The Modern Tontine: An Innovative Instrument for Longevity Risk Management in an Aging Society

Jan-Hendrik Weinert, Helmut Gründl

This version: July 6, 2017

Abstract

The changing social, financial and regulatory frameworks, such as an increasingly aging society, the current low interest rate environment, as well as the implementation of Solvency II, lead to the search for new product forms for private pension provision. In order to address the various issues, these new product forms should reduce or avoid investment guarantees and risks stemming from longevity, still provide reliable insurance benefits and simultaneously take account of the increasing financial resources required for very high ages. In this context, we examine whether a historical concept of insurance, the tontine, entails enough innovative potential to extend and improve the prevailing privately funded pension solutions in a modern way. The tontine basically generates an age-increasing cash flow, which can help to match the increasing financing needs at old ages. However, the tontine generates volatile cash flows, so that the insurance character of the tontine cannot be guaranteed in every situation. We show that partial tontinization of retirement wealth can serve as a reliable supplement to existing pension products.

Keywords: Life Insurance, Tontines, Annuities, Asset Allocation, Retirement Welfare, Aging Society
JEL Classification: D14, D91, E21, G22, J11, J14
1 Introduction

Through changes in social, financial and regulatory conditions, both life insurance policyholders and life insurers are facing big challenges. One of the large social challenges in most of the western countries is the demographic change caused by declining birth rates and an increasing longevity of the population\(^1\). Therefore, the retirement quotient rises; as a result pay-as-you-go retirement systems come under pressure while funded retirement products gain relevance. In addition, the liquidity need increases for elderly people, which is mainly driven by increasing medical expenses at old ages. According to a study by Standard Life (2013), the liquidity need of persons older than 85 years is six times higher than for persons below 65 years of age. As a consequence, the demand for funded retirement products that help to diminish the pension provision gap in an aging society increases.

Without doubt, life and health insurers already offer appropriate traditional pension, health insurance and life-care products, sometimes combined to enhanced annuities. Yet insurers are exposed to changing financial and regulatory conditions. Traditional pension products often entail minimum return guarantees, which providers try to ensure by investing extensively in fixed-income securities\(^2\). However, the current low interest environment clearly shows large solvency risks caused by the issuance of lifetime guarantees. The possible way out of this problem, i.e., to invest extensively in more profitable asset classes like stocks, is however restricted due to its higher risk and its limited ability to cover granted guarantees. Furthermore, providers of pension products are exposed to the longevity risk of their customers, which can only be partially passed on to them. Although longevity bonds provide insurance companies the opportunity to hedge longevity risk, the issuance of those products has failed in practice due to moral hazard problems\(^3\). Therefore, other solutions to deal with longevity risk are needed.

In the European Union and many other parts of the world, the regulatory conditions for providers of private pension products change substantially with the introduction of risk-based solvency regulation. The market-consistent valuation of investments as well as of technical provisions immediately reveals the addressed high risks of traditional life and health insurance products involved and therefore can cause severe financial imbalance for life insurers. In a traditional insurance context, managing these risks requires considerable equity capital backing or other

---

\(^1\) See for example Statistisches Bundesamt (2015) for a prognosis for the demographic change until the year 2060 in Germany and United States Census Bureau (2016) for a prognosis for the demographic change until the year 2050 in USA.


\(^3\) See MacMinn and Brockett (2017).
comprehensive risk management activities like re-insurance or securitization. Such risk management ultimately has to be funded by higher insurance premiums, which might make pension planning unattractive.

The changes in social, financial and regulatory conditions therefore lead to the quest for innovative instruments for private pension planning. A product innovation should optimally reduce investment guarantees and risks related to longevity, and nevertheless be able to provide reliable insurance performance. At the same time it should meet the concerns of increasing liquidity needs at old ages.

Against this backdrop, we transfer the idea of the historic tontine to a modern context and analyze whether it can help to solve the aforementioned problems. A tontine provides a mortality driven, age-increasing payout structure. Although an insurer can easily replicate such a payout structure, the tontine has the big advantage of its simplicity and low costs. While traditional insurance products entail large safety and administrative cost loadings\(^4\), a tontine can be offered at low additional costs\(^5\). This is because a tontine is a simple redistribution mechanism of the invested funds without guarantees and the need for an active management. The investment strategy of the tontinized wealth can be decided on an individual basis according to the individual risk aversion, without the issuance of guarantees. Due to the linkage between tontine returns and individual survival prospects, tontine payments are very low in younger years and increase strongly for very high ages. As we observe an increasing liquidity need at old ages, we can show that the tontine can be an appropriate instrument to serve as financial protection in the late retirement years, going along with relatively small investment volumes.

We show that a modern tontine can be an appropriate, cost-efficient complement to existing privately funded pension solutions with the ability to improve policyholders’ welfare by serving the increasing monetary demand at old ages. To this end, we take the perspective of a tontine holder who holds a tontine for pension planning purposes. We compare the tontine and its benefit structure with that of a conventional pension annuity and derive implications for the individual tontine demand. Although tontines do not provide for any investment guarantees, their features make it possible to generate cash flows that cover the age-specific needs of retirees. In an aging society, tontines can thus become an interesting instrument of old-age provision, complementary to conventional insurance products.

For assessing the advantages and disadvantages of tontines, their age-increasing payments with

\(^4\) According to Bundesanstalt für Finanzdienstleistungsaufsicht (2014) the average acquisition and administrative costs for German life insurers are 10.7% of the gross premiums.

\(^5\) See Weinert (2017a) for a cost analysis of tontines.
the ability to finance the also increasing care costs have so far not been investigated. This is the gap in the literature which we try to bridge with our contribution.

We aim at assessing the effects of tontinizing some fraction of the individual retirement wealth on the individual lifetime utility, considering an increasing liquidity need at old ages. We illustrate under which conditions there are appropriate incentives for individuals to hold some fraction of retirement wealth in tontines compared to complete annuitization.

The remainder of the article is organized as follows: Section 2 introduces the general concept of tontines. Section 3 reviews the relevant literature on tontines and increasing liquidity needs at old ages. Section 4 introduces our model framework specifying the underlying mortality dynamics, the tontine model as well as the old-age liquidity need curve and the valuation of annuities. Finally, we propose a Cumulative Prospect Theory based valuation of lifetime utility of tontines and annuities. In Section 5, we first describe the data and the calibration we adopt and provide findings for the optimal individual wealth allocation and discuss our results. Section 6 provides implications and our conclusion.

2 Tontines

The Italian Lorenzo de Tonti invented a product to consolidate the French public-sector deficit in the early 1650s, which was introduced in 1689\(^6\). His ideas were based on the pooling of persons by considering their mortality risk. The innovation was that, in exchange for a lump sum payment to the French government, one received the right to a yearly, lifelong pension, which increased over time because the yields were distributed among a smaller amount of surviving beneficiaries. The last survivor thus received the pensions of all others who died before. As Manes (1932) notes, the valuation of the original tontine was inaccurate, retirees were grouped in broad age classes, and so the contract terms were not fair in an actuarial sense. In this article, we build upon a fair tontine based on Sabin (2010) that allows participants to be of any age, of any gender, and to invest a desired amount of money\(^7\) in the tontine. Furthermore, the tontine is revolving, which means that new members can join the tontine at any age and take on the position of deceased members. Apart from that it is not allowed to leave the tontine before passing away. The tontine is a fair lottery for every member. Expected individual tontine payments equal the individual investment in the tontine, yielding an unconditional expected profit of zero. Expected tontine payments depend on the individual stake in the tontine and on the individual survival probability.

---

\(^6\) See McKeever (2009) for an overview of the history of tontines.

\(^7\) According to Sabin (2010) a fair tontine is a tontine in which the distribution to surviving participants is made in unequal portions according to a plan that provides each participant with a fair bet.
On the one hand, if a member dies, he or she loses the entire stake, while, on the other, if he or she survives he or she receives some fraction of the stake of deceased members. To be fair, expected gains and losses are equal in each period. Because the survival probability declines with age and tontine payments are only paid if one survives, the probability of receiving tontine payments decreases. To counterbalance the otherwise induced reduction in expected tontine payments, the size of the payments one receives has to increase. Through this mortality-driven feature, the expected conditional tontine payments increase with age. Mortality, therefore, is the crucial factor for determining the tontine benefit structure. For example, a man born in 1981 can expect to live 84 years, while a man born in 2013 has an increased life expectancy of 89 years. This difference of 5 years translates directly into different benefit patterns, especially at old ages. Furthermore, the whole composition of the demographic structure of the tontine members changes on the basis of the population mortality, which also impacts the actual benefit structure. Therefore, it is important to model and forecast the development of mortality and demographic structure of the tontine members. We use the one-factor model by Lee and Carter (1992) to forecast mortality, which is the standard approach to model mortality rates. Recent studies which use the Lee-Carter model are, for example, Renshaw and Haberman (2003) and Renshaw and Haberman (2008). The Lee-Carter model only considers cohort effects, i.e. changes of mortality within a cohort. However, Willets (2004) empirically substantiates cohort effects as characteristic of mortality dynamics.

While in a traditional annuity\(^8\) longevity risk is transferred from the insured to the insurer (and covered by its risk management instruments), in a tontine the risk that a single participant might live longer than expected is completely borne and shared by the other tontine holders who in this case receive less cash flows than expected. Therefore, no equity capital backing is needed to cover longevity risk, and the tontine can be offered without a risk-cost loading. However, the tontine has the disadvantage that because individual shares in the tontine as well as times of death of tontine members are random, both amount and timing of tontine payments are uncertain.

Because the tontine members carry, pool and share the total risk among each other while the offering provider does not bear it, a tontine can be offered at a cheaper price than a comparable traditional life insurance product. Additionally, it generates age-increasing benefits and is therefore able to meet increasing monetary requirements at old ages. Moreover, due to the absence of guarantees, the tontine enables participation in stock market developments. However,

\(^8\)In the following we use “annuity” synonymously for the traditional life insurance product.
the tontine generates volatile payments, which means that the insurance character of a tontine might not be ensured in every situation.

So far, tontines have been considered an alternative or historic predecessor of traditional pension insurance. However, tontines were considered to be inferior compared to traditional pension insurance, because the latter provides less volatile payments for an equal expected return\(^9\). In addition, one could think of potential moral hazard problems among tontine members. However, Milevsky (2015) found no evidence that such phenomena have ever occurred in practice.

### 3 Literature Review

In the literature there are several contributions on the evaluation of longevity-linked securities from a policyholder perspective. Sabin (2010) designs a fairly priced tontine with regard to age, gender and entry date that is equivalent to a common annuity scheme. His results exhibit a more cost efficient payout pattern compared to a typical insurer-provided annuity not just on average, but for virtually every member who lived more than just a few years. Cairns et al. (2008) summarize various instruments that deal with longevity risks of insurance companies. Their review presents various concepts of longevity-linked securities, such as mortality swaps and longevity bonds that may serve as a hedging instrument against longevity. However, the products introduced had only moderate success due to low market acceptance and the lack of a perfect hedge. Milevsky and Salisbury (2015) derive an optimal tontine design by accounting for sensitivity of both the tontine size and the longevity risk aversion for each tontine member. By doing so, they raise the question whether an optimally designed tontine with low capital requirements for the sponsor will gain more attention in times of risk-based capital standards and conclude that, due to higher volatility of the payments, the tontine provides a lower utility than a traditional life annuity. Forman and Sabin (2014) construct a fair transfer plan (FTP) to guarantee a fair bet for all participating investors of a tontine by accounting for each age, life expectancy and investment level. They show that a fairly designed tontine is superior to defined benefit plans in terms of funding and sponsoring of the pension system. They illustrate that a fairly developed tontine model would improve the situation of pension providers while serving the retirement income demand of the tontine participants. Milevsky and Salisbury (2016) follow a similar approach and allow combining heterogeneous cohorts into one pool. The tontine is made fair by allocating tontine shares at either a premium or a discount depending on the individual characteristics of the tontinists. Chen et al. (2016) propose a retirement product called tomuity,\(^9\)

in which a tontine and an annuity are combined to serve best the policyholder needs in a way that the policyholder owns a tontine in the early years of retirement which then is converted into a deferred annuity. As the tontine pays out relatively stable payouts in the early years of retirement at low costs because of the absence of guarantees, its payouts are highly volatile in the later years of retirement. Therefore, the tontine is converted into a stable deferred annuity at an optimal switching time. Weinert (2017b) extends the prevailing tontine scheme by the possibility of a premature surrender and determines the fair surrender value. Furthermore, there is a strand of literature on pooled annuity funds (Piggott et al. (2005), Donnelly et al. (2013), Donnelly et al. (2014) and Donnelly (2015)). Pooled annuity funds are very similar to tontines in a way that payouts are adjusted immediately if life expectancy changes.

4 Model Framework

We first model mortality dynamics in Germany for the upcoming decades and derive in a second step possible population pyramids. These, in turn, are the basis for the composition of the fair revolving tontine. We then estimate the risk-free benefit profile of a standard annuity and compare it to the risky benefit profile of a tontine. We assume that each individual $i$ has initial wealth endowment $W_i$ and that there is no further source of income in the future. $W_i$ can be seen as the sum of both discounted future earnings until retirement and savings up to the investment date. $W_i$ will be completely converted into pension installments. For the expected tontine benefits, we provide a closed-form solution while we determine the realized benefits by performing a Monte Carlo simulation. We then analyze to what extent tontine and traditional annuity are able to satisfy an empirically estimated, increasing old-age liquidity need function for different settings. Furthermore, we estimate an optimal portfolio consisting of annuity and tontine which maximizes expected utility according to a Cumulative Prospect Theory framework. We provide results for different demographic scenarios and mortality dynamics and show the capability of tontines as instruments for retirement planning from a policyholder perspective. Finally we incorporate subjective beliefs about individual mortality to account for different perceptions about individual life expectations, which leads to a changing optimal asset allocation for retirement planning.

4.1 Mortality Model

In a first step, we project mortality rates for a forecast horizon of $t = 1, \ldots, T$ years. Our starting point is the one-factor-model for estimating mortality rates by Lee and Carter (1992).
According to the Lee-Carter Model the one-year death probability $q_{x,t}$ of a person aged $x$ in year $t$ is specified as

$$\ln (q_{x,t}) = a_x + b_x \cdot k_t + \varepsilon_{x,t} \Leftrightarrow q_{x,t} = e^{a_x + b_x \cdot k_t + \varepsilon_{x,t}} \quad (1)$$

where $a_x$ and $b_x$ are time constant parameters for a male aged $x$ that determine the shape and the sensitivity of the mortality rate to changes in $k_t$, which is a time-varying parameter that captures the changes in the mortality rates over time. $\varepsilon_{x,t}$ is an error term with mean 0 and constant variance. As originally proposed by Lee and Carter (1992), the estimation of the time varying parameter $k_t$ can be performed by fitting a standard ARIMA model using standard time series analysis techniques. The $ARIMA(p, d, q)$ process is given by

$$k_t = (\alpha_0 + \alpha_1 k_{t-1} + \alpha_2 k_{t-2} + \ldots + \alpha_p k_{t-p} + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots + \beta_q \varepsilon_{t-q}) + \varepsilon_t = \tilde{k}_t + \varepsilon_t \quad (2)$$

in which the error term is normally distributed $\varepsilon \sim N(0, \sigma_k)$.

### 4.2 Demographic Structure

Based on the predicted mortality rates $q_{y,t}$ for the one-year death probability of a woman aged $y$ in period $t$, and $q_{x,t}$ for the one-year death probability of a man aged $x$ in period $t$, we determine the demographic structure of an economy in every period $t$. $Q_{y,t}$ ($Q_{x,t}$) is the total quantity of female (male) persons of a cohort aged $y$ ($x$) at time $t$. Equation (3) shows the updating process. Newborns, or persons in their first year of life ($y, x = 1$) are determined by the sum of the age-specific fertility rate $AGZ_y$ times the quantity of females of the respective age $y$ in each period $t$, which is weighted by the fraction of newborn females $f_0$ and males $m_0 = 1 - f_0$. From the second year of being alive, the number of persons is the probability to survive one year of someone who was one year younger in the year before times the number of persons who were one year younger in the year before:

$$Q_{y,t} = \begin{cases} f_0 \cdot \sum_{y=2}^{\Omega} Q_{y,t} \cdot AGZ_y & \text{for } y, x = 1 \\ Q_{y-1,t-1} \cdot (1 - q_{y-1,t-1}) & \text{for } y, x = 2 \ldots \Omega \end{cases}$$

$$Q_{x,t} = \begin{cases} m_0 \cdot \sum_{y=2}^{\Omega} Q_{y,t} \cdot AGZ_y & \text{for } y, x = 1 \\ Q_{x-1,t-1} \cdot (1 - q_{x-1,t-1}) & \text{for } y, x = 2 \ldots \Omega \end{cases} \quad (3)$$
We determine these quantities for all cohorts \( y, x = 1, \ldots, \Omega \) in all periods \( t = 1, \ldots, T \) for males and females to estimate the corresponding population pyramids. \( F_{y,t} \) (\( F_{x,t} \)) in Equation (4) shows the Cumulative Distribution Function (CDF) of a person aged \( y \) (\( x \)) in each \( t \), where 

\[
\kappa_{y,t} = \frac{Q_{y,t}}{\sum_{y=1}^{\Omega} Q_{y,t}} \quad \text{and} \quad \kappa_{x,t} = \frac{Q_{x,t}}{\sum_{x=1}^{\Omega} Q_{x,t}} \quad (\text{for } y, x = 1, \ldots, \Omega)
\]

are the fractions of each cohort of females and males of the total female and male population in each period \( t \).

\[
F_{y,t} = \begin{cases} 
0 & : y < 0 \\
\sum_{y=1}^{\Omega} \kappa_{y,t} & : 0 \leq y < \Omega \\
1 & : y > \Omega
\end{cases} \quad \quad F_{x,t} = \begin{cases} 
0 & : x < 0 \\
\sum_{x=1}^{\Omega} \kappa_{x,t} & : 0 \leq x < \Omega \\
1 & : x > \Omega
\end{cases}
\]

(4)

The fractions of females and males aged \( y \) and \( x \) of the total population in each \( t \) are \( f_{y,t} = \frac{\kappa_{y,t}}{\kappa_{y,t} + \kappa_{x,t}} \) and \( m_{x,t} = 1 - f_{y,t} \). The predicted mortality rates and demographic structures are the basis for calculating the tontine composition and tontine benefits, and are the starting point to analyze the impact of the demographic development on the tontine and its possible application as an alternative retirement planning product.

### 4.3 Tontine Model

We model a fair revolving tontine based on Sabin (2010) that allows tontine participants to be of any gender and age, and to invest a desired one-time initial amount of money \( B_i \) at tontine entrance. \( B_i \) then is tied in the tontine and cannot be withdrawn before the tontine member’s death. Furthermore, there is no possibility to inject additional capital for any individual in future periods. While the original model considers infinitesimal points in time, resulting in only one member being able to die at one point in time, we adjust the model to a yearly time frame allowing for multiple deaths. We further assume the number of the tontine members \( N \) as fixed: every time a participant dies, the tontine is refilled to \( N \). A new entrant \( i \) is randomly drawn from the period-corresponding demographic structure. We assume entrants at least to be of a certain age \( y_i, x_i \) and also assume an upper limit of entering the tontine of age \( \overline{y}, \overline{x} \) so 

\[
0 \leq y_i, x_i < \overline{y}, \overline{x} \leq \Omega.
\]

The random age of an individual \( i \) entering the tontine in \( t \) is expressed by

\[
y_{E,i,t}, \quad x_{E,i,t} = F_{t}^{-1}(z)
\]

(5)

where \( F_t^{-1} \) is the inverse function of \( F_t \) with \( Pr(y) = f_{y,t} \) and \( Pr(x) = m_{x,t} \) and with \( z \in \left( F_t^{-1}(y), F_t^{-1}(\overline{y}) \right) \) or \( z \in \left( F_t^{-1}(x), F_t^{-1}(\overline{x}) \right) \) where \( z \) is uniformly distributed on \( U(F_t^{-1}(y), F_t^{-1}(\overline{y})) \)
or $\mathcal{U} \left( F_{t}^{-1}(x), F_{t}^{-1}(x) \right)$. We further assume the establishment of the tontine in $t = 0$ and refrain from investing the tied capital to streamline the model and to be able to quantify solely the interrelation of mortality benefits and demographic change. For better readability, we denote the one-year death probability of individual $i$ with the beforehand assigned characteristics as $q_{i,t}$ in $t$. The index $i$ allows to identify each individual with its specific characteristics in each period. Furthermore we denote the age of a person as $x$ in the following, irrespective of the gender.

Let $\{A_{i,t}\}$ be the event that $i$ dies in $t$ with $P\left(A_{i,t}\right) = q_{i,t}$ and $\{A_{i,t}^{c}\}$ be the event that $i$ survives in $t$ with $P\left(A_{i,t}^{c}\right) = 1 - q_{i,t}$. Let $\{A_{0,t}\}$ be the event that at least someone dies in $t$ and $\{A_{0,t}\}$ be the event that no one dies in $t$. Using the inclusion-exclusion principle\(^{12}\), the probability that at least someone dies in $t$ is

$$P\left(A_{0,t}^{c}\right) = P\left(\bigcup_{i=1}^{N} A_{i,t}\right) = \sum_{j=1}^{N} \left((-1)^{j+1} \sum_{I \subseteq \{1, \ldots, N\}} P\left(\bigcap_{i \in I} A_{i,t}\right) \right),$$

and the probability that no one dies in $t$ is

$$P\left(A_{0,t}\right) = 1 - P\left(\bigcup_{i=1}^{N} A_{i,t}\right).$$

$\{A_{k,t} | A_{0,t}^{c}\}$ denotes the event that $k$ dies in $t$ conditioned that at least someone dies in $t$. Using the law of total probability yields for the probability that $k$ dies in $t$ conditioned that at least someone dies in $t$, $\rho_{k,t}$

$$\rho_{k,t} = P\left(A_{k,t} | A_{0,t}^{c}\right) = \frac{P\left(A_{k,t}\right)}{P\left(A_{0,t}^{c}\right)} = \frac{q_{k,t}}{\sum_{j=1}^{N} \left((-1)^{j+1} \sum_{I \subseteq \{1, \ldots, N\}} P\left(\bigcap_{i \in I} A_{i,t}\right) \right)}.$$

If member $i$ dies, his or her balance account $B_{i}$ is distributed to the survivors. To be fair, this reallocation takes place according to the specific characteristics of the surviving members: Older members and those with a larger stake in the tontine have to receive more. If member $k \neq i$ dies, member $i$ receives a fraction of $k$’s balance $a_{i,k,t} B_{k}$, where

$$0 \leq a_{i,k,t} \leq 1 \text{ for } i, k = 1, \ldots, N \text{ and } i \neq k.$$

\(^{12}\)See for example Graham et al. (1997).
$k$’s balance is forfeited entirely, so

$$a_{k,k,t} = -1 \text{ for } k = 1, \ldots, N. \quad (8)$$

Equation (8) states that the dying members’ stake in the tontine is distributed among the surviving members. In sum, the amount lost by $k$ equals the sum of the distributed benefits to the surviving members, so

$$\sum_{i=1}^{N} a_{i,k,t} = 0 \text{ for } k = 1, \ldots, N. \quad (9)$$

The unconditional expected benefit received by member $i$ in $t$ is the return in case no one dies and the return if at least someone dies, weighted with their corresponding probabilities, thus

$$E[r_{i,t}] = E[r_{i,t} | A_{0,t}] P(A_{0,t}) + E[r_{i,t} | A_{c0,t}] P(A_{c0,t}) \quad (10)$$

Since return is generated solely by mortality, if no one dies, there cannot be any return, thus $E[r_{i,t} | A_{0,t}] = 0$ and the expected return reduces to the second term of the right-hand side of Equation (10). The expected return conditioned that at least someone dies is the sum of the conditional death probability weighted fractions of the balance accounts over all $k$ members in $t$, thus

$$E[r_{i,t} | A_{c0,t}] = \sum_{k=1}^{N} \rho_{k,t} a_{i,k,t} B_k. \quad (11)$$

To achieve a fair tontine, each member’s expected benefit is zero in each year. This is because the expected loss of the own balance account in the case of the own death has to be offset by the expected gains one receives from other members’ deaths, so $E[r_{i,t} | A_{c0,t}] = 0$. The older $i$ is, the higher the death probability $q_{i,t}$, causing that $\rho_{i,t}$ increases as well (assuming that the composition of the tontine does not change) and leads to a higher expected loss in case of the $i$-th death. This has to be compensated by an increase in the fractions $a_{i,k,t}$ one receives in the case of other members’ death to counterbalance the aforementioned effect and to create a fair bet.

To satisfy the conditions of a fair bet for every tontine member $i = 1, \ldots, N$, one has to search for a set of $a_{i,k,t}$ that yield an expected benefit of zero for every tontine member and which fulfills conditions (6), (7), (8) and (9), yielding $E[r_{i,t}] = E[r_{i,t} | A_{c0,t}] = 0$.

As Sabin (2010) shows, such a set of $a_{i,k,t}$ exists only if no member is exposed to more than half of the total risk of the tontine. This can be achieved by introducing a ceiling of the amounts to invest $B_i$. Choosing $N$ large enough additionally reduces the threat of a single individual
holding too large a fraction of risky exposure of the tontine. Here, we implement an algorithm for the determination of the set of $a_{i,k,t}$ that is proposed by Sabin (2010), which approximately assigns constant $a_{i,k,t}$, irrespective of $k$ for $k \neq i$ and which provides best results for large $N$. In the following, we assume that the resulting $a_{i,k,t}$ satisfy conditions (6) - (9) and that no member holds more than half of risky exposure on the tontine, formally meaning that

$$\rho_{i,t}B_i \leq \frac{1}{2} \sum_{k=1}^{N} \rho_{k,t}B_k \text{ for } i = 1 \ldots N. \quad (12)$$

The expected return, conditioned that $i$ survives in $t$, is

$$E\left[r_{i,t}|A_{i,t}^c\right] = \sum_{\substack{k=1 \atop k \neq i}}^{N} \rho_{k,t}a_{i,k,t}B_k. \quad (13)$$

Because Equation (11) is solved to be zero, to yield a fair bet, $\sum_{\substack{k=1 \atop k \neq i}}^{N} \rho_{k,t}a_{i,k,t}B_k = -\rho_{i,t}a_{i,i}B_i$ and equation (13) is

$$E\left[r_{i,t}|A_{i,t}^c\right] = \mu_{i,t} = q_{i,t}B_i. \quad (14)$$

This is an interesting property since the individual expected return in case of the own survival is solely driven by the own mortality $q_{i,t}$ and the own investment in the tontine $B_i$, and does not depend on the tontine composition.

The unconditional realized benefit for $i$ in $t$ is

$$r_{i,t} = \sum_{k=1}^{N} a_{i,k,t}B_k \mathbf{1}\{A_{k,t} \cap A_{i,t}^c\} - B_i\mathbf{1}\{A_{i,t}\}$$

where the indicator function $\mathbf{1}\{...\}$ takes on the value of 1 if the respective event occurs and 0 otherwise and therefore $\mathbf{1}\{...\} \sim Ber_{p(...)}$. The realized return conditioned that $i$ survives is

$$r_{i,t}|A_{i,t}^c = \sum_{\substack{k=1 \atop k \neq i}}^{N} a_{i,k,t}B_k \mathbf{1}\{A_{k,t}\}. \quad (15)$$

Since the $a_{i,k,t}$ are approximately constant for $i$ for large $N$, $\overline{a}_{i,k,t} \approx \frac{q_{i,t}B_i}{\sum_{\substack{k=1 \atop k \neq i}}^{N} q_{k,t}B_k}$, and equation

---

13 For further algorithms to construct a tontine, see Sabin (2011).
(15) is

\[ r_{i,t} A_{i,t}^e \approx q_{i,t} B_i \frac{\sum_{k=1}^{N} B_k 1_{\{A_{k,t}\}}}{\sum_{k=1}^{N} B_k q_{k,t}}. \]

Although the volatility of the tontine converges toward zero for \( N \to \infty \), for a finite and realistic tontine size, the payouts are volatile\(^{14}\). As Appendix A.1 shows, the tontine payouts are approximately normally distributed, with

\[ \mu_{i,t} = q_{i,t} B_i \]

and

\[ \sigma_{i,t} = \frac{q_{i,t} B_i}{\sqrt{M - 1}} \left( \sum_{m=1}^{M} \left( \frac{\sum_{k=1}^{N} B_k 1_{\{A_{k,t}\}}}{\sum_{k=1}^{N} B_k q_{k,t}} - 1 \right)^2 \right)^{1/2}. \]

with \( M \) Monte Carlo simulation paths.

### 4.4 Annuity Model

The considered individual has pension wealth \( W_i \), no other wealth, no other source of income and the share of wealth which is not tontinized gets annuitized. The individual can choose to hold any positive proportion of his wealth in an annuity or in a tontine. The individual has no heirs and no desire to leave a bequest. We refer to an annuity applied by Milevsky (2006) which requires a lump sum investment of \( W_i - B_i \). The annuity then pays a stable income stream on a yearly basis until the participants’ death, starting as an immediate annuity.

The conditional survival probability of a person aged \( x \) in \( t \) of surviving \( \tau \) more years is defined as

\[ \tau p_x^t = \prod_{j=0}^{\tau-1} \left( 1 - q_{x+j}^t \right). \]

The annuity provider is risk-neutral. Therefore, because we refrain from interest rates in this model, the lump sum price \( \bar{a}_x^t \) in \( t \) of an immediate annuity which provides an income of 1 EUR per year until death is

\[ \bar{a}_x^t = \sum_{\tau=1}^{\Omega} \tau p_x^{t+\tau}. \]

\(^{14}\)For example, the tontines offered by Le Conservateur have to comprise at least 200 members to be launched. See [http://www.conservateur.fr/](http://www.conservateur.fr/).
To simplify, we denote the price of the immediate annuity as $\bar{a}_t^i$, where the individual characteristics can be identified via $i$. A lump sum investment of $W_i - B_i$ in the annuity provides a stable, lifelong and yearly income stream of

$$R_{i,t} = \frac{W_i - B_i}{\bar{a}_t^i}. \quad (18)$$

The individual now can decide on the allocation of tontinization ($B_i$) and annuitization ($W_i - B_i$) of the wealth $W_i$.

### 4.5 Old Age Liquidity Need Function

According to the Worldbank (2015), worldwide life expectancy at birth has increased between 1960 and 2015 from 52.5 to above 71.7 years. The increasing lifetime will cause the number of people over 80 years old to almost double to 9 million in Germany by the year 2060 according to forecasts by the Statistisches Bundesamt (2015). In the future, it is therefore very probable that very high ages of 100 years and even more will be achieved by a large number of people. According to the medicalisation thesis motivated by Gruenberg (1977), the additional years that people live due to demographic change are increasingly spent in bad health condition and disability. Jagger et al. (2016) find a significant increase in life expectancy between 1991 and 2011 in England. However, cognitive impairment, poor health and disability increase with age and elderly disability increased in that period. In those additional years of life, the demand for care products and medical service increases over-proportionately. Coming from 2.6 million nursing cases in Germany in 2013, Kochskämper (2015) estimates between 1.5 and 1.9 million additional nursing cases in Germany in the year 2060 due to demographic change. By the year 2030, the demand for stationary permanent care will increase by 220,000 places in Germany. While previous research finds a systematic decrease in the consumption level at retirement\(^{15}\), incorporating the nursing care costs and medical expenses into consumption yields a so called retirement smile\(^{16}\). When people retire they are mostly still healthy and have therefore time to spend on lifestyle. As they age, first physical constraints appear, they become more and more home-bound, and thus consumption declines while supplementary and medical costs are still at a low level. As people become very old they rely more on assisted living requirements, and costly long-term nursing care is needed. Therefore the typical monetary demand for a retiree

\(^{15}\)See for example Hamermesh (1984), Mariger (1987), Banks et al. (1998), Bernheim et al. (2001) or Haider and Stephens Jr (2007).

\(^{16}\)In an empirical work, Blanchett (2013) derives the trend of retirement consumption. He observes a shift in the expenditures towards increasing health care, entertainment and food.
is U-shaped. Based on own empirical research\footnote{The data used in this publication was made available to us by the German Socio-Economic Panel Study (SOEP) at the German Institute for Economic Research (DIW), Berlin.} we model an old-age liquidity need function, which accounts for demand for nursing care and medical service. The determination of the liquidity need function is based on data available on consumer spending in the SOEP from 1984 to 2013. We determine the age-specific expenditure pattern for the spending categories food, living, health, care, leisure, refurbishment and miscellaneous and finally aggregate them. The considered age ranges from 60 to 95. From the age of 95 on, we extrapolate until the age of 105 due to the limited data basis for those ages. The modeling of the nursing care costs is based on the costs of an inpatient, permanent care, which occurs in nursing homes, even though a large share of care is performed by family members and nursing care services. The inpatient care costs reflect the actual potential resources needed in a more appropriate manner because in home care, the time spent by family members is not taken into account. Moreover, many refuse inpatient care only because of lacking resources. Figure 1(a) shows our estimates of the average nursing care costs from the age of 60 to 90 per year. In addition, the average liquidity need without nursing care costs is shown. If we aggregate both components, we obtain the characterized retirement smile, which is presented in Figure 1(b).

In our model, we map the old-age liquidity need via a polynomial liquidity need function $D_t$ of order 2 which is calibrated based on our results, and assume an extrapolation up to the age of $x = 105$. The desired consumption level is driven by age $x$ in $t$ so

$$D_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 \epsilon_t$$

where the parameters $\beta_0$, $\beta_1$ and $\beta_2$ are fitted using our empirical data, and $\epsilon_t$ is the error term.
So far, we have on the one hand an income stream which is composed of both certain annuity and volatile tontine payments, and on the other hand an age-increasing liquidity need which the income stream should cover. We aim to design the payout pattern of tontine and annuity such that the liquidity need can be served optimally. Therefore, we evaluate the income stream relative to the liquidity need as reference point. An income stream larger than the liquidity need is considered to be as a gain and utility-generating, while an income stream lower than the liquidity need generates a loss, which provides disutility. In this sense, we look at the utility of the income relative to the liquidity need, rather than the absolute level of income. Since the liquidity need increases with age, a payout which is able to meet the demand in early years might not be sufficient in the later years of retirement. Therefore, to evaluate an income stream relative to a reference point, Cumulative Prospect Theory (CPT), originated by Kahneman and Tversky (1979) and enhanced by Tversky and Kahneman (1992) is highly suitable for our purpose\textsuperscript{18}. Although CPT is a descriptive rather than a normative theory, it allows us to capture the aforementioned properties and to determine an optimal, CPT-utility maximizing fraction to be invested in the tontine. To capture the life-cycle dynamics of the repeating payments until death, we use the Multi Cumulative Prospect Theory (MCPT), as applied by Ruß and Schelling (2017), where the CPT-utility is determined in every period \( t \) by considering a changing reference point which is represented by the respective liquidity need \( D_t \), and is finally aggregated with respect to survival prospects. The total utility of person \( i \) over his or her stochastic remaining lifespan is the sum of the CPT-utilities of the gains and losses \( Z_{i,t} \) of the payouts generated by the portfolio of tontine and annuity in relation to the liquidity need in each point in time \( t \), deflated by a subjective discount factor \( \delta \leq 1 \). The conditional survival probability \( \tau p_x \) of an \( x \) year old of surviving \( \tau \) more years is incorporated in the CPT-utility. It is combined with the density of the respective retirement payouts and the resulting joint probability is valued according to the CPT, thus

\[
MCPT (i) = \sum_{\tau} \delta^\tau CPT (Z_{i,t+\tau}) .
\]  

\textsuperscript{18} In this sense, Schmidt (2016) uses CPT to determine insurance demand.
The CPT-utility in each period is

\[ CPT(Z_{i,t+\tau}) = \int_{-\infty}^{0} v(z) d\left(w^- (F_{i,t+\tau} (z))\right) + \int_{0}^{\infty} v(z) d\left(-w^+ (1 - F_{i,t+\tau} (z))\right) \] (21)

in a continuous context\(^{19}\). The probability weighting function \( w^+ (F) \) for gains and \( w^- (F) \) for losses is

\[ w^+ (F) = \frac{F^\gamma}{(F^\gamma + (1 - F)^\gamma)^{1/\gamma}}, \quad w^- (F) = \frac{F^\kappa}{(F^\kappa + (1 - F)^\kappa)^{1/\kappa}} \] (22)

where \( \gamma \) and \( \kappa \) are the probability weighting parameters. The value function \( v(z) \) is given by

\[ v(z) = \begin{cases} z^a & z \geq 0 \\ -\lambda |z|^b & z < 0 \end{cases} \] (23)

where \( a, b \in (0,1) \) and \( \lambda > 1 \). The mixture cumulative distribution function, to account for the joint probability of conditional survival and the payout size, is

\[ F_{i,t+\tau} (z) = (1 - \tau p_x) 1_{[0,\infty)} + \tau p_x \int_{-\infty}^{z} dF_{N_{i,t+\tau}} (u) \] (24)

with the first moment

\[ \mu_{i,t+\tau} = q_{i,t+\tau} B_i + \frac{W_i - B_i}{a_{i,0}} - D_{i,t+\tau} \] (25)

and the standard deviation \( \sigma_{i,t+\tau} \) resulting from equation (17) for the normally distributed gains and losses. Figure 2 shows the illustration of equation (24): dying leads to \( z = 0 \) to which the death probability \( 1 - \tau p_x \) is assigned. This explains the jump in the CDF.

\(^{19}\)See for example Hens and Rieger (2010), Ågren (2006) or Ruß and Schelling (2017) who use the CPT in a continuous context.
By incorporating the mixture CDF in the analysis, equation (21) becomes

\[
CPT(Z_{i,t+\tau}) = p_x \left[ \int_{0}^{\infty} v(z) w^{-\tau}(F_{i,t+\tau}(z)) f_{\Phi(i,t)}(z) dz + \int_{0}^{\infty} v(z) w^{+\tau}(1 - F_{i,t+\tau}(z)) f_{\Phi(i,t)}(z) dz \right] + (1 - p_x) \left[ v(0^-) w^{-\tau}(F_{i,t+\tau}(0^-)) + v(0^+) w^{+\tau}(1 - F_{i,t+\tau}(0^+)) \right]
\]

with

\[
w'(F) = \frac{F^\theta - 1}{F + (1 - F)^\theta} \frac{(-\theta + 1) (F - 1) (\theta - 1) F^\theta + (F (\theta - 1) - \theta) (1 - \theta)^\theta}{\theta - 1}
\]

where \( \theta \in (\gamma, \kappa) \). The first line of equation (26) is the utility in case of survival, whereas the second line is the utility in case of death. Since the utility in case of death is zero because of \( v(0) = 0 \), equation (26) reduces to just the first line.

Finally, we numerically maximize equation (20) subject to the optimal level of tontine investment \( B_i \).

\[
\max_{B_i} MCPT(i) \quad \text{s. t. } B_i \leq W_i, B_i \geq 0
\]

### 4.7 Variation: Stochastic Liquidity Need in MCPT Valuation

From Equations (25) and (17) follows that the combined payout of tontine and annuity is normally distributed with

\[
Z_{i,t} \sim N(\mu_{i,t}, \sigma^2_{Z_{i,t}}).
\]

To cover effects stemming from uncertainty about the future liquidity need, we assume that the liquidity need itself is normally distributed with mean \( E(D_{i,t}) \) and standard deviation \( \sigma_{D_{i,t}} \), therefore

\[
\mu_{i,t} \sim N(\mu_{i,t0}, \sigma^2_{\mu_{i,t}})
\]

where \( \mu_{i,t0} = q_{i,t} B_i + \frac{\tilde{W}_i - B_i}{a_{i,0}} - E(D_{i,t}) \) and \( \sigma^2_{\mu_{i,t}} = \sigma^2_{D_{i,t}} \) is calibrated based on own empirical research. Tontine payments and liquidity need are assumed to be independent. If we write

\[
Z'_{i,t} = (Z_{i,t} - \mu_{i,t}) + \mu_{i,t}
\]
then
\[(Z'_{i,t} - \mu_{i,t}) \sim \mathcal{N}(0, \sigma_{i,t}^2)\]

and
\[Z'_{i,t} \sim \mathcal{N}(0, \sigma_{i,t}^2) + \mathcal{N}(\mu_{i,t}, \sigma_{\mu,i,t}^2) = \mathcal{N}(\mu_{i,t}, \sigma_{i,t}^2 + \sigma_{\mu,i,t}^2)\] (29)
because the sum of two normally distributed random variables is also normally distributed. While the mean expected payout remains the same, the volatility increases to the sum of the variance of the tontine payment and the variance of the liquidity need.

4.8 Variation: Subjective Mortality

To account for subjective beliefs about the own mortality risk, we adjust the objective forecasted mortality. It is important to understand the difference compared to the probability adjustment which CPT undertakes: while CPT accounts for a deviating perception of objective probabilities, the subjective mortality adjustment modifies the average probabilities subject to own perceptions about the individual health status. People who believe to live longer than the aggregate average because they feel very healthy or have an active lifestyle perceive to have lower death probabilities and thus believe to have a longer expected remaining lifetime. Therefore, the optimistic subjective death probability \(q'_{x,t}\) is lower than the average, objective death probability \(q_{x,t}\). Likewise, the pessimistic subjective death probability \(q'_{x,t}\) for persons who believe to live shorter than the overall average (because of severe illness or the awareness of a poor lifestyle) is higher than the actual death probability \(q_{x,t}\). Bissonnette et al. (2014) show that, within different groups (e.g. gender, ethnic background or education), people with similar characteristics are only slightly optimistic regarding their survival prospects compared to the average mortality within the subgroup, whereas the actual subgroup mortality itself differs tremendously from the overall population mortality. The authors conclude that the individual perceptions are very precise. Therefore, it is important to incorporate subjective survival probabilities in our analysis, because people who believe to live longer tend to live longer, and thus different retirement planning solutions are needed for different individuals. To account for subjective mortality in our model, we adjust the actual mortality rates \(q_{x,t}\) by an individual mortality multiplier \(d\), therefore the subjective mortality rate \(q'_{x,t}\) is
\[q'_{x,t} = \begin{cases} d \cdot q_{x,t} & \text{if } d \cdot q_{x,t} \leq 1 \\ 1 & \text{otherwise} \end{cases}\] (30)
where \( d \) is the realization of a random variable \( D \) and determines the subjective survival probability. For \( 0 < d < 1 \) the individual expects to live longer than the average, if \( d = 1 \) the individual self-assesses his or her lifetime of being average and if \( d > 1 \), the individual expects to live shorter than the actual mortality table predicts. Furthermore, \( q_\Omega = 1 \) which means that there is a limiting age \( \Omega \) when the individual dies with certainty. A simple modeling approach for \( d \sim D \) is shown in Appendix A.2, where \( D \) is modeled using a Gamma Distribution. Since the insurance company offering tontines and annuities uses average objective mortality rates, pricing is undertaken on the basis of average mortalities. The subjective beliefs only influence the subjective determination of individual utility.

## 5 Calibration and Results

We calibrate the Lee-Carter model based on data from the *Human Mortality Database*\(^{20}\), and forecast mortality rates for \( T = 100 \) years, beginning from 2011, which is denoted by \( t = 1 \) in the analysis. The fractions of female and male newborns to update population pyramids are calculated based on German birth statistics\(^{21}\). The age-specific birth rates are based on German birth statistics\(^{22}\) and describe the number of newborns per year of a woman in each cohort. The maximum attainable age is set to be \( \Omega = 105 \) which means that at the age of \( x = 105 \) one dies with certainty. We consider initial wealth \( W_i \) as an independent variable and measure its influence on the optimal investment behavior under various scenarios. The parameters of the polynomial liquidity need function \( D_t \) are \( \beta_0 = 163,984.686, \beta_1 = -4,000.634 \) and \( \beta_2 = 28.589 \) and fit our empirically estimated old-age liquidity need function based on SOEP data.

### 5.1 Base Case

In Table 1, we report the parameters used in the MCPT\(^{23}\) analysis, which constitute the base case. We consider \( i \) as a male individual aged 62 in the year the tontine is set up \((t = 1)\), and vary his initial endowment \( W_i \). Furthermore, we assume that the remaining tontine members \( k = 1 \ldots N, k \neq i \) behave optimally, i.e. the individual amounts \( B_k \) invested in the tontine are assumed to be the \( MCPT_k \)-utility maximizing amounts and are assumed to be uniformly


\(^{21}\) Data from 2000 - 2010, see Statistisches Bundesamt (2012).


\(^{23}\) We use the notation MCPT according to equation (20) for the sum of the periodic CPT-utilities and CPT according to equation (26) for the periodic utilities.
Based on $M = 10,000$ simulations, we calculate the realized tontine returns for individual $i$ in every period in which he is alive and thereby determine the moments of the normal approximation of tontine returns for member $i$. We set the subjective discount factor $\delta = 1$, because we assume that the future states are as important as present states for an individual who aims to secure the future standard of living\textsuperscript{25}. We calibrate the CPT-value function parameters $a, b$ and $\lambda$ according to the values proposed by Tversky and Kahneman (1992), but use the actual (i.e. not weighted) probabilities, i.e. $\gamma = \kappa = 1$\textsuperscript{26}

\begin{table}[h]
\centering
\begin{tabular}{l|l|l}
\hline
Parameter & Notation & Value \\
\hline
Forecast horizon (in years) & $T$ & 100 \\
Maximum attainable age & $\Omega$ & 105 \\
Fraction of female newborns & $\omega^f$ & 48.68 % \\
Fraction of male newborns & $\omega^m$ & 51.32 % \\
Lower boundary age at tontine entrance & $\underline{x}$ & 62 \\
Upper boundary age at tontine entrance & $\overline{x}$ & 100 \\
Size of the tontine & $N$ & 10,000 \\
Monte Carlo Paths & $M$ & 10,000 \\
Subjective discount factor & $\delta$ & 1 \\
CPT value function parameters & $a, b$ & 0.88 \\
CPT loss sensitivity factor & $\lambda$ & 2.25 \\
CPT $w^+$ parameter & $\gamma$ & 1 \\
CPT $w^-$ parameter & $\kappa$ & 1 \\
\hline
\end{tabular}
\caption{MCPT parameters Base Case}
\end{table}

Since the expected tontine return as well as the tontine volatility are driven by the individual survival probability, both increase as $i$ becomes older. Aged 62 in $t = 1$, a person investing $B_i = 30,000$ EUR in the tontine can expect to receive a first-year tontine return of 345.15 EUR which amounts to 1.15% of the initial tontine investment. In $t = 10$, his expected return is roughly 1.6 times higher than in the first year but still amounts to only 1.86% of the initial investment. The payments thus increase slowly in the early retirement years because of the slow increase of death probabilities in the early years. After 20 years, at the age of 81, the expected return is already 4.8 times as large as in the first year and the single payment in this year amounts to 5.51% of the initial investment. For very high ages, the payments increase

\textsuperscript{24}For every amount invested in the tontine $B_k$, there exists a corresponding amount of initial wealth $W_k$ for which $B_k$ is optimal. Since we chose the optimal amounts of $B_k$ randomly, we implicitly specify their initial wealth levels $W_k$. The chosen values for $B_k$ correspond to the range of optimal amounts invested in the tontine for average individuals for the considered range of initial wealth $W_i$. Thereby we provide the optimal investment decision for hypothetical levels of $W_k$, which are roughly in the same range as the bandwidth of $W_i$.

\textsuperscript{25}In this sense, Parsonage and Neuburger (1992) and Van der Pol and Cairns (2000) provide empirical evidence that it is feasible to assume a subjective discount rate of zero for the discounting of future health benefits.

\textsuperscript{26}This is in line with the asset pricing literature. See for example Barberis et al. (2001) and Levy and Levy (2004), who also use the actual probabilities. In section 5.7 we calibrate $\gamma$ and $\kappa$ according to Tversky and Kahneman (1992).
tremendously: at the age of 91, in $t = 30$, the expected tontine return is almost 16.3 times as large as in the first year and yields a rate of return of 18.73%. Every year of further survival then yields even steeper increasing returns, being 47.69% of the initial investment at the age of 101 in $t = 40$, and finally 100% at the maximum attainable age of 105 in $t = 44$. Neglecting interest rate effects, one can expect to recoup the initial investment in year 26 at the age of 87. Since the standard deviation of the tontine returns depends on the individual mortality, volatility increases similarly with age. In comparison, an immediate fairly priced annuity with a lump sum investment of 30,000 EUR would yield an income of yearly 1458.96 EUR. The investor could recoup the initial investment already after 21 years at the age of 82. This is because the annuity provides stable payments whereas tontine payments increase with age. Figure 3 shows the yearly expected payout patterns of a tontine and an annuity with investment volume normalized to unity.

Based on these considerations, we determine the CPT-utility $CPT(Z_{i,t})$ in each period for different levels of $W_i$ for member $i$. Figure 4(a) shows the CPT-utility of member $i$ for an initial wealth endowment of $W_i = 572,000$ EUR for different portfolio compositions at each point in time $t$. This represents the expected contribution of the CPT-utility on the aggregated MCPT-utility. Since survival probabilities decline with age, the impact of each CPT-utility declines with age and finally converges toward zero. We first consider the case in which person $i$ completely annuitizes his initial wealth (solid line). As annuity payments are constant, an increasing liquidity need $D_{i,t}$ causes declining CPT-utilities in time. In early years, the liquidity need can be met. As the liquidity need rises and exceeds the available funds, CPT-utility decreases and becomes negative. As age increases, the declining survival probability causes a lower CPT utility which reduces the impact of late periods on MCPT-utility. For the very late years, the low survival probabilities outweigh the negative CPT-utility, yielding that, finally, the impact of very
late years on MCPT-utility approaches zero. Second, we consider the complete tontinization of initial wealth (dotted line). Because tontine payments are driven by mortality, payments are very low in the early years and increase in age, thus the liquidity need cannot be met for early ages and can easily satisfy $D_{i,t}$ in later years. Again, very low survival probabilities in later years reduce the impact on MCPT-utility. Furthermore, since tontine and annuity payments proceed adversely, a portfolio of both can help to generate payout patterns which enable to finance the increasing liquidity need appropriately. The dashed line shows the CPT-utilities of a payout pattern of a portfolio consisting of 10% tontine and 90% annuity. While still being able to satisfy the liquidity need in the early years, it is also able to provide almost sufficient funds in the later years. The sum of the CPT-utilities yields the MCPT-utility.

Figure 4(b) exemplarily shows the MCPT-utility for different fractions of $W_i = 572,000$ EUR being invested in the tontine ($B_i$). The remaining fraction $W_i - B_i$ is annuitized. Starting from
a situation of complete annuitization, \( i \) can increase his MCPT-utility by investing a positive fraction in the tontine, and finally maximizes his MCPT-utility if he invests 10.84\% of \( W_i \) in the tontine. An optimal fraction exists because of two counteracting effects: up to an optimal point, a higher investment in the tontine increases the later years’ CPT-utilities more than it decreases the early years’ CPT-utilities, yielding an increasing MCPT-utility. Beyond this optimal point, the decrease in CPT-utility in early years outweighs the increase in CPT-utility in the late years, yielding a declining MCPT-utility. These effects are resulting from the fact that up to the age of 80, the annuity provides a higher return than the tontine, while beyond the age of 80, the tontine outperforms the annuity. Therefore, one unit of additional investment in the tontine decreases CPT-utilities until the age of 80 and increases CPT-utilities beyond the age of 80, finally yielding an optimal MCPT-utility maximizing tontine investment level.

Figure 4(c) shows the optimal, MCPT-utility maximizing fractions to be invested in the tontine for different levels of initial wealth \( W_i \). If \( W_i < 537,000 \) EUR, it is optimal not to invest in the tontine. This is because even for complete annuitization, annuity payments are so low that the CPT-utility losses in early years, caused by investing in the tontine, are large and cannot be offset by the CPT-utility gains in later years, caused by increasing tontine payouts. As \( W_i \) increases, the optimal fraction to invest in the tontine increases very sharply up to \( W_i = 572,000 \) EUR and decreases thereafter. In the wealth region \( 537,000 \leq W_i \leq 753,000 \) the reduction in early years CPT-utilities due to shifting from annuity to tontine investment is overcompensated by the increase in late years’ CPT-utilities. This is because the marginal CPT-utility in early years is lower for higher \( W_i \), and therefore more wealth can be shifted from the annuity to the tontine investment. The optimal tontine fraction decreases beyond the peak at \( W_i = 572,000 \) EUR because for higher \( W_i \) marginal CPT-utility decreases for late years’ consumption and less wealth in relative terms is needed to increase late years’ CPT-utilities. In other words, the CPT-utilities in early years’ do not decline much, while late years’ consumption can be financed with the additional tontine payments. For \( W_i > 753,000 \) EUR it is again optimal not to invest at all in the tontine. At this wealth level, the annuity payments are sufficient to satisfy the liquidity need in early as well as in later years. An investment in the tontine thus would reduce early consumption possibilities and therefore reduce early years’ CPT-utility, while the gain from later consumption would be very small because later years’ liquidity need can already be met by the annuity payments. Therefore, the tontine would take away funds in early years in which survival prospects are high and therefore negatively impact utility. In turn, the tontine would provide funds in states when the additional tontine payments are not needed because funds
from the annuity payments are already sufficient. In addition, these funds hardly contribute to MCPT-utility because of low survival prospects at high ages.

Figure 4(d) shows the optimal MCPT-utility for different levels of $W_i$ compared to the MCPT-utility under complete annuitization. For $W_i < 537,000$ EUR and $W_i > 753,000$ EUR, complete annuitization provides the highest MCPT-utility. As seen before, in these domains tontine investment reduces the MCPT-utility. For $537,000 \leq W_i \leq 573,000$ EUR, the highest MCPT-utility can be achieved by investing a positive fraction of wealth in the tontine. The highest MCPT-utility increase can be generated at $W_i = 584,000$ EUR. This can be seen in the gray shaded area, which corresponds to the scale on the right hand side of the figure.

5.2 Variation: Equal Treatment of Gains and Losses

If we set the parameters of the value function of the CPT to $a = b = 0.5$, we receive a square root utility, by which gains and losses are treated equally. As Figure 5 shows, the resulting fractions of tontine investment are generally similar compared to the base case setting. It is striking that at $W_i = 590,995.07$ EUR the optimal fraction to invest in the tontine immediately jumps from 0 to 10.80% and decreases thereafter, until it finally reaches 0 again at $W_i = 950,158.80$ EUR. Compared to the base case, investment in the tontine is optimal for higher $W_i$. Figure 6 provides an explanation for these results.

Figure 6(a) shows the MCPT-utility values for $W_i = 570,000$ EUR for different levels of tontine investment. The highest MCPT-utility can be achieved if no investment in the tontine takes place. If the tontine investment increases, the MCPT-utility first decreases, and at roughly 12% there is a little peak with a local maximum where MCPT-utility slightly increases, but decreases

![Figure 5: Equal treatment of gains and losses-calibration (ETGL) maximizing fractions to invest in the tontine for different levels of $W_i$](image-url)
(a) generally decreasing utility, little hump

(b) hump grows and moves to the left

(c) indifference, hump grows even more and moves further to the left

(d) hump surmounts no-tontine utility, moves further to the left

(e) no decreasing utility, optimal level

(f) decreasing utility, optimal level has reached 0

Figure 6: MCPT-utility for different levels of $W_i$ and increasing fractions of tontine investment relative to $W_i$ for square-root utility calibration of the base case
thereafter again\textsuperscript{27}. As shown in Figure 6(b), for a 10,000 EUR higher initial wealth \( W_i = 580,000 \) EUR, two changes in the shape of the MCPT-utility curve can be observed: first the hump increases, meaning that the positive influence on MCPT-utility of tontine investments increases, and second, the hump moves a bit to the left compared to the previous wealth level, meaning that the local maximum results for lower fractions of tontine investment, compared to the situation presented in figure 6(a). As initial wealth reaches the threshold value \( W_i = 590,995.065 \) EUR (Figure 6(c)), the hump is as large as that the MCPT-utility with 10.82\% tontine investment equals the MCPT-utility without tontine investment. Therefore, the individual is indifferent between no tontine investment and 10.82\% tontine investment. For a tontine investment between 0 and 10.82\%, the MCPT-utility is lower compared to the maximum MCPT-utility. For tontine investments larger than 10.82\%, the MCPT-utility decreases as well. As \( W_i \) further increases, the peak further moves to the left and surmounts the MCPT-utility without tontine investment (Figure 6(d)). Gradually, the local minimum between no tontine investment and optimal tontine investment disappears (Figure 6(e)). Finally, as \( W_i \) is very large, the slope around the local maximum is very flat and finally disappears when the MCPT-utility maximizing fraction to invest in the tontine hits 0 again (Figure 6(f)).

5.3 Variation: Tontine Size

For an increased tontine size of \( N = 100,000 \) (compared to \( N = 10,000 \) in the base case), the volatility of the tontine payments declines (Table 2). As presented in Figure 7, less volatile tontine payments make it optimal to invest in the tontine for lower \( W_i \) than in the base case scenario. Similarly, for higher \( W_i \), investing in the tontine remains beneficial with an increased pool size. This is because less volatile payments generally enhance CPT-utilities. Therefore, it is optimal for both a lower and a higher \( W_i \) to invest more in the tontine compared to \( N = 10,000 \).

<table>
<thead>
<tr>
<th>( t )</th>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>44</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_{i,t}^{N=10,000} )</td>
<td>345.15</td>
<td>559.72</td>
<td>1,652.04</td>
<td>5,619.14</td>
<td>14,307.30</td>
<td>30,000</td>
</tr>
<tr>
<td>( \sigma_{i,t}^{N=10,000} )</td>
<td>19.05</td>
<td>25.30</td>
<td>63.63</td>
<td>227.70</td>
<td>572.00</td>
<td>1169.77</td>
</tr>
<tr>
<td>( \sigma_{i,t}^{N=100,000} )</td>
<td>5.93</td>
<td>8.74</td>
<td>23.00</td>
<td>78.43</td>
<td>187.74</td>
<td>370.75</td>
</tr>
</tbody>
</table>

Table 2: Properties of normally distributed tontine returns for \( B_i = 30,000 \) in \( t \) in EUR for tontine size \( N = 100,000 \) vs. \( N = 10,000 \)

\textsuperscript{27} The reason for this peak lies in the tradeoff between early years and later years consumption possibilities as explained in the previous base case.
Variation: Stochastic Liquidity Need

If we assume a stochastic liquidity need, the fact whether cash flows lead to gains or losses with respect to the liquidity need is affected by the volatility of the tontine payments as well as by the volatility of the liquidity need. A stochastic liquidity need increases the overall volatility and therefore CPT-utilities decline. First, we set the variance of the liquidity need at $\sigma^2_{\mu_i,t} = \sigma^2_{D_i,t} = 2,500^2$ which could reflect both, uncertainty about health care costs and uncertainty of becoming frail and dependent. As a consequence, the wealth level at which it becomes optimal to invest in the tontine increases compared to the base case (see Figure 8). The reason for this lies in the fact that in early years the more volatile nature of gains and losses makes it desirable to hold more funds to cover the liquidity need. Every unit taken away from the annuity in the early years causes a huge decline in early years’ CPT-utilities. Therefore, it is optimal only for a higher $W_i$ to invest in the tontine. The opposite effect applies to high $W_i$. The volatile liquidity need brings about situations of high liquidity need in which payments from the tontine can support its coverage. Therefore, it is optimal for a higher $W_i$ to hold some fraction in the tontine. Furthermore, the optimal fraction to invest in the tontine is smaller compared to the base case, because the tontine investment itself adds another layer of volatility to the payments, which decreases utility. As we further increase the volatility of the liquidity need to $\sigma^2_{\mu_i,t} = \sigma^2_{D_i,t} = 4,000^2$, we can observe a boost of both effects. A higher $W_i$ is required to start investing in the tontine in order to lower the risk of experiencing a utility-harming drop far below the liquidity need. As the level of $W_i$ is relatively high, the tontine investment loses its efficiency compared to the resulting annuity payments, yielding a lower optimal fraction to be invested in the tontine.
5.5 Variation: Subjective Mortality

If we adjust the mortality according to equation (30) by \( d = 0.8 \), the individual expects to live longer than average. This means that future periods have a greater impact on MCPT-utility because survival probabilities decline less fast. Therefore, later years’ CPT-utilities are higher compared to the base case. This situation is presented in Figure 9(a), where the dotted lines represent the CPT-utility paths for the base case and the solid lines represent the CPT-utility paths for the subjective, improved mortality. As a result, positive and negative subjective CPT-utilities both have a higher impact on total MCPT-utility compared to the base case, indicating that it might be more favorable to invest a higher fraction in the tontine because it is more likely to experience the later years’ CPT-utilities. Figure 9(b) shows that it is optimal to invest in the tontine for lower \( W_i \) because, by investing in the tontine, later years’ CPT-utilities gain more relevance and are higher although early years’ CPT-utilities are reduced. By investing more intensely in the tontine, overall MCPT-utility can be increased. Furthermore, it is optimal to invest in the tontine up to a higher \( W_i \), compared to the base case. This is because marginal CPT-utility in later years increases as survival probabilities increase. Early years’ CPT-utility losses can be overcompensated by later years’ CPT utility gains. Furthermore, later years’ CPT-utility losses also have a higher impact on the MCPT and, therefore, a tontine investment in higher \( W_i \) regions can help to mitigate the otherwise resulting underfunding problem. In addition, it is optimal to invest a higher fraction in the tontine for all \( W_i \)’s for which it is optimal to invest in the base case. This is due to the increased probability of experiencing CPT-utilities in the late years.
5.6 Variation: Changing Liquidity Need

In this section we change the shape of the liquidity need. First, we parallel shift the standard liquidity need curve up by 10,000 EUR. Second, we assume an exponential growth of the standard liquidity need curve by $D_t^{exp} = 1.01^t D_t$.

Since the standard liquidity need curve represents the average liquidity need unconditioned on the health status, an exponential growth can be interpreted as the liquidity need conditional on bad health. Figures 10(a) and 10(b) show the resulting liquidity need curves and the optimal fractions to invest in the tontine for the different liquidity need curves. If we assume a parallel, upward shift of the liquidity need curve by 10,000 EUR, two characteristics of the optimal investment choice can be observed. First, the optimal investment pattern shifts to the right, which means that it is optimal to invest in the tontine only for higher initial wealth endowment $W_i$. Second, the optimal fractions to invest in the tontine are lower compared to the base case. The reason for these two properties lies in the increasing liquidity need in every period. For a relatively low $W_i$, it is not optimal to invest in the tontine because the loss in CPT-utilities in the early years due to a reduction of annuitized wealth exceeds the CPT-utility gains in later years. Only if there is sufficient initial wealth it is optimal to invest in the tontine. The maximal tontine investment in the parallel shift case is lower compared to the peak in the base case. This is a consequence of a substantially higher liquidity need in the early years which can not be covered by tontines. If we assume an exponentially increasing liquidity need, investment in the tontine starts for a higher $W_i$ compared to the base case and below the parallel shift case at approximately $W_i = 600,000$ EUR. Furthermore, the peak of the optimal amount to be invested in the tontine is almost twice as large compared to
the base case. Optimal positive fractions of tontine investment persist longer for high $W_i$. This is because in the early years the liquidity need in the exponential case is relatively close to the base case and disproportionately increases with age, compared to the base case. Therefore, the CPT-utility decrease in early years is relatively low when investing some fraction in the tontine, while the CPT-utility gains of the tontine investment in the late years are very high. Thus, with large amounts of money invested in the tontine, the early years’ CPT-utilities do not suffer much, while later years’ CPT-utilities benefit strongly. As a consequence, larger amounts to be invested in the tontine are optimal to satisfy the liquidity need best. To sum up, the tontine is most powerful if the liquidity need is low in the early years and high in the later years of retirement.

5.7 Variation: Cumulative Prospect Theory Probability Weights

If we calibrate the probability weighting function of the Cumulative Prospect Theory according to the originally proposed values by Tversky and Kahneman (1992), $\gamma = 0.61$ and $\kappa = 0.69$, we can observe in Figure 11 that the optimal fractions to invest in the tontine are on a similar level as using objective probabilities. However, using the distorted weights makes it optimal to begin to invest in the tontine for a slightly lower $W_i$, and that it is optimal to invest a larger fraction in the tontine for any wealth level, where it is optimal in the base case to invest at least some fraction. Furthermore, it is still optimal for a higher available wealth $W_i$ to invest...
in the tontine. The reason for the slightly different tontine investment pattern is that small probabilities are over- and large probabilities are under-weighted. Small gains and losses occur with relatively equal probabilities and thus are hardly distorted. However, large gains and losses are low probability events and are therefore perceived to occur less frequent. Since a large loss harms more than a large gain benefits, the lower perception of those extreme events yields that the optimal tontine investment curve with distorted probabilities lies slightly above the base case optimality curve.

6 Summary and Conclusion

The changing social, financial and regulatory framework, such as an increasingly aging society, the current low-interest environment, as well as the implementation of risk-based capital standards in the insurance industry, lead to the search for new product forms for private pension provision. These product forms should reduce or avoid investment guarantees and risks stemming from longevity, provide reliable insurance benefits and reflect in the payout pattern the increasing financial resources required for very high ages. We propose the traditional tontine to serve as such "product innovation", especially in combination with a traditional life annuity.

To assess the effects of tontine investments on policyholders’ welfare, we develop a model by which individual old-age liquidity needs and payouts stemming from annuity and tontine investment can be evaluated and result in an optimal retirement planning decision, based on individual preferences, characteristics and subjective mortality beliefs.

To show the effects of a tontine investment on retirement planning, we model the development of the changing population structure for the next decades in Germany. Based on the changing mortality dynamics, we describe a fair revolving tontine. To assess its advantages and disad-
vantages compared to a traditional life annuity, we derive a targeted consumption level from empirical data and combine the tontine payout structure with the payout structure of a traditional annuity to optimally cover the desired consumption. Our results reveal that a portfolio of annuity and tontine can provide the highest expected Multi Cumulative Prospect Theory utility. While the annuity pays a stable, constant pension, the tontine provides volatile, age-increasing payouts. The results of our analyses prove to be sensitive with respect to the initial wealth endowment, the shape of the liquidity need curve and the subjective expectation about the remaining lifetime of an individual. For very low and very high endowments, complete annuitization is optimal, whereas for medium endowments of initial wealth, it might be optimal to invest a small fraction in the tontine, depending on individual circumstances. Taking these circumstances into account, our results indicate that, from a policyholder perspective, a tontine can be a beneficial supplement to existing retirement planning solutions.

Future research could incorporate investment risk and analyze its effects on the optimal tontine investment decision. The level of investment risk might significantly change the optimal allocation of retirement wealth. In this context, the integration of reinvestment opportunities of tontine and annuity returns might yield an additional determining factor for the optimal retirement planning decision. Another interesting area for further research is the analysis from an insurer’s perspective. It will be interesting to analyze whether the supply of tontines can reduce the insurer’s capital requirements, or reduce safety loadings included in annuity prices, since the tontine involves no longevity risk for the provider, and might partially substitute annuities.

References


A Appendix

A.1 Tontine Volatility

The variance of the return conditional that $i$ survives in $t$ is

$$Var \left[ r_{i,t} | A_{i,t}^c \right] = E \left[ r_{i,t}^2 | A_{i,t}^c \right] - \left( E \left[ r_{i,t} | A_{i,t}^c \right] \right)^2$$

$$= E \left[ \sum_{k=1}^{N} a_{i,k,t} B_k 1_{ \{ A_k,t \} } \right]^2 - \left( q_{i,t} B_i \right)^2.$$

If we assume constant $a_{i,k,t}$'s, the variance can be expressed as

$$Var \left[ r_{i,t} | A_{i,t}^c \right] = q_{i,t}^2 B_i^2 \frac{\sum_{k=1}^{N} B_k^2 E \left[ 1_{ \{ A_k,t \} } \right]^2}{\sum_{k=1}^{N} B_k^2 q_{k,t}^2 + \sum_{k=1}^{N} \sum_{k \neq j} B_k B_j q_{k,t} q_{j,t}}$$

$$+ q_{i,t}^2 B_i^2 \frac{\sum_{k=1}^{N} B_k^2 q_{k,t}^2 + \sum_{k=1}^{N} \sum_{k \neq j} B_k B_j q_{k,t} q_{j,t}}{\sum_{k=1}^{N} B_k^2 q_{k,t}^2 + \sum_{k=1}^{N} \sum_{k \neq j} B_k B_j q_{k,t} q_{j,t}} - q_{i,t}^2 B_i^2.$$

For $N \to \infty$

$$Var \left[ r_{i,t+1} | A_{i,t}^c \right] = q_{i,t}^2 B_i^2 - q_{i,t}^2 B_i^2 = 0$$

because $E \left[ 1_{ \{ A_k,t \} } \right] = q_{k,t}$. Nevertheless, the tontine we employ is large, but still risky\(^{28}\). For a fixed tontine size $N$, the tontine payouts are still volatile and follow a Poisson binomial distribution in each $t$ which we approximate by a normal distribution\(^{29}\) $N(\mu_{i,t}, \sigma_{i,t})$ for large $N$ where $\mu_{i,t} = q_{i,t} B_i$ and

$$\sigma_{i,t} = \frac{q_{i,t} B_i}{\sqrt{M-1}} \left( \sum_{m=1}^{M} \left( \frac{\sum_{k=1}^{N} B_k 1_{ \{ A_k,t \} }^m}{\sum_{k=1}^{N} B_k q_{k,t}} - 1 \right)^2 \right)^{1/2}.$$

with $M$ Monte Carlo simulation paths.

\(^{28}\)At a tontine size of $N = 100,000,000$ members, the volatility would be negligible. Of course, a tontine of this size is not realistic. Therefore, in our numerical approach, we employ a tontine size of $N = 10,000$ members which might be practically realizable and which comprises significant volatility.

\(^{29}\)See for example Volkova (1996), Hong et al. (2009), Hong and Meeker (2010) and Hong (2011).
A.2 Modeling Subjective Mortality

Hoermann and Ruß (2008) propose a gamma distribution for modeling $D \sim \Gamma(\alpha, \beta, \gamma)$ with density function

$$f_{(\alpha, \beta, \gamma)}(d) = \frac{1}{\Gamma(\alpha) \beta^\alpha} (d - \gamma)^{\alpha-1} e^{-\frac{d-\gamma}{\beta}},$$

expected value

$$E(D) = \alpha \beta + \gamma$$

and variance

$$Var(D) = \alpha \beta^2$$

for $d \geq \gamma$, $\gamma \in \mathbb{R}$ and $\alpha, \beta > 0$. 