frac{\partial U}{\partial a}(x_1, x_2 - \theta, \eta q_2 x_2 + y)\eta q_2 = \lambda(p_2 + t_2), \tag{40}$$

$$\frac{\partial U}{\partial a}(x_1, x_2 - \theta, \eta q_2 x_2 + y) = -\lambda, \tag{41}$$

$$x_1 + (p_2 + t_2)x_2 = y - T, \tag{42}$$

with $\lambda > 0$ the Lagrange multiplier. Let $x_1(t_2, q_2, T, \theta)$, $x_2(t_2, q_2, T, \theta)$ and $y(t_2, q_2, T, \theta)$ denote the solution. Using these equations, we can use the last equation to eliminate y and the solution for the goods must satisfy

$$\frac{\partial U}{\partial x_2}(x_1, x_2 - \theta, x_1 + (\eta q_2 + p_2 + t_2)x_2 + T) - (\eta q_2 + p_2 + t_2)\frac{\partial U}{\partial x_1}(x_1, x_2 - \theta, x_1 + (\eta q_2 + p_2 + t_2)x_2 + T)$$

$$\frac{\partial U}{\partial a}(x_1, x_2 - \theta, x_1 + (\eta q_2 + p_2 + t_2)x_2 + T) + \frac{\partial U}{\partial x_1}(x_1, x_2 - \theta, x_1 + (\eta q_2 + p_2 + t_2)x_2 + T)$$

The solution satisfies this system of equations (and therefore the solution does not change) if one changes q_2 and t_2 simultaneously with $\triangle q_2$ and $\triangle t_2$ such that $\eta \triangle q_2 + \triangle t_2 = 0$, while keeping T constant. Therefore also (indirect) utility will not change, irrespective of type. But the revenues do change. The revenues are

$$n_L(t_2 x_2(t_2, q_2, T, \theta_L) + T) + n_H(t_2 x_2(t_2, q_2, T, \theta_H) + T).$$

As the quantities do not change by the previous operation, the revenues will change with

$$(n_L x_2(t_2, q_2, T, \theta_L) + n_H x_2(t_2, q_2, T, \theta_H))\triangle t_2.$$

So, if the planner reduces queuing and increases taxes (or decreases subsidies), then no one's utility changes, but revenues increase. As there is no self-selection constraint required—the lump-sum tax implies that it is satisfied—I conclude that it is not useful to introduce queuing.

H Proof of proposition 6

Suppose that the planner can provide two separate queuing schemes, one for each ability type. Suppose that bundling is possible, i.e., the queuing scheme depends on the earnings chosen. In practice, it means that if you high earnings, then you get access to the good in a different way, i.e., with a different queuing rate and tax rate. The societal problem is to choose the extended set of policy instruments to maximize the utility level of the low ability individuals

$$\max_{t_{2L},q_{2L},c_L,y_L,t_{2H},q_{2H},c_H,y_H} V(t_{2L},q_{2L},c_L,y_L,\theta_L)$$

subject to the following three constraints

$$V(t_{2H},q_{2H},c_H,y_H,\theta_H)-\underline{V}\geq 0, \tag{43}$$

$$V(t_{2H},q_{2H},c_H,y_H,\theta_H)-V(t_{2L},q_{2L},c_L,y_L,\theta_H)\geq 0, \tag{44}$$

$$n_L(t_{2L}x_2(t_{2L},q_{2L},c_L,y_L,\theta_L)+y_L-c_L)+n_H(t_{2H}x_2(t_{2H},q_{2H},c_H,y_H,\theta_H)+y_H-c_H)-\underline{R}\geq 0. \tag{45}$$

We first focus on the high ability individuals. Using α, β, and γ as the Lagrange multipliers, the partial derivative of the Lagrangian with respect to the tax rate t_{2H} is

$$(\alpha+\beta)\frac{\partial V_H}{\partial t_{2H}}(\cdot)+\gamma(n_H x_{2H}(\cdot)+n_H t_{2H}\frac{\partial x_{2H}}{\partial t_{2H}}(\cdot)).$$

Using equations (14) and (19), we get

$$-(\alpha+\beta)\frac{\partial V_H}{\partial c}(\cdot)x_{2H}(\cdot)+\gamma(n_H x_{2H}(\cdot)+n_H t_{2H}(\frac{\partial \tilde{x}_{2H}}{\partial t_{2H}}(\cdot)-\frac{\partial x_{2H}}{\partial c}(\cdot)x_{2H}(\cdot)). \tag{46}$$

Because equation (26) must also be zero in this set-up, we must have

$$-(\alpha+\beta)\frac{\partial V_H(\cdot)}{\partial c}=\gamma n_H(t_{2H}\frac{\partial x_{2H}(\cdot)}{\partial c}-1),$$

which we can use to rewrite (46) as

$$n_H t_{2H}\frac{\partial \tilde{x}_{2H}}{\partial t_{2H}}(\cdot).$$

Because the first-order condition requires this derivative to be equal to zero, the only solution is $t_{2H}=0$.

Similarly, the partial derivative with respect to the queuing rate q_{2H} is

$$(\alpha+\beta)\frac{\partial V_H}{\partial q_{2H}}(\cdot)+\gamma n_H t_{2H}\frac{\partial x_H}{\partial q_{2H}}(\cdot),$$

which, using $t_{2H}=0$ and equation (15) reduces to $(\alpha+\beta)\theta_H\eta\frac{\partial V_H}{\partial y}(\cdot)x_{2H}(\cdot)$, which is strictly negative. So, the corner solution $q_{2H}=0$ is optimal. Both results together confirm the usual "no distortion at the top"-result: high ability types pay taxes, but are not distorted at the (labour and consumption) margin.

We now focus on the low ability types. The partial derivative of the Lagrangian with respect to the tax rate t_{2L} is

$$\frac{\partial V_L}{\partial t_{2L}}(\cdot) - \beta\frac{\partial V_M}{\partial t_{2L}}(\cdot) + \gamma(n_L x_{2L}(\cdot) + n_L t_{2L}\frac{\partial x_{2L}}{\partial t_{2L}}(\cdot)).$$

Using equations (14) and (19), we get

$$-\frac{\partial V_L}{\partial c}(\cdot)x_{2L}(\cdot)+\beta\frac{\partial V_M}{\partial c}(\cdot)x_{2M}(\cdot)+\gamma(n_L x_{2L}(\cdot)+n_L t_{2L}(\frac{\partial \tilde{x}_{2L}}{\partial t_{2L}}(\cdot)-\frac{\partial x_{2L}}{\partial c}(\cdot)x_{2L}(\cdot))). \tag{47}$$

Because equation (24) must also be zero in this set-up, we can plug in

$$-\frac{\partial V_L}{\partial c}(\cdot) = -\beta\frac{\partial V_M}{\partial c}(\cdot) + \gamma n_L(t_{2L}\frac{\partial x_{2L}}{\partial c}(\cdot) - 1),$$

in (47), to obtain

$$\beta\frac{\partial V_M}{\partial c}(\cdot)(x_{2M}(\cdot) - x_{2L}(\cdot)) + \gamma n_L t_{2L}\frac{\partial \tilde{x}_{2L}}{\partial t_{2L}}(\cdot).$$

Evaluated at $q_{2L} = 0$, we have $x_{2M}(\cdot) = x_{2L}(\cdot)$ for a separable utility function; then, the first-order condition reduces to $\gamma n_L t_{2L}\frac{\partial \tilde{x}_{2L}}{\partial t_{2L}}(\cdot) = 0$, which implies that $t_{2L} = 0$ must hold.

Similarly, the partial derivative of the Langrangian with respect to the queuing rate q_{2L} is

$$\frac{\partial V_L}{\partial q_{2L}}(\cdot) - \beta\frac{\partial V_M}{\partial q_{2L}}(\cdot) + \gamma n_L t_{2L}\frac{\partial x_{2L}}{\partial q_{2L}}(\cdot).$$

We can use equation (15) to get

$$\theta_L\eta\frac{\partial V_L}{\partial y}(\cdot)x_{2L}(\cdot) - \beta\theta_H\eta\frac{\partial V_M}{\partial y}(\cdot)x_{2M}(\cdot) + \gamma n_L t_{2L}\frac{\partial x_{2L}}{\partial q_{2L}}(\cdot).$$

Because now equation (13) must be equal to zero, we can replace $\frac{\partial V_L}{\partial y}(\cdot)$ by $\beta\frac{\partial V_M(\cdot)}{\partial y} - \gamma n_L(t_{2L}\frac{\partial x_{2L}(\cdot)}{\partial y} + 1)$, leading to

$$\beta\theta_L\eta\frac{\partial V_M(\cdot)}{\partial y}x_{2L}(\cdot)-\beta\theta_H\eta\frac{\partial V_M}{\partial y}(\cdot)x_{2M}(\cdot)-\theta_L\eta\gamma n_L t_{2L}\frac{\partial x_{2L}(\cdot)}{\partial y}x_{2L}(\cdot)-\theta_L\eta\gamma n_L x_{2L}(\cdot)+\gamma n_L t_{2L}\frac{\partial x_{2L}}{\partial q_{2L}}(\cdot).$$

Evaluated at $q_{2L} = 0$, we must have $x_{2M}(\cdot) = x_{2L}(\cdot)$ and $t_{2L} = 0$, so the expression reduces to

$$-\beta\eta\frac{\partial V_M(\cdot)}{\partial y}(\theta_H - \theta_L)x_{2L}(\cdot) - \eta\gamma n_L\theta_L x_{2L}(\cdot).$$

So, because the first term is always strictly positive, we find that introducing queuing for the low ability types can be optimal if θ_L is sufficiently low.

I Proof of proposition 7

There are different ways to introduce a concern for merit goods and/or specific egalitarianism. One way to do so is to measure indirect well-being as $V(t_2, q_2, c_i, y_i, \theta_i) + \delta x_2(t_2, q_2, c_i, y_i, \theta_i)$ for type i individuals. Choosing $\delta > 0$ implies that good 2 is a merit good for both types and inequality is not only about inequality in utilities, but also about inequalities in the quantities of good 2.[32]

The societal problem is now to choose the policy instruments to maximize income of the low ability individuals

$$\max_{t_2, q_2, c_L, y_L, c_H, y_H} V(t_2, q_2, c_L, y_L, \theta_L) + \delta x_2(t_2, q_2, c_L, y_L, \theta_L)$$

subject to a minimal (exogenous) utility level $\underline{V}$ reserved for the high ability types

$$V(t_2, q_2, c_H, y_H, \theta_H) + \delta x_2(t_2, q_2, c_H, y_H, \theta_H) - \underline{V} \geq 0, \tag{48}$$

subject to the self-selection constraint (to guarantee that high ability types do not mimick low ability types, which in turn guarantees that the earnings can be implemented via an earnings tax scheme),

$$V(t_2, q_2, c_H, y_H, \theta_H) - V(t_2, q_2, c_L, y_L, \theta_H) \geq 0, \tag{49}$$

and subject to a budget constraint (to guarantee that the total tax revenues can finance some exogenous revenue requirement $\underline{R}$),

$$n_L(t_2 x_2(t_2, q_2, c_L, y_L, \theta_L) + y_L - c_L) + n_H(t_2 x_2(t_2, q_2, c_H, y_H, \theta_H) + y_H - c_H) - \underline{R} \geq 0. \tag{50}$$

In particular, note that the extra term in the well-being specification does not enter the self-selection constraint, because individual behaviour is driven by utilities only.

Again, we use $\mathscr{L}$ to denote the Langrange function and α, β and γ as the Lagrange multipliers. The derivatives of the Langrange function now become

[32] A more explicit way to introduce inequality would be to drop the minimal utility constraint for the high ability types and to maximize a welfare function W which is strictly increasing in the (extended) well-being levels.

$$
\begin{aligned}
\frac{\partial \mathscr{L}}{\partial c_L} &= \frac{\partial V_L(\cdot)}{\partial c} + \delta \frac{\partial x_{2L}(\cdot)}{\partial c} - \beta \frac{\partial V_M(\cdot)}{\partial c} + \gamma n_L (t_2 \frac{\partial x_{2L}(\cdot)}{\partial c} - 1), && (51)\\
\frac{\partial \mathscr{L}}{\partial y_L} &= \frac{\partial V_L(\cdot)}{\partial y} + \delta \frac{\partial x_{2L}(\cdot)}{\partial y} - \beta \frac{\partial V_M(\cdot)}{\partial y} + \gamma n_L (t_2 \frac{\partial x_{2L}(\cdot)}{\partial y} + 1), && (52)\\
\frac{\partial \mathscr{L}}{\partial c_H} &= (\alpha+\beta)(\frac{\partial V_H(\cdot)}{\partial c} + \delta \frac{\partial x_{2H}(\cdot)}{\partial c}) + \gamma n_H (t_2 \frac{\partial x_{2H}(\cdot)}{\partial c} - 1), && (53)\\
\frac{\partial \mathscr{L}}{\partial y_H} &= (\alpha+\beta)(\frac{\partial V_H(\cdot)}{\partial y} + \delta \frac{\partial x_{2H}(\cdot)}{\partial y}) + \gamma n_H (t_2 \frac{\partial x_{2H}(\cdot)}{\partial y} + 1), && (54)\\
\frac{\partial \mathscr{L}}{\partial t_2} &= \frac{\partial V_L(\cdot)}{\partial t_2} + \delta \frac{\partial x_{2L}(\cdot)}{\partial t_2} + (\alpha+\beta)(\frac{\partial V_H(\cdot)}{\partial t_2} + \delta \frac{\partial x_{2H}(\cdot)}{\partial t_2}) - \beta \frac{\partial V_M(\cdot)}{\partial t_2} \\
&\quad + \gamma(n_L(x_{2L}(\cdot) + t_2 \frac{\partial x_{2L}(\cdot)}{\partial t_2}) + n_H(x_{2H}(\cdot) + t_2 \frac{\partial x_{2H}(\cdot)}{\partial t_2})), && (55)\\
\frac{\partial \mathscr{L}}{\partial q_2} &= \frac{\partial V_L(\cdot)}{\partial q_2} + \delta \frac{\partial x_{2L}(\cdot)}{\partial q_2} + (\alpha+\beta)(\frac{\partial V_H(\cdot)}{\partial q_2} + \delta \frac{\partial x_{2H}(\cdot)}{\partial q_2}) - \beta \frac{\partial V_M(\cdot)}{\partial q_2} \\
&\quad + \gamma(n_L t_2 \frac{\partial x_{2L}(\cdot)}{\partial q_2} + n_H t_2 \frac{\partial x_{2H}(\cdot)}{\partial q_2}). && (56)
\end{aligned}
$$

The derivative with respect to the tax rate can—under (additive) separability and without queuing such that $x_{2M}(\cdot) = x_{2L}(\cdot)$—be written as

$$
\delta(\frac{\partial x_{2L}(\cdot)}{\partial t_2} + (\alpha+\beta)\frac{\partial x_{2H}(\cdot)}{\partial t_2}) + \gamma t_2 (n_L \frac{\partial \tilde{x}_{2L}(\cdot)}{\partial t_2} + n_H \frac{\partial \tilde{x}_{2H}(\cdot)}{\partial t_2}).
$$

So, if good 2 is a normal good, then the derivative is negative and it is optimal to subsidize good 2 ($t_2^* < 0$). The derivative with respect to the queuing rate (evaluated at zero queuing and given an optimal negative tax rate) becomes

$$
\underbrace{(1-\beta)\frac{\partial V_L(\cdot)}{\partial q_2}}_{?} + \underbrace{(\alpha+\beta)\frac{\partial V_H(\cdot)}{\partial q_2}}_{<0} + \underbrace{\beta(\frac{\partial V_L(\cdot)}{\partial q_2} - \frac{\partial V_M(\cdot)}{\partial q_2})}_{<0} +
$$

$$
\underbrace{\delta(\frac{\partial x_{2L}(\cdot)}{\partial q_2} + (\alpha+\beta)\frac{\partial x_{2H}(\cdot)}{\partial q_2})}_{\leq 0} + \underbrace{\gamma(n_L t_2 \frac{\partial x_{2L}(\cdot)}{\partial q_2} + n_H t_2 \frac{\partial x_{2H}(\cdot)}{\partial q_2})}_{\geq 0}.
$$

The question mark comes from the fact that the first-order condition for c_L can be rewritten as

$$
(1-\beta)\frac{\partial V_M(\cdot)}{\partial c} = \underbrace{\gamma n_L (1 - t_2 \frac{\partial x_{2L}(\cdot)}{\partial c})}_{\geq 0} + \underbrace{(\frac{\partial V_M(\cdot)}{\partial c} - \frac{\partial V_L(\cdot)}{\partial c})}_{=0} - \delta \frac{\partial x_{2L}(\cdot)}{\partial c},
$$

and the last (extra) term to the right does no longer allow to derive the sign of

$1 - \beta$.