Fees in Tontines

An Chen*, Montserrat Guillen§ and Manuel Rach*

* Institute of Insurance Science, Ulm University, Helmholtzstr. 20, 89069 Ulm, Germany.
§ Dept. Econometrics, Universitat de Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain.
E–Mail: an.chen@uni-ulm.de; mguillen@ub.edu; manuel.rach@uni-ulm.de

Abstract

This paper studies the incorporation of an explicit fee in modern tontine schemes and investigates how it affects the attractiveness of these products from the point of view of both the insurer and tontine participants. We consider a single initial fee and a variable fee, where the latter can be designed to meet different liquidity needs and risk aversions. We find that the "indifference fee", the fee for the tontine at which the participants are indifferent to choosing between the tontine or an annuity, comes extremely close to that of the annuity. This presents insurers with a large spectrum of fees on which to base their charge for managing the tontine. Tontine providers and participants can agree on a fee level which makes tontines an attractive alternative to annuities for tontine participants and at the same time, allows insurers to retain a large part of the fee as a profit, given the low risk contained in a tontine.[1]

Keywords: annuity, tontine, optimal retirement products, fees

JEL: G22, J32

[1]Montserrat Guillen acknowledges the financial support given by the Spanish Ministry of Science (grant PID2019-105986GB-C21) and ICREA Academia. An Chen and Manuel Rach acknowledge the financial support given by the DFG for the research project "Zielrente: die Lösung zur alternden Gesellschaft in Deutschland" (Grant number 418318744).

1 Introduction

Tontines are financial products, where a part of the returns is obtained via the inheritance of wealth from non-surviving participants. Tontines date back to the 17th century (Weir, 1989) when they were staged as public financial instruments to raise funds. Starting with Piggott et al. (2005) and Sabin (2010), modern forms of tontines have regained popularity in recent years because they spread the risk of living a long life by combining features of a group annuity with a sort of mortality lottery (see also Valdez et al. (2006), Stamos (2008), Qiao and Sherris (2013), Hanewald et al. (2013), Donnelly et al. (2013, 2014), Milevsky and Salisbury (2015), Li and Rothschild (2020) and Bernhardt and Donnelly (2019)). However, all the existing literature overlooks the fees required to administrate modern tontines,[2] and most of the literature examines these products from the perspective of tontine participants.

Our aim is to analyze the effec on the benefits of a tontine to its participants and the insurert of paying a small emolument to the insurer managing the tontine product. We also look into different ways of organizing the payment, and evaluate both the interests of an insurer in organizing a modern tontine and the consequences on the expected utility of the tontine participants. Intuitively, if insurers or financial managers are not provided with a high enough incentive to organize the tontine, e.g. by providing them an explicit fee, they will not be particularly interested in trading the tontine product. However, if fees paid to the managers are too high, pension savers would probably think that tontines were not an attractive alternative to life annuities. By considering possible forms of administration fees in a tontine scheme, we are able to introduce a realistic framework in the debate on choosing between tontines versus annuities, and we also show that there is a viable incentive for insurers to manage a tontine. We find a large margin for benefits, which could be the most attractive incentive for launching modern tontine products to the market.

The motive for the current analysis is to assess the viability of fees payable by policyholders to service providers, i.e. insurers or financial managers, in order to manage the tontine administration. Once this is established, we can then proceed to an analysis of what can happen once these products can be underwritten. Organizing a tontine is a type of service provision, and this provision can be paid for in various ways. For example,

[2] These costs might include but not be limited to the following: (a) organizing a modern tontine, (b) managing the investments, (c) distributing dividends and (d) informing participants. Indeed, in a digitalized world, where communicating with the customer is the future of mobile banking, it suggests a significant amount of back-office and regular online service to communicate with clients.

fees can be paid at the beginning, or for a stipulated time, or until the participant dies. Inspired by the literature on variable annuities (e.g. MacKay et al. (2017) and Cui et al. (2017)), we introduce two specific ways of charging fees for the tontines, and compare the two schemes: a (possibly) time-varying proportional fee of the payoff charged over time and a fixed up-front fee. Before the tontines or tontine-like products are launched, it is likely that the tontine providers will check the regularly used fee structures for similar products existing on the market. As both fixed and variable fees are used for regular and variable annuities, we believe that both fee types can also be adopted for tontine products. Forman and Sabin (2016) introduce a so-called survivor fund, a product very similar to tontines, and mention that there would be a one-time administrative fee at the end of the contract and regular management and record-keeping fees. We find that a constant fraction of the payoff being deducted over time can yield the same level of expected utility to policyholders as a fixed up-front fee, where the market value of both fees is identical. We also design a time-decreasing proportional fee deducted over time, and in our numerical exercise, we find that it delivers a slightly lower level of expected utility than the fixed up-front fee for the chosen parameters.

Another motive for this study is the so-called annuity puzzle. This is a term used to describe the discrepancy between the theoretical demand for annuities (see, for instance, Yaari (1965) and Peijnenburg et al. (2016)) and the fact that few households voluntarily purchase an annuity (see, for example, Hu and Scott (2007), Inkmann et al. (2010) and Lockwood (2012)). Much of the relevant literature offers explanations regarding the annuity puzzle (see e.g. Beshears et al. (2014), Agnew et al. (2008), Brown et al. (2017) and Zhang et al. (2020)), using empirical and behavioral approaches. One interpretation is that individuals perceive annuities as too expensive. In our paper, we focus on the design of the fee structure for tontines to show that these can become an attractive alternative to annuities, both for policyholders and for insurers.

To gain further insights into the fee charged for tontine products, we compare tontines with annuities. Under a fixed up-front fee and for a given annuity, we then determine the critical indifference loading for the tontine which makes a policyholder indifferent to choosing between a fee-based tontine or the annuity. This fee can be interpreted as the maximum willingness to pay for the tontine product, at which the policyholders are prepared to buy the tontine products when given the choice between an annuity or a tontine. This indifference fee turns out to be very close to the fee of the annuity. For example, we find, under our parameter choices, that policyholders with a relative risk aversion level of 10 are indifferent to choosing between a tontine with a fee of 12.5% and an annuity with a fee of 14%. For risk aversions below 10, the corresponding indifference fee of the tontine is even larger and comes closer to the 14% of the annuity. The high maximum willingness to pay from our analysis suggests that tontine providers have significant scope to set a reasonable fee for tontines. Organizing and managing a tontine can become quite mechanical, and can be easily integrated into any "fintech" investment platform or an "insurtech" company that would simply have to design a particular form of dividend distribution, based only on surviving participants on top of a smart long-term investment strategy. Furthermore, as negligible safety loadings will be charged for the longevity risk in a tontine product (see also Chen et al. (2019)), any fee charges which lie between the essential management and indifference fee are likely to be reasonable for a tontine product. The difference between the indifference fee and the actual fee charged will become a benefit to the tontine participants, as a saving in fees. This would boost their income and on average would compensate the instability of the tontine income. Forman and Sabin (2015, 2016) also point out that tontines and survivor funds could have significantly lower fees than annuities. By setting certain profit margins for tontine participants, more customers would be attracted to tontine products.

To further examine the comparative advantages of fee-based tontines over fee-based annuities, we carry out a mean-variance, and a quantile analysis for the insurer, starting with an identical number of contracts in each case. We find that the expected profit of the annuity is slightly larger, because the insurer may charge higher fees for annuities. However, the variance of the annuity payoff exceeds that of the tontine payoff drastically over a longer period of the retirement phase. From this observation and the quantile analysis, we can conclude that fee-based tontines are less volatile and easier to manage from an insurer's point of view. A comparison of the coefficients of variation of the annuity and tontine payoffs confirms these observations, as the coefficient of variation of the tontine payoff is below that of the annuity payoff over a longer period of the retirement phase. Finally, under the indifference fee (which makes policyholders indifferent to choosing between a tontine or an annuity), we find that reserves of tontines are lower than reserves of annuities at practically all times during the retirement phase. All in all, our results suggest that insurers could largely benefit from issuing tontines with fees, as the fees obtained from tontines, especially compared to annuities, are an (almost) certain profit for insurers. Given that the actual administration expenses of tontines could be fairly low in an increasingly digitalized world, insurers could expect to retain a greater part of the fees as profits.

Unlike Chen et al. (2020) and Chen and Rach (2019), which build on previous studies

(Milevsky and Salisbury (2015)) by coming up with some innovative combinations of tontines and annuities, the current paper contributes to the existing literature by analyzing different fee structures in tontines and their impact on benefit for the policyholders and insurers in comparison to annuities. It is widely accepted that policyholders prefer tontines to annuities with rather low safety loadings (see e.g. Stamos (2008), Hanewald et al. (2013), Donnelly et al. (2013) and Milevsky and Salisbury (2015)). Our results are consistent with this literature in showing that the indifference fee of the tontine can come extremely close to the fee of the annuity. However, this literature focuses exclusively on the perspective of the policyholder, whereas our paper additionally shows that insurers might generate sizeable and secure profits from underwriting fee-based tontines. Nevertheless, our results also carry a certain risk by giving insurers significant freedom to set fees for tontines which might be too similar to the fees of annuities. Hence, when insurers start issuing tontines in the (near or distant) future, it seems reasonable to suggest that regulators should observe and, if necessary, control the corresponding fees.

The remainder of our article is organized in the following way: In Section 2, we describe the model setup, in particular the mortality model, the tontine structure and the organization of fees, and derive the utility-maximizing tontine payoffs under each fee structure. In Section 3, taking the insurer's perspective, we compare fee-based tontines to fee-based annuities. To be precise, we derive the indifference loading of the tontine which makes policyholders indifferent to choosing between annuities or tontines, and analyze the variance, coefficient of variation and quantiles of the tontine and annuity payoffs over time. The section ends with an analysis of the reserves of annuities and tontines. Section 4 concludes the article and is followed by an appendix with one proof.

2 Optimal tontines with fees

In this section, we describe the model setup used throughout the paper. Following e.g. Yaari (1965) and Milevsky and Salisbury (2015), we consider a continuous-time setting.

2.1 A simple stochastic mortality model

In this paper, we disregard financial market risk in order to focus exclusively on mortality risk. We assume a rather simple stochastic mortality risk model which allows us to distinguish two sources of mortality risk: unsystematic (or idiosyncratic) and systematic (or aggregate) mortality risk. Unsystematic mortality risk stems from the fact that lifetimes of individuals are unknown but still follow some mortality law. This risk component can be diversified through pooling, i.e. it tends to disappear if the portfolios are large enough. The systematic mortality risk, also called longevity risk in the context of retirement products, stems from the fact that we are not able to determine the actual "true" mortality law with certainty. This risk component affects all policies in the same way. In the context of retirement products, this component can be identified as the so-called longevity risk, that is, the risk of an overall unanticipated decline in mortality rates (see, for example, Pitacco et al., 2009). When it is present, even with a large portfolio, there is a residual part of risk that cannot be eliminated.

For any x-year-old policyholder, the best-estimate t-year survival probability is denoted by ${}_tp_x$ which can be computed from some continuous-time mortality law. We incorporate uncertainty in this mortality law in a similar way to, for example, Lin and Cox (2005) by applying a random shock ϵ to the best-estimate survival probabilities. Given the shock ϵ, the shocked survival curve is given by ${}_tp_x^{1-\epsilon}$. This shock covers the systematic mortality risk described above. We assume that ϵ is a continuous random variable taking values in $(-\infty, 1)$ almost surely whose density is denoted by $f_\epsilon(\cdot)$. The special case without longevity shock is simply obtained by setting $\epsilon = 0$.

2.2 Tontine: Payoff and premium

We denote the remaining future lifetime of the policyholder by ζ_ϵ, with ϵ being the random longevity shock as defined in Subsection 2.1. At time 0, the policyholder invests his initial wealth v in a retirement product. In a tontine contract, mortality risk is shared among a homogeneous pool of $n \geq 1$ policyholders who are of the same age.[3] The unsystematic mortality risk can initially be diversified by a large enough pool size n. The systematic risk, however, cannot be diversified, as this type of risk affects all the policyholders in the same way. At older ages, as the pool size decreases, the remaining policyholders are left with both systematic and unsystematic risk. We follow the specific tontine design described by Milevsky and Salisbury (2015), denoting by $N_\epsilon(t)$ the number of policyholders alive at time t. As mentioned earlier, we consider two explicit fee structures for the tontine product, i.e. a fixed initial fee and a variable fee. Depending on the fee structure, the survival benefits received by each policyholder are different. In the case of a fixed initial fee, each policyholder receives a stream of survival benefits

[3]That is, the insurer carries the longevity risk of the last survivor in the pool.

$nd^F(t)/N_\epsilon(t)$, where $d^F(t)$ is a deterministic payment stream specified at the beginning of the contract:[4]

$$b^F(t) := \mathbb{1}_{\{\zeta_\epsilon > t\}} \frac{nd^F(t)}{N_\epsilon(t)}, \tag{1}$$

where we follow the notation of Milevsky and Salisbury (2015). In the case of a variable fee, the fraction $\alpha(t) \in (0,1)$ of the survival benefits is used as a time-varying proportional fee. In other words, the policyholder receives $(1-\alpha(t))nd^V(t)/N_\epsilon(t)$ at time t, while the insurer retains $\alpha(t)nd^V(t)/N_\epsilon(t)$ from each living policyholder. In total, in the case of a variable fee, this yields the following continuous payment stream to a policyholder for each $t \geq 0$:[5]

$$b^V(t) := \mathbb{1}_{\{\zeta_\epsilon > t\}} \frac{(1-\alpha(t))nd^V(t)}{N_\epsilon(t)}. \tag{2}$$

In particular, as in an annuity product, the tontine provides an income for life, where the income provided by the tontine (2) depends on the number of survivors $N_\epsilon(t)$.

The valuation of tontine products has been well studied in recent literature, e.g. Milevsky and Salisbu (2015) and Chen et al. (2019). Working from these recent studies, we apply the fact that the remaining number of individuals being alive at time $t > 0$ follows a binomial distribution, that is, $(N_\epsilon(t) - 1 \mid \zeta_\epsilon > t, \epsilon) \sim \mathrm{Bin}\,(n-1, {}_tp_x^{1-\epsilon})$, conditional on the considered policyholder still being alive and given ϵ. Here, we make use of the (conditional) independence of the pool members. Following equation (2.5) in Chen et al. (2019), for the fixed and continuous fees, the corresponding net premiums, P_0^F and P_0^V, for the tontine payoffs in (1) and (2) are given by

$$\begin{aligned}
P_0^F &= \int_0^\infty e^{-rt}\, d^F(t) \int_{-\infty}^{1} \left(1 - \left(1 - {}_tp_x^{1-\varphi}\right)^n\right) f_\epsilon(\varphi)\, \mathrm{d}\varphi\, \mathrm{d}t \\
P_0^V &= \int_0^\infty e^{-rt}\, (1-\alpha(t))d^V(t) \int_{-\infty}^{1} \left(1 - \left(1 - {}_tp_x^{1-\varphi}\right)^n\right) f_\epsilon(\varphi)\, \mathrm{d}\varphi\, \mathrm{d}t,
\end{aligned}$$

where r is the constant risk-free rate.

[4] For the case $N_\epsilon(t), = 0$ we define $b^F(t) := 0$.

[5] Similar to the fixed initial fee, for the case $N_\epsilon(t) = 0$, we define $b^V(t) := 0$.

2.3 Optimization problem in the case of a fixed fee

First, we assume that the insurer charges a single fixed fee for the tontine at time 0. This implies that, in addition to the net premium, the policyholder shall pay a single up-front fee M_0/n. M_0 can be considered as the total expenses the insurer will charge from all the policyholders to bring the tontine into operation, and is split equally among the policyholders.

Assuming that the policyholder is interested in maximizing his/her expected lifetime utility of a tontine with the payout function defined in (1), the expected utility is given by

$$\begin{aligned} U\big(\{d^F(t)\}_{t\geq 0}\big) &:= \mathbb{E}\left[\int_0^\infty e^{-\rho t} u\left(\frac{n d^F(t)}{N_\epsilon(t)}\right) \mathbb{1}_{\{\zeta_\epsilon > t\}} \mathrm{d}t\right] \\ &= \int_0^\infty e^{-\rho t}\, \kappa_{n,\gamma,\epsilon}({}_tp_x)\, u\left(d^F(t)\right) \mathrm{d}t\,, \end{aligned} \tag{3}$$

where $u(x) = \frac{x^{1-\gamma}}{1-\gamma}$, $\gamma > 0$, $\gamma \neq 1$, and

$$\begin{aligned} \kappa_{n,\gamma,\epsilon}({}_tp_x) &= \mathbb{E}\left[\left(\frac{n}{N_\epsilon(t)}\right)^{1-\gamma} \mathbb{1}_{\{\zeta_\epsilon > t\}}\right] \\ &= \sum_{k=1}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^{\gamma} \int_{-\infty}^{1} \left({}_tp_x^{1-\varphi}\right)^k \left(1 - {}_tp_x^{1-\varphi}\right)^{n-k} f_\epsilon(\varphi)\, \mathrm{d}\varphi. \end{aligned}$$

Throughout the paper, we assume a constant relative risk aversion utility function to describe a policyholder's preferences for each tontine payout. Furthermore, the individual uses a subjective discount factor ρ which might differ from the risk-free rate r.

After calculating the expected lifetime utility, the optimization problem in the case of a fixed fee can be presented as follows:

$$\max_{d^F(t)} \int_0^\infty e^{-\rho t}\, \kappa_{n,\gamma,\epsilon}({}_tp_x)\, u\left(d^F(t)\right) \mathrm{d}t \text{ subject to } P_0^F + \frac{M_0}{n} \leq v. \tag{4}$$

The optimization problem is conducted under a budget constraint. In this optimization problem, we have assumed that the policyholder's initial wealth v is fully used to invest in the tontine, whose total costs (net premium P_0^F plus the additional fee M_0/n) shall not exceed the initial wealth level.

Theorem 2.1. *The solution to problem* (4) *is given by*

$$d^{F*}(t) = \frac{e^{\frac{(r-\rho)t}{\gamma}} \left(\kappa_{n,\gamma,\epsilon}({}_tp_x)\right)^{1/\gamma}}{\lambda_F^{1/\gamma} \left(\int_{-\infty}^{1} \left(1 - \left(1 - {}_tp_x^{1-\varphi}\right)^n\right) f_\epsilon(\varphi)\,\mathrm{d}\varphi\right)^{1/\gamma}}, \tag{5}$$

where the optimal Lagrangian multiplier λ_F *is given by*

$$\lambda_F = \left(\frac{1}{v - \frac{M_0}{n}} \left(\int_0^{\infty} e^{\left(\frac{1}{\gamma}-1\right)rt - \frac{1}{\gamma}\rho t} \frac{\left(\kappa_{n,\gamma,\epsilon}({}_tp_x)\right)^{1/\gamma}}{\left(\int_{-\infty}^{1} \left(1 - \left(1 - {}_tp_x^{1-\varphi}\right)^n\right) f_\epsilon(\varphi)\,\mathrm{d}\varphi\right)^{1/\gamma - 1}} \,\mathrm{d}t\right)\right)^{\gamma}.$$

The optimal level of expected utility is given by

$$U_F = \frac{1}{1-\gamma} \cdot \lambda_F \cdot \left(v - \frac{M_0}{n}\right).$$

Proof. Due to the strictly increasing utility, the optimal solution suggests that the budget constraint is binding. In other words, we can reduce the budget constraint to an equality, i.e. $P_0^F + M_0/n = v$. We can reformulate this as

$$v - \frac{M_0}{n} = P_0^F.$$

This allows us to solve (4) by applying Theorem 2(b) in Chen et al. (2019) using the reduced initial wealth $v - \frac{M_0}{n}$. □

The case with no fees can obviously be obtained by setting the fee M_0 equal to zero. Clearly, for all $\gamma > 0$, $\gamma \neq 1$, the optimal level of expected utility decreases in M_0. In other words, the larger the fee M_0, the lower the expected utility of the policyholder.

2.4 Optimization problem in the case of a variable fee

Instead of a fixed fee, we now move to a time-varying proportional fee, i.e. a function $\alpha(t)$ which depends on t. In this case, the policyholder receives $(1 - \alpha(t))nd^V(t)/N_\epsilon(t)$ at time t, while the insurer retains $\alpha(t)nd^V(t)/N_\epsilon(t)$ from each living policyholder. Thus, the utility maximization problem for the tontine payoff defined in (2) becomes:

$$
\begin{aligned}
&\max_{d^V(t)} \int_0^\infty e^{-\rho t}\, \kappa_{n,\gamma,\epsilon}({}_tp_x)\, u\left((1-\alpha(t))d^V(t)\right) \mathrm{d}t \\
&\text{subject to } v \geq \int_0^\infty e^{-rt}\alpha(t)d^V(t) \int_{-\infty}^{1} \left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right) f_\epsilon(\varphi)\,\mathrm{d}\varphi\,\mathrm{d}t \\
&\qquad + \int_0^\infty e^{-rt}(1-\alpha(t))d^V(t) \int_{-\infty}^{1} \left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right) f_\epsilon(\varphi)\,\mathrm{d}\varphi\,\mathrm{d}t \\
&\qquad := \text{present value of the variable fees} + \text{net premium of (2).}
\end{aligned} \tag{6}
$$

As in the case of a fixed fee, the optimization problem is conducted subject to a budget constraint: The total initial wealth v of the policyholder is used to cover the net premium of the tontine payoff in (2) and the variable fees.

Theorem 2.2 provides the solution to Problem (6).

Theorem 2.2. *The solution of problem* (6) *is given by*

$$
d^{V*}(t) = \frac{1}{1-\alpha(t)} \cdot \frac{e^{\frac{(r-\rho)t}{\gamma}}\left((1-\alpha(t))\kappa_{n,\gamma,\epsilon}({}_tp_x)\right)^{1/\gamma}}{\lambda_V^{1/\gamma}\left(\int_{-\infty}^1 \left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right) f_\epsilon(\varphi)\,\mathrm{d}\varphi\right)^{1/\gamma}}, \tag{7}
$$

where the optimal Lagrangian multiplier λ_V is given by

$$
\lambda_V = \left(\frac{1}{v}\left(\int_0^\infty e^{\left(\frac{1}{\gamma}-1\right)rt-\frac{1}{\gamma}\rho t}\frac{1}{1-\alpha(t)} \cdot \frac{\left((1-\alpha(t))\kappa_{n,\gamma,\epsilon}({}_tp_x)\right)^{1/\gamma}}{\left(\int_{-\infty}^1 \left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right) f_\epsilon(\varphi)\, d\varphi\right)^{1/\gamma-1}}\,\mathrm{d}t\right)\right)^{\gamma}.
$$

The optimal level of expected utility is given by

$$
U_V = \frac{1}{1-\gamma} \cdot \lambda_V \cdot v.
$$

Proof: See Appendix A.1.

Some questions naturally arise at this point: When is one fee structure superior to the other? Does a relation exist between M_0 and $\alpha(t)$ to make the insurer/policyholder indifferent between the two fee structures? From the insurer's perspective, we look at the initial value of the total fee incomes the insurer obtains. From the fixed fee scheme, the total fee income of the insurance is given by M_0, while from the variable fee scheme, the initial value of the total fee income is determined by

$$\mathbb{E}\left[\int_0^\infty e^{-rt}\alpha(t)\,\frac{n\,d^V(t)}{N_\epsilon(t)}\cdot N_\epsilon(t)\cdot \mathbb{1}_{\{N_\epsilon(t)>0\}}\mathrm{d}t\right]$$
$$=\int_0^\infty e^{-rt}\,\alpha(t)\,n\,d^V(t)\int\limits_{-\infty}^{1}\left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right)f_\epsilon(\varphi)\,\mathrm{d}\varphi\,\mathrm{d}t. \tag{8}$$

The insurer, in total, obtains $N_\epsilon(t)\cdot\frac{\alpha(t)nd^V(t)}{N_\epsilon(t)}\mathbb{1}_{\{N_\epsilon(t)>0\}}$ variable payments from the surviving policyholders. Note that if

$$M_0\geq\int_0^\infty e^{-rt}\,\alpha(t)\,n\,d^V(t)\int\limits_{-\infty}^{1}\left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right)f_\epsilon(\varphi)\,\mathrm{d}\varphi\,\mathrm{d}t, \tag{9}$$

the fixed fee scheme is more (weakly) beneficial than its variable counterpart. Apparently, the reverse argument holds when "$\leq$", instead of "$\geq$", appears in the above condition. Furthermore, the condition in (9) will be reduced to some more intuitive relations between M_0 and the variable fraction $\alpha(t)$ if a specific form for $\alpha(t)$ is assumed.

From the policyholder's point of view, we further analyze and compare his/her expected lifetime utility from the fixed and variable fee scheme. Considering two fee structures with identical initial values, the policyholder would choose the structure which results in a higher expected discounted lifetime utility.

In the following, we present some special cases for the time-varying fee $\alpha(t)$.

2.4.1 Variable fee structure with a constant α

At any instant $t\geq 0$, we assume $\alpha(t)\equiv\alpha\in(0,1)$. Let us first write down the optimal withdrawal payment for this case, which can be obtained by replacing $\alpha(t)$ by α in (7).

Corollary 2.3. *The solution to problem* (6) *for the special case* $\alpha(t)\equiv\alpha\in(0,1)$, $\forall t$, *is given by*

$$d^{V*}(t)=(1-\alpha)^{1/\gamma-1}\frac{e^{\frac{(r-\rho)t}{\gamma}}\left(\kappa_{n,\gamma,\epsilon}({}_tp_x)\right)^{1/\gamma}}{(\lambda_V^c)^{1/\gamma}\left(\int_{-\infty}^1\left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right)f_\epsilon(\varphi)\,\mathrm{d}\varphi\right)^{1/\gamma}},$$

where the optimal Lagrangian multiplier λ_V^c *is given by*

$$\lambda_V^c = (1-\alpha)^{1-\gamma}\left(\frac{1}{v}\left(\int_0^\infty e^{\left(\frac{1}{\gamma}-1\right)rt-\frac{1}{\gamma}\rho t}\frac{\left(\kappa_{n,\gamma,\epsilon}({}_tp_x)\right)^{1/\gamma}}{\left(\int_{-\infty}^1\left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right)f_\epsilon(\varphi)\,\mathrm{d}\varphi\right)^{1/\gamma-1}}\,\mathrm{d}t\right)\right)^\gamma.$$

The optimal level of expected utility is given by

$$U_V^c = \frac{1}{1-\gamma}\cdot\lambda_V^c\cdot v.$$

In this case, the insurer obtains $N_\epsilon(t)\cdot\frac{\alpha n d^{V*}(t)}{N_\epsilon(t)}\mathbb{1}_{\{N_\epsilon(t)>0\}}$ variable payments from the surviving policyholders. The initial value of these payments, collected from all the policyholders, can be computed as:

$$\int_0^\infty e^{-rt}\,\alpha\,n\,d(t)\int_{-\infty}^{1}\left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right)f_\epsilon(\varphi)\,\mathrm{d}\varphi\,\mathrm{d}t = \frac{n\cdot\alpha}{1-\alpha}P_0^V$$

Condition (9) suggests a critical value of α which makes the insurer indifferent to choosing between the fixed or variable fee (M_0 is the total fixed fee):

$$\alpha^* = \frac{M_0}{nP_0^V+M_0} = \frac{M_0/n}{P_0^V+M_0/n} = \frac{M_0}{nv}. \tag{10}$$

where we have used $P_0^V + M_0/n = v$ which is the budget constraint of Problem (4), i.e. we assume that for both the fixed and the variable fee problem the policyholder has the same initial wealth v.

Next, we want to examine whether the indifference fee fraction α^*, resulting from the perspective of the insurer, is also the critical fee that makes the policyholder indifferent to choosing between the two fee schemes. For this, we look at the expected lifetime utility of the policyholder, given α^*.

Proposition 2.4. *Given* $\alpha^* = \frac{M_0}{nv}$, *we obtain*

(i) $(1-\alpha^*)d^{V*}(t) = d_F^*(t)$;

(ii) $U_V^c = U_F$.

Proof. We prove both points in detail, where (ii) follows directly from (i):

(i) Note that the optimal payoffs $d^{F*}(t)$ and $d^{V*}(t)$ (cf. Corollary 2.3) take similar forms. We have

$$\begin{aligned} d^{V*}(t) =& (1-\alpha^*)^{1/\gamma-1}\frac{\lambda_F^{1/\gamma}}{(\lambda_V^c)^{1/\gamma}}d^{F*}(t) \\ =& (1-\alpha^*)^{1/\gamma-1}\frac{v}{v-\frac{M_0}{n}}(1-\alpha^*)^{1-1/\gamma}d^{F*}(t) \\ =& \frac{v}{v-\frac{M_0}{n}}d^{F*}(t) \overset{\alpha^*=\frac{M_0}{nv}}{=} \frac{1}{1-\alpha^*}d^{F*}(t). \end{aligned}$$

(ii) The two utility levels differ only in the Lagrangian multiplier and the initial wealth. In particular, we have

$$\begin{aligned} U_V^c =& \frac{1}{1-\gamma}\cdot\lambda_V^c\cdot v \\ =& \frac{1}{1-\gamma}(1-\alpha^*)^{1-\gamma}v^{-\gamma}\left(v-\frac{M_0}{n}\right)^{\gamma}\cdot\lambda_F\cdot v \\ \overset{\alpha^*=\frac{M_0}{nv}}{=}& \frac{1}{1-\gamma}\cdot\lambda_F\cdot\left(v-\frac{M_0}{n}\right) = U_F. \end{aligned}$$

□

Proposition 2.4 shows that the indifference fee fraction α^*, resulting from the perspective of the insurer, makes the policyholder indifferent to choosing between the fixed or variable fee scheme, in terms of expected lifetime utility. For other cases, we can make the following remarks.

- If a variable fee with $\alpha > \alpha^* = \frac{M_0}{nv}$ is charged, the insurer benefits from higher variable fee incomes, while it makes the policyholder worse off than when a fixed single fee is applied.
- If a variable fee with $\alpha < \alpha^* = \frac{M_0}{nv}$ is charged, it makes the policyholder better off than with a fixed single fee, but the insurer's fee income deteriorates.

In our paper, the optimization problems are done under a budget constraint, which can be considered as a risk-neutral insurer's participation constraint, where the insurer's utility is characterized by a linear function. If we allow the insurer to also be risk averse, the above indifference result might not hold.

We use the remainder of this subsection to illustrate a numerical example. Table 1 provides the base case parameters used in the majority of the subsequent numerical analyses.

Initial wealth $v = 100$	Pool size $n = 1000$	Risk aversion $\gamma = 4$
Risk-free rate $r = 0.01$	Subjective discount rate $\rho = r$	Fee $M_0 = 5000$
Initial age $x = 65$	Gompertz law $m = 88.721,\ \beta = 10$	Longevity shock $\epsilon \sim \mathcal{N}_{(-\infty,1)}\left(-0.0035, 0.0814^2\right)$

Table 1: Base case parameter setup.

The parameters we choose coincide mostly with the parameters chosen in Chen et al. (2020):

- For the pool size n, we follow Qiao and Sherris (2013) who assume a similar pool size for their analysis of group self-annuitization schemes.
- The risk-free interest rate is chosen close to zero, based on the prevailing low interest rate environment in many developed countries. In Germany, for instance, the average risk-free rate of investment in 2019 was equal to 1.1% (cf. Statista, 2019).
- For the mortality law we assume the well-known Gompertz law (see Gompertz, 1825) which is frequently used in actuarial science (see Milevsky, 2020). To be precise, we assume that for any x and $t \geq 0$, the force of mortality is given by

$$\mu_{x+t} = \frac{1}{\beta} e^{\frac{x+t-m}{\beta}},$$

  and the t-year survival probability of an x-year old is given by

$${}_tp_x = e^{-\int_0^t \mu_{x+s} \mathrm{d}s} = e^{e^{\frac{x-m}{\beta}}\left(1-e^{\frac{t}{\beta}}\right)},$$

  where $m > 0$ denotes the modal age at death and $\beta > 0$ is the dispersion coefficient. For the values of m and β, we follow Milevsky and Salisbury (2015).
- Concerning the longevity shock ϵ, we follow Chen et al. (2019) and assume that it follows a truncated normal distribution on the interval $(-\infty, 1)$. The parameters used for this distribution also stem from Chen et al. (2019).

- The fee level of $M_0 = 5000$ results in a fee of $M_0/n = 5$ for each individual retiree. Note that due to the choice of the initial wealth level, this amounts to 5% of the initial wealth. The critical proportion α^* that results from this choice can be calculated as $\alpha^* = 0.05$.

In Figure 1, we provide the optimal tontine payoff (5) for the initial fee M_0 given in Table 1 and for $M_0 = 0$. We have added $M_0 = 0$ to compare fee-based tontines with those with no fees.

Figure 1: Optimal payoff (5) for two fee levels $M_0 = 5000$ and $M_0 = 0$. The parameters are chosen as in Table 1.

As the optimal payoff (5) is proportional to the initial investment v, a fixed fee leads to a reduced payoff at all times compared to a payoff with no fee. Note that this payoff is identical to the payoff resulting under a constant proportional fee charged over time which is therefore omitted in Figure 1.

2.4.2 Time-decreasing fee structure

In addition to the constant α, we consider a time-decreasing fee structure, with the reasoning that more medical/care expenses are needed for more advanced ages, e.g. $\alpha(t) = \alpha_0 e^{-gt}$, with $\alpha_0 \in (0,1)$ and $g \geq 0$. Furthermore, we assume that fees are only paid until a certain age $x+\tau$, e.g. until 75, in which case the fee becomes

$$\alpha(t) = \begin{cases} \alpha_0 e^{-gt} & t \leq \tau \\ 0 & t > \tau. \end{cases} \tag{11}$$

In contrast to the case of $\alpha(t) = \alpha$, the specific form of (11) does not allow us to derive a proposition, such as Proposition 2.4. In other words, we can only rely on numerical procedures to compare the fixed with the variable fee scheme.

Note that with this specific form for $\alpha(t)$, condition (9) becomes

$$M_0 \geq \int_0^\tau e^{-(r+g)t} \alpha_0 \cdot n \cdot d^V(t) \int_{-\infty}^{1} \left(1 - \left(1 - {}_tp_x^{1-\varphi}\right)^n\right) f_\epsilon(\varphi)\,\mathrm{d}\varphi\,\mathrm{d}t. \tag{12}$$

Note that the optimal withdrawal plan $d^V(t)$ depends on α_0, which is why we need to rely on numerical procedures (e.g. the bisection method) to find an α_0 which achieves equality in (12). We rely on the base case parameters summarized in Table 1 for the parameterized fee structure (11):

Growth rate	Duration	Initial fee
$g = 0.02$	$\tau = 10$	$\alpha_0 = 0.109$

Table 2: Base case parameter setup for the parameterized fee structure (11).

For these parameters, we obtain a critical value of the initial fee $\alpha_0 = 0.109$. In Figure 2, we provide the optimal tontine payoff (7) multiplied by $(1-\alpha(t))$ and compare it to the optimal payoff (5) under a constant fee α, both relying on the base case parameters presented in Table 1.

We observe that the payoff resulting under the time-varying fee is lower (higher) than the payoff resulting under the constant fee before (after) the duration of the fee payments is reached. At time 10, there is a jump in the payoff resulting under the time-varying fee, since the payoff switches from a payoff with fee to a payoff with no fee.

2.4.3 Comparison of constant and time-decreasing fee structure

In this subsection, we want to work out whether a time-decreasing fee structure or a constant fee structure is more beneficial to the retirees. To compare the utility levels

Figure 2: Optimal payoffs (7) multiplied by $(1-\alpha(t))$ and (5) for the fee level $M_0 = 5000$. The parameters are chosen as in Table 1 and 2.

resulting under the two schemes in an appropriate way, we introduce the certainty equivalent CE as the level of the deterministic lifelong retirement payment that yields the same expected utility as some given retirement plan (which in our case is the tontine under a constant and a time-decreasing fee fraction). That is, CE is determined by

$$\mathbb{E}\left[\int_0^\infty e^{-\rho t}\mathbb{1}_{\{\zeta_\epsilon > t\}} u(\mathrm{CE})\,\mathrm{d}t\right] = \int_0^\infty e^{-\rho t}{}_tp_x\, m_\epsilon(-\ln {}_tp_x) u(\mathrm{CE})\,\mathrm{d}t = U_V,$$

or equivalently,

$$\mathrm{CE} = \left((1-\gamma)U_V\left(\int_0^\infty e^{-\rho t}{}_tp_x\, m_\epsilon(-\ln {}_tp_x)\,\mathrm{d}t\right)^{-1}\right)^{\frac{1}{1-\gamma}},$$

where U_V is the expected discounted lifetime utility of the individual resulting under a variable fee structure (either constant α or decreasing $\alpha(t)$). Note that the certainty equivalent is an increasing function in U_V.

In Table 3, we compare the certainty equivalents resulting under a constant proportional fee α and a time-decreasing fee $\alpha(t)$. We observe that a constant fee delivers a higher CE for all constellations of subjective discount factors and relative risk aversions. As the risk aversion increases, the difference between the certainty equivalents diminishes, implying that retirees with a lower risk aversion suffer larger losses under the time-decreasing fee

	$\gamma = 0.5$	$\gamma = 4$	$\gamma = 8$
$\rho = 0$	(5.208,5.184)	(5.145,5.143)	(5.107,5.106)
$\rho = r$	(5.172,5.155)	(5.141,5.140)	(5.106,5.105)
$\rho = 0.02$	(5.207,5.194)	(5.147,5.145)	(5.110,5.109)

Table 3: Certainty equivalents $(\mathrm{CE}_V^c, \mathrm{CE}_V^{\mathrm{dec}})$ obtained under a constant and a time-decreasing proportional fee structure dependent on the risk aversion and the subjective discount factor. The remaining parameters are chosen as in Tables 1 and 2.

structure. However, note that the losses are only very small as the retiree adjusts his/her optimal payoff to each fee structure by choosing the appropriate optimal tontine payoff.

In Table 4, we show a similar analysis as in Table 3 for an alternative fee of $M_0 = 10000$. We make similar observations as in Table 3. The main difference is that all the certainty

	$\gamma = 0.5$	$\gamma = 4$	$\gamma = 8$
$\rho = 0$	(4.934,4.800)	(4.874,4.867)	(4.838,4.835)
$\rho = r$	(4.900,4.811)	(4.871,4.864)	(4.838,4.834)
$\rho = 0.02$	(4.933,4.870)	(4.876,4.870)	(4.841,4.838)

Table 4: Certainty equivalents $(\mathrm{CE}_V^c, \mathrm{CE}_V^{\mathrm{dec}})$ obtained under a constant and a time-decreasing proportional fee structure dependent on the risk aversion and the subjective discount factor. The remaining parameters are chosen as in Tables 1 and 2 except for $M_0 = 10000$.

equivalents have decreased and that the differences between the certainty equivalent of the constant and the time-decreasing fee have increased, both due to the higher fee $M_0 = 10000$.

3 Comparison to annuities

In this section, we compare tontines containing fees to annuities containing fees. For simplicity, we focus on the case with a fixed initial fee which is equal to considering a constant proportional fee over time. Note that the optimal payoffs and utility levels of the annuity have been included in all the previous sections and can simply be obtained by setting $n = 1$ in the tontine solutions.

3.1 Indifference fee

Typically, annuities contain higher fees than tontines. If we denote the initial fee charged for the annuity by Δ_0, we want to determine $M_0^*/n < \Delta_0$ such that the tontine and the annuity deliver the same levels of expected utility. A similar indifference loading for the annuity has already been considered by Milevsky and Salisbury (2015). It can be obtained by equating the optimal levels of expected utility of the corresponding optimal solutions:

$$\begin{aligned}\lambda_F\left(v - \frac{M_0^*}{n}\right) &= \left(v - \frac{M_0^*}{n}\right)^{1-\gamma}\left(\int_0^\infty e^{\left(\frac{1}{\gamma}-1\right)rt-\frac{1}{\gamma}\rho t}\frac{\left(\kappa_{n,\gamma,\epsilon}\left({}_tp_x\right)\right)^{1/\gamma}}{\left(\int_{-\infty}^1\left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right)f_\epsilon(\varphi)\,\mathrm{d}\varphi\right)^{1/\gamma-1}}\,\mathrm{d}t\right)^\gamma \\ &= (v-\Delta_0)^{1-\gamma}\left(\int_0^\infty e^{\left(\frac{1}{\gamma}-1\right)rt-\frac{1}{\gamma}\rho t}\int_{-\infty}^1 {}_tp_x^{1-\varphi} f_\epsilon(\varphi)\,\mathrm{d}\varphi\,\mathrm{d}t\right)^\gamma.\end{aligned}$$

Solving this equation for M_0^*/n delivers the indifference fee[6]

$$\frac{M_0^*}{n} = v - (v-\Delta_0)\frac{\left(\int_0^\infty e^{\left(\frac{1}{\gamma}-1\right)rt-\frac{1}{\gamma}\rho t}\int_{-\infty}^1 {}_tp_x^{1-\varphi} f_\epsilon(\varphi)\,\mathrm{d}\varphi\,\mathrm{d}t\right)^{\frac{\gamma}{1-\gamma}}}{\left(\int_0^\infty e^{\left(\frac{1}{\gamma}-1\right)rt-\frac{1}{\gamma}\rho t}\frac{\left(\kappa_{n,\gamma,\epsilon}({}_tp_x)\right)^{1/\gamma}}{\left(\int_{-\infty}^1\left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right)f_\epsilon(\varphi)\,\mathrm{d}\varphi\right)^{1/\gamma-1}}\,\mathrm{d}t\right)^{\frac{\gamma}{1-\gamma}}} \tag{13}$$

as critical fee of the tontine. In the following studies, the following values for the fee of the annuity are considered:

- Chen et al. (2019) derive a loading of 4.8% for the annuity using parameters similar to ours. Therefore, we set the minimum fee at $\Delta_0 = 0.04 \cdot v$, i.e. 4% of the initial wealth of the policyholder is used to finance the fees and the remaining 96% is used to finance the benefits.
- Sabin (2010) writes that annuities are 14% higher than fair. For this reason, we set the maximum fee of the annuity at $\Delta_0 = 0.14 \cdot v$, i.e. 14% of the initial wealth of the policyholder is used to finance the fees and the remaining 86% is used to finance the benefits.

[6]Note that we have used the term "indifference fee" in Section 2 already, where it describes the critical fee level which makes the insurer indifferent to choosing between the variable and single up-front fee. As in this section, we focus only on the case with a single up-front fee, we use the term "indifference fee" again to label the critical fee level which makes the policyholder indifferent to choosing between the tontine and the annuity.

- Additionally, we consider the intermediate values $\Delta_0 = 0.07 \cdot v$ and $\Delta_0 = 0.1 \cdot v$.

Figure 3 provides the indifference fee (13) dependent on the relative risk aversion γ for the different annuity fees presented above. In all panels, we observe that the indifference fee decreases in the risk aversion. This result is what one would expect because individuals with a higher risk aversion seek the stable payments of an annuity more than individuals with a lower risk aversion, meaning that the latter ones are willing to pay a higher fee for a tontine than individuals with a higher risk aversion. A minor observation is that the critical tontine fee increases as the fee of the annuity increases. A quite unexpected outcome of Figure 3 is, however, the rather high magnitude of the indifference fee. Considering the base case parameters (Panel (a)) with an annuity fee of 14 as an example, for the highest risk aversion that we consider ($\gamma = 10$), the insurer may still charge a fee below 12.5 to make the tontine more attractive than the annuity (with a fee of 14). In other words, insurers would have to reduce their fees by at most $(14 - 12.5)/14 = 0.11 = 11\%$ compared to annuities in order to make tontines more appealing than annuities. For the lowest annuity fee $\Delta_0 = 4$, the critical tontine fee for the highest risk aversion is 2.34.

Under a higher standard deviation of the shock ϵ (Panel (b)), which translates into a higher longevity risk for the policyholders, the indifference fee decreases more steeply in the risk aversion and the indifference fees are lower than in the base case. This means that policyholders, particularly those with relatively high risk aversions, are less willing to buy a tontine with such a higher risk. Note that for an annuity loading of $\Delta_0 = 4$ the indifference fee of the tontine even falls below zero for the highest risk aversions. In this case, this means that the policyholder needs to be paid by the insurer to trade in tontines, instead of paying a fee.

If, however, the standard deviation is lowered (compared to the base case parameters, Panel (c)), the decrease of the indifference fee in the risk aversion is less pronounced and the indifference fees increase compared to the base case parameters. In other words, under a lower systematic mortality risk, individuals are more willing to buy tontines.

As already pointed out in the introduction, organizing and managing a tontine can become quite mechanical and consequently would not require a high management fee. Furthermore, compared to annuities where the annuity providers bear the longevity risk fully, the longevity risk in a tontine product is shared among the tontine participants. This implies that negligible safety loadings would be charged for the longevity risk in a tontine product (see also Chen et al. (2019)). The high indifference fee resulting from this part of our analysis suggests that tontine providers have ample scope for setting a reasonable fee for tontines. Any fee charges which lie between the essential management and indifference fee could be considered reasonable. The difference between the indifference fee and the actual fee charged will become a benefit to tontine participants through savings in fees. This would boost their income and on average would compensate the instability of the tontine income. By offering a range of profit margins to tontine participants, more customers would be attracted to tontine products.

3.2 Mean and variance analysis

In this subsection, we take a closer look at the expected profit and the volatility of the payoffs from the insurer's perspective. For simplicity, we again focus on the case with a fixed initial fee which is equal to considering a constant proportional fee over time. We assume that the indifference fee which makes the policyholder indifferent to choosing between a tontine or an annuity is applied in the subsequent analyses. Note that the payoffs of the tontine and the annuity are, from the insurer's perspective, given by

$$B_T(t) = nd^F(t)\mathbb{1}_{\{N_\epsilon(t)>0\}} \tag{14}$$

$$B_A(t) = c(t)N_\epsilon(t). \tag{15}$$

respectively, where $c(t)$ is the deterministic payout function of the annuity. The expected present value of future profits of the tontine is given by

$$\mathbb{E}\left[nv - \int_0^\infty e^{-rt} nd^F(t)\mathbb{1}_{\{N_\epsilon(t)>0\}}\mathrm{d}t\right] = nv - nP_0^F = M_0^*,$$

whereas the expected present value of n annuities is given by

$$\mathbb{E}\left[nv - \int_0^\infty e^{-rt} c(t)N_\epsilon(t)\mathrm{d}t\right] = nv - nP_0^{F,A} = n\Delta_0.$$

Hence, the expected value of future profits of the annuity exceeds that of the tontine. However, the profits of the annuity might have to be used as a security buffer due to the longevity risk contained in annuities, whereas the risks contained in tontines are much less pronounced. We illustrate this by an analysis of the variance of the payoffs of the two retirement products. The variance of the tontine payoff is given by

$$\mathrm{Var}\left(nd^F(t)\mathbb{1}_{\{N_\epsilon(t)>0\}}\right) = (nd^F(t))^2\mathrm{Var}\left(\mathbb{1}_{\{N_\epsilon(t)>t\}}\right)$$

$$= (nd^F(t))^2 \left(\mathbb{E}\left[\mathbb{1}_{\{N_\epsilon(t)>t\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{N_\epsilon(t)>t\}}\right]^2\right)$$
$$= (nd^F(t))^2 \int\limits_{-\infty}^{1} \left(1 - \left(1 - {}_tp_x^{1-\varphi}\right)^n\right) f_\epsilon(\varphi)\,\mathrm{d}\varphi \left(\int\limits_{-\infty}^{1} \left(1 - {}_tp_x^{1-\varphi}\right)^n f_\epsilon(\varphi)\,\mathrm{d}\varphi\right).$$

Using the law of total variance, the variance of the annuity payoff can be determined as

$$\begin{aligned}
\mathrm{Var}\left(c(t)N_\epsilon(t)\right) &= (c(t))^2\mathrm{Var}\left(N_\epsilon(t)\right)\\
&= (c(t))^2\left(\mathrm{Var}\left(\mathbb{E}\left[N_\epsilon(t)\mid\epsilon\right]\right) + \mathbb{E}\left[\mathrm{Var}\left(N_\epsilon(t)\mid\epsilon\right)\right]\right)\\
&= (c(t))^2\left(\mathrm{Var}\left(n{}_tp_x^{1-\epsilon}\right) + \mathbb{E}\left[n{}_tp_x^{1-\epsilon}\left(1-{}_tp_x^{1-\epsilon}\right)\right]\right)\\
&= (c(t))^2\left(n^2\left(\mathbb{E}\left[\left({}_tp_x^{1-\epsilon}\right)^2\right] - \mathbb{E}\left[{}_tp_x^{1-\epsilon}\right]^2\right) + \mathbb{E}\left[n{}_tp_x^{1-\epsilon}\left(1-{}_tp_x^{1-\epsilon}\right)\right]\right)\\
&= (c(t))^2\left(n^2\left(\mathbb{E}\left[{}_tp_x^{2(1-\epsilon)}\right] - \mathbb{E}\left[{}_tp_x^{1-\epsilon}\right]^2\right) + n\left(\mathbb{E}\left[{}_tp_x^{1-\epsilon}\right] - \mathbb{E}\left[{}_tp_x^{2(1-\epsilon)}\right]\right)\right),
\end{aligned}$$

where $\mathbb{E}\left[{}_tp_x^{a(1-\epsilon)}\right] = {}_tp_x^a \cdot m_\epsilon(-a\ln{}_tp_x)$ for $a = 1, 2$. In Figure 4, we provide the variances of the annuity and the tontine derived above over time for different values of the annuity fee Δ_0, where we rely on the optimal tontine payoff $d^{F*}(t)$.

Although the variance of the annuity does not dominate the variance of the tontine at all times (for the four different fees, respectively), it seems clear from Figure 4 that the tontine is easier to manage from an insurer's perspective. As the variance measures deviations from the expectation upwards and downwards, we can also conclude that a tontine contains far less potential for profits and less risk for losses than annuities. Nevertheless, as we have seen in the previous analyses, the insurer may still charge a relatively high fee for the tontine which is practically a guaranteed profit, given the low variance of the tontine payoff. Additionally, we observe that the variance of the annuity decreases in the fee Δ_0. Naturally, a higher volatility implies a higher variance of the payoffs, in particular for the annuity.

As a final quantity of interest in this subsection, we look at the coefficient of variation (CV) at each time. This is defined as the standard deviation divided by the expectation, i.e. in the case of the tontine we obtain

$$\mathrm{CV}_T(t) = \frac{\sqrt{\mathrm{Var}(B_T(t))}}{\mathbb{E}\left[B_T(t)\right]}$$

$$= \frac{\sqrt{\int\limits_{-\infty}^{1} \left(1 - \left(1 - {}_tp_x^{1-\varphi}\right)^n\right) f_\epsilon(\varphi)\,\mathrm{d}\varphi \left(\int\limits_{-\infty}^{1} \left(1 - {}_tp_x^{1-\varphi}\right)^n f_\epsilon(\varphi)\,\mathrm{d}\varphi\right)}}{\int\limits_{-\infty}^{1} \left(1 - \left(1 - {}_tp_x^{1-\varphi}\right)^n\right) f_\epsilon(\varphi)\,\mathrm{d}\varphi}$$

$$= \sqrt{\frac{\int\limits_{-\infty}^{1} \left(1 - {}_tp_x^{1-\varphi}\right)^n f_\epsilon(\varphi)\,\mathrm{d}\varphi}{\int\limits_{-\infty}^{1} \left(1 - \left(1 - {}_tp_x^{1-\varphi}\right)^n\right) f_\epsilon(\varphi)\,\mathrm{d}\varphi}}.$$

For the annuity, however, we obtain

$$\begin{aligned} \mathrm{CV}_A(t) &= \frac{\sqrt{\mathrm{Var}(B_A(t))}}{\mathbb{E}\left[B_A(t)\right]} \\ &= \frac{\sqrt{n^2\left(\mathbb{E}\left[{}_tp_x^{2(1-\epsilon)}\right] - \mathbb{E}\left[{}_tp_x^{1-\epsilon}\right]^2\right) + n\left(\mathbb{E}\left[{}_tp_x^{1-\epsilon}\right] - \mathbb{E}\left[{}_tp_x^{2(1-\epsilon)}\right]\right)}}{n\mathbb{E}\left[{}_tp_x^{1-\epsilon}\right]} \\ &= \frac{\sqrt{n\left(\mathbb{E}\left[{}_tp_x^{2(1-\epsilon)}\right] - \mathbb{E}\left[{}_tp_x^{1-\epsilon}\right]^2\right) + \left(\mathbb{E}\left[{}_tp_x^{1-\epsilon}\right] - \mathbb{E}\left[{}_tp_x^{2(1-\epsilon)}\right]\right)}}{\sqrt{n}\cdot\mathbb{E}\left[{}_tp_x^{1-\epsilon}\right]}. \end{aligned}$$

Note that the coefficients of variation of the annuity and the tontine do not depend on the fee levels M_0 and Δ_0. Figure 5 provides the difference $\mathrm{CV}_A(t) - \mathrm{CV}_T(t)$ of the coefficients of variation of the annuity and tontine payoff over time as derived above.[7]

We observe that the CV of the tontine is lower than the CV of the annuity at all times considered (except for time 0, where both are equal to zero). Naturally, a higher standard deviation implies a higher difference between the CVs, as the standard deviation of the annuity is affected more strongly by the higher systematic mortality risk.

3.3 Quantile analysis

In this subsection, we further analyze the annuity and tontine payoffs by considering the 0.01-, 0.99- and 0.5-quantile (i.e. the median) of the tontine and annuity payoffs (14) and (15). The setting and the parameters are similar to the previous subsections and

[7] Note that, in Figure 5, the time axis ranges only from 0 to 50 instead of 55, because the denominators in both CVs come extremely close to zero for $t > 50$, which delivers unstable results. The probability for a policyholder of surviving 50 years is $\mathbb{E}\left[{}_{50}p_x^{1-\epsilon}\right] = 1.90\cdot 10^{-6}$, i.e. it would also not be an unrealistic assumption to set the highest maximum age at 115.

we rely again on the optimal tontine payoff $d^{F*}(t)$. Figure 6 provides the corresponding quantiles of the annuity and tontine payoffs. As the plots for the different annuity fees $\Delta_0 \in \{4, 7, 10, 14\}$ look rather similar (the only difference is that the lower fees shift the curves upwards), here we only provide the plot for $\Delta_0 = 14$. We observe that the tontine payoff is basically stable over time, whereas the annuity payoff allows for a lot of variability upwards and downwards. The only time when the tontine payoff is not fully stable is after 40 to 50 years, depending on the standard deviation (when the policyholders are aged 105). Naturally, as the standard deviation increases, the quantile range of the payoffs of both products also becomes larger. For the annuity, this increase in the quantile range affects all time points up to time 50, whereas in the tontine, only the time range 40 to 50 is affected.

These observations confirm the conclusions drawn from the previous subsection that the annuity payoff is much more volatile over time and, consequently, harder to manage from a risk manager's perspective.

3.4 Analysis of the reserves

Another question of interest to the insurer is how high the reserves need to be for different products. In the tontine design introduced by Milevsky and Salisbury (2015) which we consider in this paper, the insurer promises the payments $nd^F(t)$ to the pool of policyholders. Hence, the insurer has to set up reserves for the tontine and, of course, also for the annuity. To determine the best-estimate reserve at a time $t > 0$, we follow Börger (2010) and assume that the development of the mortality follows best-estimate assumptions until time t. To be precise, the number of living policyholders at time t is given by $n_t p_x$. Multiplying the per-contract reserves of the annuity and tontine given in Chen et al. (2019) by n, we obtain the following expressions for the overall best-estimate reserve of the tontine and the annuity:

$$
\begin{aligned}
{}_tV_x^T &= n_t p_x \int_t^\infty e^{-r(s-t)} \int_{-\infty}^1 \left(1 - (1 - {}_{s-t}p_{x+t})^n\right) f_\epsilon(\varphi) \mathrm{d}\varphi \cdot d^F(s) \,\mathrm{d}s, \\
{}_tV_x^A &= n_t p_x \int_t^\infty e^{-r(s-t)} {}_{s-t}p_{x+t} \cdot m_\epsilon(-\ln {}_{s-t}p_{x+t}) c(s) \,\mathrm{d}s.
\end{aligned}
$$

In Figure 7, we show the reserves of the annuity and the tontine over time using the optimal payoff $d^{F*}(t)$. For each fee Δ_0, we observe that the annuity requires a higher reserve than the tontine at all times except $t = 0$, which is due to the fact that the fee charged for the tontine is lower than that of the annuity. Additionally, we observe that the reserves of the annuity and tontine decrease over time, with the reserve of the tontine decreasing more rapidly than the reserve of the annuity. Different volatility levels of the shock ϵ do not seem to have a major impact on the reserves of the retirement product. One thing that we can observe is that the reserve at time 0 (i.e. the initial value) of the tontine exceeds that of the annuity to a greater extent if a higher volatility is present. This is a direct result of the lower tontine indifference fee obtained under higher volatility levels (see Figure 3).

4 Conclusions

We have analyzed the impact of fees in modern tontines, where two different fee structures are considered: a (possibly) time-varying proportional fee of the payoff charged over time and a fixed, up-front fee. We find that a constant fraction of the payoff being deducted over time yields the same level of expected utility to policyholders as a fixed- up-front fee. A time-decreasing proportional fee deducted over time delivers an only slightly lower level of expected utility than the fixed up-front fee. Hence, tontine managers can charge higher costs in the younger ages, when individuals are less likely to suffer from the lack of financial resources compared to older ages, when individuals generally need a higher income to cover care expenditure. Our framework also allows for a comparison with annuities and reveals that tontine managers could charge a little less than annuity managers even for highly risk-averse individuals and consumers would still prefer tontine schemes, the advantage being that tontine managers do not need to be concerned about longevity risk.

From a mean-variance analysis of the payoffs of tontines and annuities from the insurer's perspective, we find that the variance of the annuity payoff significantly exceeds that of the tontine payoff for a greater part of the retirement phase. An analysis of the coefficients of variation of the annuity and tontine payoff and a quantile analysis confirm the substantially reduced risk of the tontine products to the insurer. Furthermore, under the indifference fee (which makes policyholders indifferent to choosing between a tontine or an annuity), we find that reserves of tontines are lower than those of annuities at practically all times during the retirement phase. These results indicate that insurers might generate sizeable and secure profits from issuing fee-based tontines.

As an extension to this article, the problem could be reformulated, possibly as a game in which the insurer chooses the fee and the participants choose payment streams. Determining the optimal fee structure would be an interesting question for future research.

References

Agnew, J. R., Anderson, L. R., Gerlach, J. R., and Szykman, L. R. (2008). Who chooses annuities? An experimental investigation of the role of gender, framing, and defaults. *American Economic Review*, 98(2):418–22.

Bernhardt, T. and Donnelly, C. (2019). Modern tontine with bequest: innovation in pooled annuity products. *Insurance: Mathematics and Economics*, 86:168–188.

Beshears, J., Choi, J. J., Laibson, D., Madrian, B. C., and Zeldes, S. P. (2014). What makes annuitization more appealing? *Journal of Public Economics*, 116:2–16.

Börger, M. (2010). Deterministic shock vs. stochastic value-at-risk – an analysis of the Solvency II standard model approach to longevity risk. *Blätter der DGVFM*, 31(2):225–259.

Brown, J. R., Kapteyn, A., Luttmer, E. F., and Mitchell, O. S. (2017). Cognitive constraints on valuing annuities. *Journal of the European Economic Association*, 15(2):429–462.

Chen, A., Hieber, P., and Klein, J. K. (2019). Tonuity: A novel individual-oriented retirement plan. *ASTIN Bulletin: The Journal of the IAA*, 49(1):5–30.

Chen, A. and Rach, M. (2019). Options on tontines: An innovative way of combining tontines and annuities. *Insurance: Mathematics and Economics*, 89:182–192.

Chen, A., Rach, M., and Sehner, T. (2020). On the optimal combination of annuities and tontines. *ASTIN Bulletin: The Journal of the IAA*, 50(1):95–129.

Cui, Z., Feng, R., and MacKay, A. (2017). Variable annuities with VIX-linked fee structure under a Heston-type stochastic volatility model. *North American Actuarial Journal*, 21(3):458–483.

Donnelly, C., Guillen, M., and Nielsen, J. P. (2013). Exchanging uncertain mortality for a cost. *Insurance: Mathematics and Economics*, 52(1):65–76.

Donnelly, C., Guillen, M., and Nielsen, J. P. (2014). Bringing cost transparency to the life annuity market. *Insurance: Mathematics and Economics*, 56:14–27.

Forman, J. B. and Sabin, M. J. (2015). Tontine pensions. *University of Pennsylvania Law Review*, 173(3):755–831.

Forman, J. B. and Sabin, M. J. (2016). Survivor funds. *Pace Law Review*, 37(1):204–291.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical transactions of the Royal Society of London*, 115:513–583.

Hanewald, K., Piggott, J., and Sherris, M. (2013). Individual post-retirement longevity risk management under systematic mortality risk. *Insurance: Mathematics and Economics*, 52(1):87–97.

Hu, W.-Y. and Scott, J. S. (2007). Behavioral obstacles in the annuity market. *Financial Analysts Journal*, 63(6):71–82.

Inkmann, J., Lopes, P., and Michaelides, A. (2010). How deep is the annuity market participation puzzle? *The Review of Financial Studies*, 24(1):279–319.

Li, Y. and Rothschild, C. (2020). Selection and redistribution in the Irish tontines of 1773, 1775, and 1777. *Journal of Risk and Insurance*, 87(3):719–750.

Lin, Y. and Cox, S. H. (2005). Securitization of mortality risks in life annuities. *Journal of Risk and Insurance*, 72(2):227–252.

Lockwood, L. M. (2012). Bequest motives and the annuity puzzle. *Review of Economic Dynamics*, 15(2):226–243.

MacKay, A., Augustyniak, M., Bernard, C., and Hardy, M. R. (2017). Risk management of policyholder behavior in equity-linked life insurance. *Journal of Risk and Insurance*, 84(2):661–690.

Milevsky, M. A. (2020). Calibrating Gompertz in reverse: What is your longevity-risk-adjusted global age? *Insurance: Mathematics and Economics*, 92:147–161.

Milevsky, M. A. and Salisbury, T. S. (2015). Optimal retirement income tontines. *Insurance: Mathematics and Economics*, 64:91–105.

Peijnenburg, K., Nijman, T., and Werker, B. J. (2016). The annuity puzzle remains a puzzle. *Journal of Economic Dynamics and Control*, 70:18–35.

Piggott, J., Valdez, E. A., and Detzel, B. (2005). The simple analytics of a pooled annuity fund. *Journal of Risk and Insurance*, 72(3):497–520.

Pitacco, E., Denuit, M., Haberman, S., and Olivieri, A. (2009). *Modelling longevity dynamics for pensions and annuity business*. Oxford University Press.

Qiao, C. and Sherris, M. (2013). Managing systematic mortality risk with group self-pooling and annuitization schemes. *Journal of Risk and Insurance*, 80(4):949–974.

Sabin, M. J. (2010). Fair tontine annuity. *Available at SSRN:* https://ssrn.com/abstract=1579932.

Stamos, M. Z. (2008). Optimal consumption and portfolio choice for pooled annuity funds. *Insurance: Mathematics and Economics*, 43(1):56–68.

Statista (2019). Average risk-free investment rate in Germany 2015–2019. Website. Available online on https://www.statista.com/statistics/885774/average-risk-free-rate-germany/; accessed on October 23, 2019.

Valdez, E. A., Piggott, J., and Wang, L. (2006). Demand and adverse selection in a pooled annuity fund. *Insurance: Mathematics and Economics*, 39(2):251–266.

Weir, D. R. (1989). Tontines, public finance, and revolution in France and England, 1688–1789. *The Journal of Economic History*, 49(1):95–124.

Yaari, M. E. (1965). Uncertain lifetime, life insurance, and the theory of the consumer. *The Review of Economic Studies*, 32(2):137–150.

Zhang, J., Purcal, S., and Wei, J. (2020). Optimal life insurance and annuity demand under hyperbolic discounting when bequests are luxury goods. *Insurance: Mathematics and Economics*.

A Proofs

A.1 Proof of Theorem 2.2

For a similar proof see also Milevsky and Salisbury (2015) and Chen et al. (2019). The Lagrangian function is given by:

$$\begin{aligned}\mathcal{L} = &\int_0^\infty e^{-\rho t}\kappa_{n,\gamma,\epsilon}({}_tp_x)u((1-\alpha(t))d^V(t))\mathrm{d}t\\ &+\lambda_V\left(v-\int_0^\infty e^{-rt}\int_{-\infty}^1\left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right)\mathrm{d}\varphi\, d^V(t)\,\mathrm{d}t\right).\end{aligned}$$

The first-order condition is:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial d^V(t)} &= e^{-\rho t}\kappa_{n,\gamma,\epsilon}({}_tp_x)u'((1-\alpha(t))d^V(t))(1-\alpha(t))-\lambda_V e^{-rt}\int_{-\infty}^1\left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right)\mathrm{d}\varphi=0\\ \Leftrightarrow\ d^{V*}(t) &= \frac{1}{1-\alpha(t)}\left(\frac{e^{(\rho-r)t}\lambda_V\int_{-\infty}^1\left(1-(1-{}_tp_x^{1-\varphi})^n\right)\mathrm{d}\varphi}{(1-\alpha(t))\kappa_{n,\gamma,\epsilon}({}_tp_x)}\right)^{-1/\gamma}.\end{aligned}$$

Plugging these terms into the budget constraint delivers an explicit formula for the Lagrangian multiplier:

$$v=\int_0^\infty e^{-rt}\int_{-\infty}^1\left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right)\mathrm{d}\varphi\,\frac{1}{1-\alpha(t)}\left(\frac{e^{(\rho-r)t}\lambda_V\int_{-\infty}^1\left(1-(1-{}_tp_x^{1-\varphi})^n\right)\mathrm{d}\varphi}{(1-\alpha(t))\kappa_{n,\gamma,\epsilon}({}_tp_x)}\right)^{-\frac{1}{\gamma}}\mathrm{d}t.$$

This delivers the following formula for the Lagrangian multiplier:

$$\begin{aligned}\lambda_V = \Bigg(&\frac{1}{v}\int_0^\infty e^{-rt}\int_{-\infty}^1\left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right)\mathrm{d}\varphi\\ &\frac{1}{1-\alpha(t)}\left(\frac{e^{(\rho-r)t}\int_{-\infty}^1\left(1-(1-{}_tp_x^{1-\varphi})^n\right)\mathrm{d}\varphi}{(1-\alpha(t))\kappa_{n,\gamma,\epsilon}({}_tp_x)}\right)^{-\frac{1}{\gamma}}\mathrm{d}t\Bigg)^{\gamma}.\end{aligned}$$

The overall expected discounted lifetime utility is then given by:

$$\begin{aligned}&\mathbb{E}\left[\int_0^{\zeta_\epsilon} e^{-\rho t}u\left(\frac{n(1-\alpha(t))d^{V*}(t)}{N_\epsilon(t)}\right)\mathrm{d}t\right]\\ &=\int_0^\infty e^{-\rho t}\kappa_{n,\gamma,\epsilon}({}_tp_x)u\left(\left(\frac{e^{(\rho-r)t}\lambda_V\int_{-\infty}^1\left(1-(1-{}_tp_x^{1-\varphi})^n\right)\mathrm{d}\varphi}{(1-\alpha(t))\kappa_{n,\gamma,\epsilon}({}_tp_x)}\right)^{-1/\gamma}\right)\mathrm{d}t\end{aligned}$$

$$
\begin{aligned}
&= \frac{\lambda_V^{1-\frac{1}{\gamma}}}{1-\gamma}\left(\int_0^\infty e^{-\rho t}\kappa_{n,\gamma,\epsilon}({}_tp_x) \left(\frac{e^{(\rho-r)t}\int_{-\infty}^1 \left(1-(1-{}_tp_x^{1-\varphi})^n\right) \mathrm{d}\varphi}{(1-\alpha(t))\kappa_{n,\gamma,\epsilon}({}_tp_x)} \right)^{1-\frac{1}{\gamma}} \mathrm{d}t \right) \\
&= \frac{\lambda_V^{1-\frac{1}{\gamma}}}{1-\gamma}\left(\int_0^\infty e^{-rt}\int_{-\infty}^1 \left(1-\left(1-{}_tp_x^{1-\varphi}\right)^n\right) \mathrm{d}\varphi \frac{1}{1-\alpha(t)} \left(\frac{e^{(\rho-r)t}\int_{-\infty}^1 \left(1-(1-{}_tp_x^{1-\varphi})^n\right) \mathrm{d}\varphi}{(1-\alpha(t))\kappa_{n,\gamma,\epsilon}({}_tp_x)} \right)^{-\frac{1}{\gamma}} \mathrm{d}t \right) \\
&= \frac{\lambda_V^{1-\frac{1}{\gamma}}}{1-\gamma}\cdot v \cdot \lambda_V^{\frac{1}{\gamma}} \\
&= \frac{\lambda_V}{1-\gamma} v.
\end{aligned}
$$

$\square$

(a) $\sigma = 0.0814$

(b) $\sigma = 2 \cdot 0.0814$

(c) $\sigma = 0.5 \cdot 0.0814$

Figure 3: Indifference fee for the tontine dependent on the relative risk aversion γ. The remaining parameters are chosen as in Table 1.

(a) $\sigma = 0.0814$

(b) $\sigma = 2 \cdot 0.0814$

(c) $\sigma = 0.5 \cdot 0.0814$

Figure 4: Variance of the annuity and tontine payoffs from the insurer's perspective over time. The fee charged for the tontine is the indifference fee (see Figure 3). The remaining parameters are chosen as in Table 1.

(a) $\sigma = 0.0814$

(b) $\sigma = 2 \cdot 0.0814$

(c) $\sigma = 0.5 \cdot 0.0814$

Figure 5: Difference of the coefficients of variation of the annuity and tontine payoffs $\mathrm{CV}_A(t) - \mathrm{CV}_T(t)$ from the insurer's perspective over time. The remaining parameters are chosen as in Table 1.

(a) Tontine, $\sigma = 0.0814$

(b) Annuity, $\sigma = 0.0814$

(c) Tontine, $\sigma = 2 \cdot 0.0814$

(d) Annuity, $\sigma = 2 \cdot 0.0814$

(e) Tontine, $\sigma = 0.5 \cdot 0.0814$

(f) Annuity, $\sigma = 0.5 \cdot 0.0814$

Figure 6: Quantiles of the annuity and tontine payoffs from the insurer's perspective

(a) $\sigma = 0.0814$

(b) $\sigma = 2 \cdot 0.0814$

(c) $\sigma = 0.5 \cdot 0.0814$

Figure 7: Reserves of the annuity and tontine over time. The fee charged for the tontine is the indifference fee. The remaining parameters are chosen as in Table 1.