Journal of Empirical Finance 58 (2020) 1–18

Contents lists available at ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

Disaggregation and the equity premium puzzle

Matthew S. Wilson[1]

Economics Department, Binghamton University, 4400 Vestal Parkway East, Binghamton, NY 13902, USA

ARTICLE INFO

JEL classification:
E4
G1

Keywords:
Equity premium puzzle
Stocks
Bonds
Interest rates
Consumer Expenditure Survey

ABSTRACT

Standard macroeconomic models cannot explain why stocks so greatly outperform bonds. However, this result depends on the use of aggregate consumption data. If markets are incomplete, then a representative agent might not exist and it is necessary to use consumption data at the household rather than aggregate level. In the household data, I fail to reject the Euler equation when the coefficient of relative risk aversion is as low as 2.7–3.8. This result is robust in a very general framework and I prove that many of the tests used in the literature are biased.

1. Introduction

Mehra and Prescott (1985) could not explain the equity premium in a model that had many standard features: a representative agent, CRRA utility, and no frictions. The coefficient of relative risk aversion would have to be above 20 to fit the data, but that leads certainty equivalents that are unrealistically small (Mankiw and Zeldes, 1991). Thus, at least one of these foundational assumptions must be incorrect.

Many attempts have been made to solve the equity premium puzzle. Kocherlakota (1996) demonstrated that it depends upon complete markets, CRRA utility, and the absence of frictions. If there are complete markets and CRRA utility, then several aggregation theorems prove that there exists a representative agent who also has CRRA utility (Kocherlakota, 1996).

Some papers have relaxed the assumption of CRRA utility, but they have been largely unsuccessful in solving the puzzle. However, even if this worked, the larger issue would still remain. The puzzle is not just an unanswered question about asset pricing; it is a challenge to the foundations of macroeconomics. CRRA is still used very frequently. Showing that the puzzle can be solved with a different utility function would not justify all the models that rely upon CRRA. Of course, it is possible that the high equity premium is proof that CRRA is wrong. However, for the purposes of defending standard macroeconomics, it would be desirable if this assumption could be maintained. Ideally, a solution to the puzzles would keep CRRA utility. However, there would be some frictions or market incompleteness that is enough to solve the puzzle, but can safely be ignored when studying the business cycle or other questions of interest.

Kocherlakota (1996) demonstrates that models of incomplete markets have typically not been successful in addressing the puzzle.[2] Few would argue that markets are truly complete, but by several measures they are quite close.[3] Thus, market incompleteness may be of little assistance in solving the equity premium puzzle.

E-mail address: MSwilson@binghamton.edu.

[1] George Evans, Bruce McGough, an anonymous referee, and seminar participants at the University of Oregon Macro Group provided helpful feedback and discussion. Any remaining errors are my own.

[2] Barro (2006) developed Rietz's (1988) idea that the puzzle is solved if we allow for "disaster states" – a small, uninsurable risk that consumption plummets. This view has gained some traction, but Constantinides (2008) found a serious error. In the data, consumption falls gradually over four years in the average disaster state. However, in the model, the full drop in consumption takes place suddenly in a single year, exaggerating its effects. Julliard and Ghosh (2012) show that once this error is fixed, the equity premium is still a puzzle.

[3] For instance, see Martin and Klein (2010). They develop an index of consumption smoothing and use it on CEX data. If there are complete markets, then consumption will be very smooth and the index will be zero. The opposite extreme is autarky, which corresponds to an index value of one. In the data, they

https://doi.org/10.1016/j.jempfin.2020.05.002
Received 7 November 2018; Received in revised form 7 April 2020; Accepted 4 May 2020
Available online 7 May 2020

Nevertheless, my paper focuses on the role of incomplete markets and I find that the model is not rejected. Households have identical CRRA preferences. However, income will be heterogeneous, and there is no market for insurance against income shocks. Due to this missing market, we cannot appeal to Arrow–Debreu; there is no representative agent. Thus, the model will be tested using household consumption data rather than aggregate consumption. There are two key differences between these datasets. First, household consumption is more volatile. Second, in some cases, the negative covariance between the equity premium and marginal utility is stronger in the household data. I conclude that the Euler equation is not rejected, even when the coefficient of relative risk aversion is as low as 2.7–3.8.

Mankiw (1986) was one of the earliest papers in this field. He considered a very simple case in which there are two time periods and two states of the world. In the beginning, everyone is the same, but the bad state affects some agents more severely. E.g., a recession has a bigger impact on people who lose their job. There is no insurance against recessions. A parameter governs whether the bad state has a mild impact on a lot of people or a large effect on a few. Aggregate consumption is not affected by how narrowly this shock is concentrated. However, the model's equity premium does respond to it except in special cases. Thus, using aggregate consumption data to study the equity premium puzzle is inappropriate. Therefore, my paper and many others use micro data instead.

However, Constantinides and Duffie (1996) found an elegant way to continue using macro data. As in my model, there are uninsurable idiosyncratic shocks to income. However, their framework is far more complex. For certain functional forms, the Euler equation can be rewritten in terms of aggregate variables in spite of these shocks. Their results depend critically on the choice of functional forms. Some other papers have also claimed success, but they require adding an extensive list of assumptions to the standard macro framework. Constantinides and Ghosh (2017) rely heavily on functional forms that are designed to yield analytical results. Heaton and Lucas (1996) assume that transactions costs for stocks are substantially higher than for bonds and that labor shocks are very persistent. A strength of my paper is that the assumptions made about the idiosyncratic shocks are quite unrestrictive. This is important for the larger question behind the equity premium puzzle. It is much more than an unresolved asset pricing issue; it is a challenge to the fundamental framework of macroeconomics. To defend that framework, it is preferable to solve the puzzle with a minimal alteration of its basic assumptions.

Mankiw and Zeldes (1991) find that for stockholders, there is a much tighter correlation between consumption growth and the equity premium. As a result, stocks lose some of their appeal; they are less suitable for smoothing consumption. However, the puzzle is not fully resolved. Unfortunately, they also rely on data from the Panel Survey of Income Dynamics (PSID). A well-known issue with the PSID is that the only consumption data it tracks is food consumption. This is a questionable proxy for total consumption. For this reason, I use data from the Consumer Expenditure Survey (CEX) instead.

Vissing-Jørgensen (2002) also explores whether there are differences between stockholders and non-stockholders. However, her focus is on estimating the elasticity of intertemporal substitution (EIS) for each of these groups. On asset pricing, she writes, "it is too early to precisely determine the extent to which [limited asset market participation] helps resolve the equity premium puzzle". I considered whether to restrict the sample to stock and bondholders, but I rejected it for two reasons. First, there are limitations in the data. The CEX asks if respondents own securities, but it does not distinguish between stocks and bonds. If a household has only one of those assets but not the other, then only one of the Euler equations binds.[4] This is an issue since we have to combine both Euler equations in order to study the puzzle. Second, the larger goal of this paper is to defend the standard framework. To do so, I have to address the puzzle with only a minimal departure from usual assumptions. Adding limited asset market participation would be a very small alternation, but it would be even better if it were unnecessary. I can get reasonable results for risk aversion without having to introduce this extra feature into the model.

Kocherlakota and Pistaferri (2009) attempt to resolve both the equity premium and risk-free rate puzzles in three countries. Their model adds incomplete markets, private information, and idiosyncratic shocks. The US consumption data in the paper comes from the CEX, which tracks far more categories of consumption than the PSID does. They find that the equity premium puzzle can be solved if the coefficient of relative risk aversion is approximately five. This is a great improvement over the standard result, which requires a CRRA parameter in excess of 20. However, it is still too high to be plausible. The other issue with Kocherlakota and Pistaferri's (2009) results is that in order to solve the risk-free rate puzzle, the quarterly discount factor has to be below 0.5. This is far too low to be plausible. Basu et al. (2011) extend Kocherlakota and Pistaferri's (2009) model to explore other asset pricing puzzles and they modify the sample design. However, they concede that the results remain unsatisfactory. My paper focuses on the equity premium and uses lower values of risk aversion.

Jacobs (1999) also tries to resolve both puzzles. Using GMM, he estimates relative risk aversion from the two Euler equations: one for stocks and one for bonds. The results are mixed. In many specifications, the estimated risk aversion is reasonable, but sometimes the Euler equation is still rejected. He also adds demographic variables to the Euler equations. Though theoretical macro models typically do not do this, he offers a good justification: the demographic variables help the model fit the data better and they do not significantly alter the risk aversion estimates. One shortcoming of the paper is that it uses PSID data. This means that food consumption is treated as total consumption, but instrumental variables help deal with this measurement error issue.

However, for many papers on the equity premium, the conclusions are questionable due to the techniques used to obtain them. The standard approach is to compute the model's equity premium; this usually involves estimating means and covariances. As we all know, the sample mean and covariance are unbiased estimators of their population counterparts. However, the *ratio* of a sample

find that the index is 0.1278. This is high enough to reject complete markets but still close enough to zero that complete markets seem like a reasonable approximation.

[4] Under a different set of assumptions – which I use in this paper – the Euler equation is always binding. This is because households can hold negative quantities of stocks and bonds by shorting them. Thus, households that own no stocks or bonds are included in the sample.

covariance to a sample mean is biased. This is due to Jensen's inequality, as demonstrated in Appendix A. This problem affects Cogley (2002). After taking Taylor series approximations, he divides a covariance by a sample mean. As in my model, agents cannot buy insurance against shocks to household income. He claims that the equity premium is still a puzzle. His conclusion appears to be very robust. The main result stands even after he makes two possible corrections for measurement error, explores different criteria for excluding outliers, and considers alternative assumptions about how long people hold financial assets. However, in Section 4.3, I show that I would have wrongly rejected the Euler equation if I had used his test.

Some of the tests in Balduzzi and Yao (2007) also involve taking the ratio of a covariance to a mean. They observe that there are several different ways to construct the stochastic discount factor. Most papers aggregate over the intertemporal marginal rate of substitution (i.e. $\sum \left(\frac{c_{i,t+1}}{c_{it}}\right)^{-\sigma}$). Balduzzi and Yao (2007) note that aggregating over the marginal utilities (i.e. $\frac{\sum c_{i,t+1}^{-\sigma}}{\sum c_{it}^{-\sigma}}$) is equally valid. In their results, the marginal utility pricing kernel tends to perform better. Under some but not all specifications, the model is not rejected when risk aversion is low. I apply their test to my dataset and show that it is biased.

Brav et al. (2002) use a different test. They take a Taylor series approximation of the stochastic discount factor and rewrite it in terms of the cross-sectional mean, variance, and skewness of household consumption growth. One feature of this approach is that it nests both complete and incomplete markets. If markets are complete, then only the cross-sectional mean will matter. If instead markets are incomplete, then the higher order moments can be significant. They reject market completeness. In many specifications with incomplete markets, the model can match the equity premium when relative risk aversion is between two and four. These values are very reasonable. Unfortunately, they take the ratio of a household's consumption growth to the sample mean of consumption growth. Though the sample mean is an unbiased estimator, the same cannot be said of the *inverse* of the sample mean. The proof is in Appendix A.

Many papers rely upon GMM to estimate risk aversion (e.g. Jacobs, 1999; Vissing-Jørgensen, 2002; Kocherlakota and Pistaferri, 2009), but Toda and Walsh (2015) demonstrate that this procedure is not valid in the disaggregated data. Household consumption and its growth rate follow the power law. For these distributions, not all of the usual statistical rules apply. For most distributions, the sample moments converge to their population counterparts. However, for power law distributions, the moments might be infinite, so the sample estimates do not converge. Thus, GMM is not appropriate. In addition, GMM is vulnerable to type II errors. In the "sequel", Toda and Walsh (2017) use simulations to show these issues and find even more problems with GMM. Though I use a version of the method of moments, I demonstrate that my hypothesis tests are still valid if the measurement error is sufficiently small.

My paper shows that there is no puzzle in the disaggregated data and makes several important contributions. First, the framework that I use is far more general. The results do *not* require extensive assumptions about the processes of the idiosyncratic shocks. The sole departure from the standard baseline of frictionless markets and identical CRRA preferences is that I allow for uninsurable income shocks. This is important for addressing the larger question behind the puzzle. The puzzle challenges the foundations of macroeconomics. It suggests that macro models fit the data too poorly to be realistic. To defend those foundations, the puzzle has to be dealt with in way that minimally alters the usual macro framework. Solving the puzzle in a model that is substantially different does not justify all the research that relies on the standard framework. Second, I show which tests are unbiased and robust to measurement error and which ones are not. Though there has been much discussion of the equity premium puzzle, relatively little attention has been paid to the statistical tests in these papers. Depending on the specification, the CRRA parameter can be as low as 2.7–3.8 – all very plausible values.

Before discussing the data sources, a review of the puzzle is in order.

2. The puzzle

All households have identical CRRA preferences, and solve the problem Eq. (1)

$$\max E_{t_0} \sum_{t=t_0}^{\infty} \beta^t \left(\frac{c_{it}^{1-\sigma} - 1}{1-\sigma} \right) \textit{subject to}\, c_{it} + b_{it} + s_{it} = y_{it} + R_t^s s_{i,t-1} + R_t^b b_{i,t-1}. \tag{1}$$

Here c_{it} is consumption per capita by household i at time t; b_{it} and s_{it} are its bond and stock holdings, respectively. Income is y_{it}. The gross interest rate is R_t^b for bonds and R_t^s for stocks. The household chooses c_{it}, s_{it}, and b_{it} while treating R_t^b, R_t^s, and y_{it} as exogenous. The Euler equations are Eqs. (2) and (3)

$$c_{it}^{-\sigma} = \beta E_t R_{t+1}^s c_{i,t+1}^{-\sigma} \tag{2}$$

$$c_{it}^{-\sigma} = \beta E_t R_{t+1}^b c_{i,t+1}^{-\sigma}. \tag{3}$$

With the Law of Iterated Expectations, these can be rewritten as

$$E\left(\left(\frac{c_{i,t+1}}{c_{it}} \right)^{-\sigma} (R_{t+1}^s - R_{t+1}^b) \right) = 0 \tag{4}$$

Let C_t denote aggregate consumption per capita in time t. Define h_{it} to be the ratio of household to aggregate consumption per capita: $h_{it} \equiv \frac{c_{it}}{C_t}$. This can be plugged into Eq. (4).

$$E\left(\left(\frac{h_{i,t+1} C_{t+1}}{h_{it} C_t} \right)^{-\sigma} (R_{t+1}^s - R_{t+1}^b) \right) = 0 \tag{5}$$

Table 1
The equity premium puzzle, based on seasonally adjusted quarterly aggregate data. The model predicts that the mean of M is zero.

Coeff. of relative risk aversion (σ)	Mean M	Std. error	P-value
1	0.0191	0.0081	0.020
2	0.0190	0.0080	0.020
3	0.0189	0.0080	0.021
4	0.0187	0.0080	0.022
5	0.0186	0.0080	0.022
10	0.0180	0.0080	0.026
20	0.0168	0.0079	0.036
30	0.0157	0.0079	0.051

Table 2
Summary statistics, based on seasonally adjusted quarterly aggregate data.

	C_{t+1}/C_t	R_t^s	R_t^b
Sample mean	1.0030	1.0210	1.0017
Standard error	0.0005	0.0081	0.0009
Correlation with C_{t+1}/C_t	1.0000	0.1923	0.1004

All households have identical preferences, but there is no need to assume that the realizations of their income y_{it} will be identical. If there are complete markets, then households will trade state-contingent claims to insure themselves against income risk. Since they can perfectly insure themselves against idiosyncratic shocks, in equilibrium, only aggregate shocks can affect their consumption. Thus, $h_{i,t+1}/h_{it}$ will be equal to one. The proof is in Appendix B.

The complete markets assumption is critical: with it, Eq. (5) becomes Eq. (6)

$$E\left(\left(\frac{C_{t+1}}{C_t}\right)^{-\sigma}(R_{t+1}^s - R_{t+1}^b)\right) = 0 \tag{6}$$

A representative agent exists, and the problem can be studied using aggregate consumption data. Without complete markets, there is no guarantee that $\frac{h_{i,t+1}}{h_{it}}$ will be equal to one, and therefore I need consumption data at the household rather than aggregate level. For the moment, continue to assume that markets are complete; later I will relax that assumption.

I use quarterly data on aggregate personal consumption expenditures and divide by population. As usual, the discount factor β is calibrated to 0.99. As in Mehra and Prescott (1985) and Kocherlakota (1996), values for R_t^b come from the interest rate on 90-day Treasury bills and R_t^s is based on the S&P 500 Index. The return on stocks includes both capital gains and dividends. All of these values are deflated by the CPI.

To match the household level data that I will use later, the time interval is 1989Q1–2013Q4, excluding 1996Q1 (Section 3 explains why this quarter was dropped). Reinserting 1996Q1 does not change the aggregate results by much. Following Kocherlakota (1996), I will exploit the fact that the unconditional expectation can be estimated with the sample average. Thus, in Eq. (7) below, the sample mean of M should not be significantly different from zero.

$$E\left(\left(\frac{C_{t+1}}{C_t}\right)^{-\sigma}(R_{t+1}^s - R_{t+1}^b)\right) \equiv E(M) = 0 \tag{7}$$

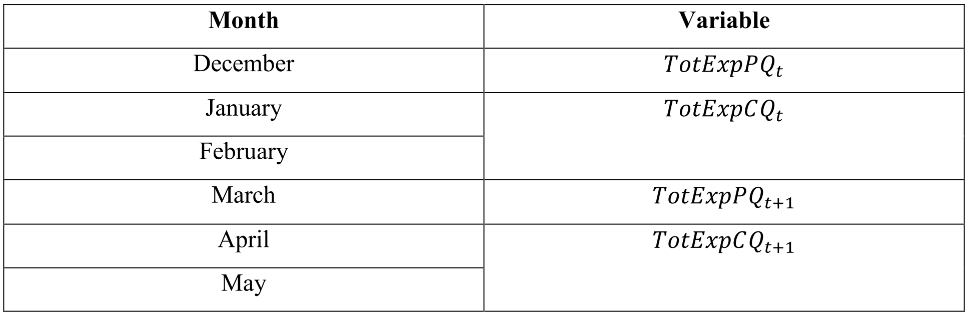
The equity premium is $R_{t+1}^s - R_{t+1}^b$; Table 1 demonstrates why it is so puzzling.

Relative risk aversion (σ) has to be as high as 30, and even then we just barely fail to reject the Euler equation. Though there is evidence of serial correlation, correcting for it does not alter the overall picture. I used the Prais–Winsten estimator to address the autocorrelation and found that σ had to be at least 20 for us to not reject the Euler equation. The conclusion is clear: the data can be reconciled with the Euler equation only if we allow risk aversion to be implausibly high.

Overall, reasonable values for σ imply that M is positive. This means that households should be buying more stocks and fewer bonds. As Table 2 shows, on average stocks yield substantially more than bonds. However, stocks are more tightly correlated with consumption growth. This means that they are not a great vehicle for smoothing consumption; stocks tend to be high when the marginal utility of consumption is low. Stocks are also much more volatile. Nevertheless, the magnitude of the equity premium outweighs all these concerns; why are not people buying more stocks?

The means[5] in Table 2 appear to be much lower than in Mehra and Prescott (1985). This is because my data is quarterly rather than annual. The annual real return on stocks, $\left(R_t^s\right)^4 - 1$, is about 8.7%; for bonds it is 0.7%. The puzzle is still present if annual instead of quarterly data is used in my sample period of 1989–2013.

[5] The means in the table are arithmetic. The geometric means are nearly identical except for stocks. Its 1.0177 geometric mean implies a 7.3% annual return. In any case, there is a large gap between stock and bond returns.



Month	Variable
December	$TotExpPQ_t$
January	$TotExpCQ_t$
February	
March	$TotExpPQ_{t+1}$
April	$TotExpCQ_{t+1}$
May	

Fig. 1. Consumption variables in the CEX for a household with interview t at the end of February.

All of the empirical results in this section depend on the complete markets assumption. If instead markets are incomplete, then a representative agent might not exist. This suggests that the appropriate variable is consumption at the household level rather than in aggregate.

3. Data

Fortunately, there is a data set that has detailed information on household-level consumption. This is the Consumer Expenditure Survey (CEX). The Interview Survey of the CEX is a panel survey of about 5000 households per quarter. On the first of five interviews, basic demographic information is gathered (BLS, 2013). The next four interviews obtain data on the household's consumption in the last three months (BLS, 2013). Each quarter, about 20% of the sample completes its fifth and final interview (BLS, 2013). It then leaves the sample and is replaced with a new household (BLS, 2013). Households may refuse to answer some questions or even skip one of the interviews. However, even if it stops responding after the first interview, it is not replaced until the quarter after the fifth interview.

Though households are often interviewed in the middle of a quarter, the CEX[6] is designed so that it is easy to reconstruct consumption for the entire quarter. Consider a household whose second interview is at the end of February. It is asked about its consumption in the last three months. That is broken down into two variables; the first is total expenditures in the current quarter (called "totexpcq" in the dataset) (BLS, 2013). This is all the consumption from January 1 to the date of the interview. The other variable is total expenditures from the previous quarter ("totexppq") (BLS, 2013). This contains all the consumption from the date of the previous interview three months earlier until December 31. Its third interview will be three months after its second. This will be in the end of May. Similarly, in this interview the total expenditures from the previous quarter will be all of March's consumption, while total expenditures from the current quarter will be April and May's consumption. Fig. 1 summarizes.

The formula for household i's consumption in the quarter t is straightforward.

$$c_{it} = totexpcq_{it} + totexppq_{i,t+1} \tag{8}$$

These total expenditure variables include rent if the family does not own its home. However, families that own their home do not pay any rent, so their spending does not reflect their consumption of housing. Fortunately, there is a variable to account for this. The CEX asks these households how much money they would get if they rented out their house (BLS, 2013); this captures the opportunity cost of living there. This variable, "renteqvx",[7] is monthly rent (BLS, 2013), so I multiply by three to obtain quarterly consumption of housing.

$$c_{it} = totexpcq_{it} + totexppq_{i,t+1} + 3 \cdot renteqvx_{it} \tag{9}$$

There is no seasonal adjustment in the CEX, so I deflate using the non-seasonally adjusted CPI. Note that there are at most four interviews that gather consumption data, and to construct each quarter's consumption, we need two consecutive interviews. Thus,

[6] Some sources refer to it as the CE. Strictly speaking, the CEX interviews "consumer units" rather than "households". From the documentation (BLS, 2013): "A consumer unit comprises either: (1) all members of a particular household who are related by blood, marriage, adoption, or other legal arrangements; (2) a person living alone or sharing a household with others or living as a roomer in a private home or lodging house or in a permanent living quarters in a hotel or motel, but who is financially independent; or (3) two or more persons living together who use their income to make joint expenditures. Financial interdependence is determined by the three major expense categories: housing, food, and other living expenses. To be considered financially independent, at least two of the three major expense categories have to be provided entirely or in part by the respondent". For the model, what matters is whether agents choose their consumption individually or as part of a group which I call a household. The CEX's definition of a consumer unit captures that idea, so I refer to consumer units as households throughout the paper.

[7] "Renteqvx" was not included in the 1993Q3–1995Q1 data sets. For those periods, I calculated consumption using Eq. (8). If consumption growth rates are calculated naively, this approach would create a large drop in 1993Q3 when "renteqvx" vanishes from the data, and a large spike in 1995Q2 when it reappears. These fluctuations would reflect changes in the CEX rather than changes in consumption, which is what we care about. However, there is a solution that completely fixes this problem. Eq. (8), which excludes "renteqvx", can be used in every period; this bridges the gap. For growth between 1993Q2 and 1993Q3, I used Eq. (8) for consumption in both periods and did the same for 1995Q1 and 1995Q2. However, I was able to use the more accurate Eq. (9) when finding growth after 1995Q1 and before 1993Q3.

Table 3
Real per capita growth rate of aggregate consumption.

	CEX	PCE
Mean	0.0031	0.0042
Standard error	0.0037	0.0040
F stat	0.07	
p-value	0.79	

we can compute no more than two consecutive quarters of consumption growth $c_{i,t+1}/c_{it}$ for each household. Nevertheless, tests using this data set will have high power, because there is a very large number of households. The PCE has a much smaller sample size, so all else equal, it will be harder to solve the puzzle in the CEX data. The likelihood that the Euler equation is not rejected will fall as the sample size rises. This is because the power of the test increases when the sample size is bigger.

The CEX sample is not fully representative. Even if the original sample reflected the general population, not all households are equally likely to respond. The CEX corrects for this by assigning weights to each household (BLS, 2013). The weights and sampling were redesigned in 1996Q1. The changes were dramatic enough that the BLS recommends not comparing results before and after the change (BLS, 1997). I dropped this quarter so that no household's consumption growth straddled the sample redesign period.

In one of the robustness checks, I use nondurable consumption instead of total consumption. Many papers use data on nondurables and there is a good reason for that. For durables, the expenditure is made only once, but actual consumption persists for several quarters. Both the aggregate and CEX data have no way to account for this since they only track expenditures. To match the BEA's definition of nondurables as closely as possible, I subtract household furniture and appliances, new and used vehicle purchases, education, reading, cash contributions, pension contributions, and personal insurance expenditures from total consumption. The Euler equation is still not rejected.

Outliers and measurement error are always a concern in papers that work with the CEX. The next section discusses the implications of measurement error in much more detail. Here is how I addressed outliers in the data. A handful of households had negative consumption or nondurable consumption that exceeded total consumption. These are obviously the result of errors, so I dropped them from the sample. Some of the observations are "topcoded". For instance, if a household had more than $51,323 in credit card debt, that would be recorded as just $51,323 (BLS, 2014a). This is done to protect the household's anonymity (BLS, 2014b), though it does introduce measurement error. Fortunately, most of the topcoding affects the income variables rather than the consumption variables. Income data is not necessary for testing the Euler equation. I excluded households with topcoded consumption. A handful of households had nondurable consumption that seemed implausibly low. Even a very frugal household would have difficulty living off less than $700 per person for three months of food and housing. Recall that housing includes the opportunity cost of not renting out your house. I dropped these observations as well. The results are not sensitive to the choice of the $700 cutoff. I experimented with other cutoff values and the results are similar. Reinserting these observations and only dropping the households with negative consumption does not make much of a difference either. Other criteria for excluding outliers (e.g., households with very low consumption growth) do not have much impact.

Questions have been raised about whether the CEX data is comparable with aggregates. This is an important concern to consider. If discrepancies between the CEX data and aggregate data are driving my results, then the findings here would be of little significance for macroeconomics. Measurement error is a potential source of discrepancies. Since the CEX Interview Survey is conducted once every three months, it does not ask about small purchases that people would be unlikely to remember (BLS, 2014a). Examples of such items are "housekeeping supplies, personal care products, and nonprescription drugs" (BLS, 2014b). However, these goods are a component of aggregate consumption, so it is possible that the CEX does not align with the aggregates. Nevertheless, Branch, 1994 finds that the CEX closely matches the Personal Consumption Expenditure (PCE) series. Branch (1994) and BLS (2014a) report that the CEX captures as much as 95% of total consumption. Still, 95% is not 100%, so the natural question is whether these discrepancies affect the results.

To address this issue, I calculated the growth rate of aggregate consumption in the CEX and compared it to the non-seasonally adjusted PCE. The results are in Table 3.

We are very far from rejecting the null that the growth rates are the same. The CEX compares well with aggregates. Notably, seasonal adjustment removes a large amount of volatility. For the seasonally adjusted PCE, the mean growth rate is 0.0030 and the standard error is just 0.0005. For most macroeconomic projects, the seasonally adjusted data is preferable; we prefer to focus on economic fundamentals without having to disentangle them from random temperature fluctuations or a surge in consumption before Christmas. However, this will have an impact on the equity premium puzzle. Seasonal adjustment artificially smooths the consumption data. The relevant variable is *actual* consumption growth, not consumption growth in a hypothetical season-free world. The seasonally adjusted data will lead us to overestimate the degree to which agents smooth their consumption, biasing σ upwards. As expected, Ferson and Harvey (1992) obtained lower estimates of σ when they used non-seasonally adjusted data. A number of other papers have argued that seasonal factors are important in asset pricing, though they tend to also reject time separability (Heaton, 1995; Ferson and Harvey, 1992; Garrett et al., 2005; Clive et al., 2000; Chang and Huang, 1990). Thus, seasonal factors alone are not enough to resolve the puzzle. In my sample, the puzzle still exists in the non-seasonally adjusted PCE data. As a result, we turn to another attempt to solve the puzzle: disaggregated household data.

Table 4
The equity premium puzzle, based on quarterly CEX data for household consumption. The model predicts that the mean of m is zero.

Coeff. of relative risk aversion (σ)	Mean m	Std. error	P-value
1	0.0201	0.0002	0.000
2	0.0276	0.0009	0.000
3	0.0627	0.0147	0.000

Table 5
The equity premium puzzle, based on quarterly CEX data for household consumption and clustered standard errors. The model predicts that the mean of m is zero.

Coeff. of relative risk aversion (σ)	Mean m	Std. error	P-value
1	0.0201	0.0094	0.035
2	0.0276	0.0127	0.032
2.8	0.0501	0.0245	0.043
2.9	0.0558	0.0281	0.050
3	0.0627	0.0328	0.059
4	0.3390	0.3660	0.357
5	4.7021	7.9716	0.557

4. Results

4.1. Euler equation test

Following Kocherlakota (1996), I test the Euler equation as in Section 2, except that this time I use disaggregated household consumption growth. Table 4 presents the results for several different calibrations of the CRRA parameter (σ). There are several reasons why this approach is superior to estimating σ. First of all, it is conceivable that the Euler equation is still rejected even for the σ that is relatively best. Jacobs (1999) acknowledges this issue. Secondly, remember that the puzzle is about much more than risk aversion and rates of return — it is about the foundations of macroeconomics. The σ that is the best fit for the portfolio data might not be the best for other macro issues. Imagine a macroeconomist calibrating or estimating relative risk aversion in their model. They may be primarily interested in the business cycle or some other macro question; the equity premium is secondary. However, they would like to check if their CRRA parameter is compatible with the equity premium data. Otherwise, the foundations of the model would be questionable. Tables showing the results for different values of σ would allow them to do this. Lastly, Toda and Walsh (2015) show that many procedures for estimating σ from the Euler equations are not valid.

$$E\left(\left(\frac{c_{i,t+1}}{c_{it}}\right)^{-\sigma}(R^s_{t+1}-R^b_{t+1})\right)\equiv E(m)=0 \tag{10}$$

At first, it appears that the puzzle is not solved at all. In fact, the results actually appear *worse* here than in the aggregate data (Table 1). However, it is possible that the OLS standard errors are inappropriate in this case; at each point in time, the observations across households might not be independent. I tested Eq. (10) again and used clustered standard errors.

The Euler equation is not rejected when the CRRA parameter (σ) is as low as 2.9. If Mehra and Prescott had studied the equity premium with this dataset, they would not have rejected the model or found that there was a puzzle to explain. Notably, the point estimates of m are larger than in the aggregate data (Table 1). However, the clustered standard errors are much higher. This expands the confidence interval, so the Euler equation is not rejected.

A natural question is whether we should cluster over time or over households. It is easy to imagine that observations from the same household would not be independent. However, this is not a concern due to the sample design. Recall from Section 3 that I can get – at most – two observations from each household for Eq. (10). Many households only yield one observation. This is because the BLS cannot force them to complete all five interviews. If there were more observations per household, then I would certainly explore clustering over households. However, when each household produces just one or two observations, then clustering over them makes little sense.

The case for clustering over time is stronger. There are about 100 clusters with an average of more than 2000 observations each. In each time period, all households may be subject to macroeconomic shocks. For example, a recession might affect the income of many households. Thus, consumption across households will not be independent. This does not violate the assumptions made earlier. When setting up the model, I assumed that households faced uninsurable, idiosyncratic income shocks. There is no requirement that these shocks are uncorrelated.

For further analysis, I rewrite Eq. (10) as follows.

$$E\left(\left(\frac{c_{i,t+1}}{c_{it}}\right)^{-\sigma}\right)E\left(R^s_{t+1}-R^b_{t+1}\right)+Cov\left(\left(\frac{c_{i,t+1}}{c_{it}}\right)^{-\sigma},(R^s_{t+1}-R^b_{t+1})\right)=0 \tag{11}$$

Table 6
Sample moments in CEX and PCE data when the CRRA parameter equals 2.9.

CRRA parameter $\sigma = 2.9$	$Cov\left(\left(\frac{c_{i,t+1}}{c_{it}}\right)^{-\sigma}, (R^s_{t+1} - R^b_{t+1})\right)$
Aggregate data (PCE)	−0.0002
Household data (CEX)	−0.0003

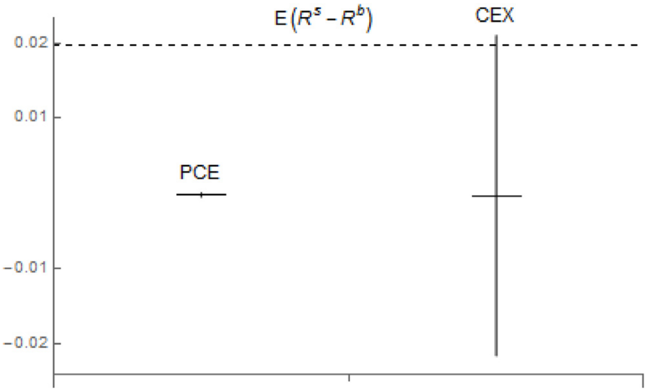


Fig. 2. 95% confidence intervals for $Cov\left(\left(\frac{c_{i,t+1}}{c_{it}}\right)^{-\sigma}, (R^s_{t+1} - R^b_{t+1})\right)$ in the aggregate data (PCE) and household data (CEX).

To go further, we have to pick a value for the CRRA parameter σ. Since the Euler equation was not rejected when $\sigma = 2.9$, I will use that value in Table 6.

In the PCE data, the high equity premium was somewhat (though not completely) mitigated by the covariance term. The covariance between excess returns and the marginal utility of consumption growth was negative. E.g., stocks were high when marginal utility was low. This diminished the appeal of stocks. It implies that they are not good for smoothing consumption. I will call this the *covariance effect.*

To visualize it more easily, imagine an agent who lost a job during a recession. Consumption would fall, so marginal utility would rise. This may be the time to cash in your stocks in order to stabilize your consumption. Unfortunately, stocks tend to fall when your marginal utility is high; this is the covariance effect. The same recession that took away your job is also harming your investments. Therefore, equities are not a good way to insure against shocks.

However, on average, stocks performed much better than bonds; this is the *premium effect.* It increases the appeal of stocks. In order to solve the puzzle, the premium effect and the covariance effect have to cancel out. However, in the aggregate data, that did not happen. The premium effect overwhelmed the covariance effect. That is how the equity premium puzzle arose.

In the disaggregated CEX data, the point estimate of the covariance is slightly more negative. However, this change, by itself, would not be sufficient to solve the puzzle. In the point estimates, the premium effect still outweighs the covariance effect. So why is the Euler equation not rejected? It is because the volatility of household consumption is higher. This is a consequence of market incompleteness. Since households cannot purchase insurance against shocks, their consumption fluctuates more. The standard errors increase. Therefore, the confidence interval for the covariance term widens, as shown in Fig. 2.

Remember that the equity premium $E\left(R^s_{t+1} - R^b_{t+1}\right)$ is about 0.02 since the data is quarterly, not annual (see Table 2 for summary statistics). For the aggregate data, there are no values in the confidence interval that are even close to working. The premium effect consistently dominates the covariance effect. However, that is not always true in the household data. The confidence interval is wide enough that the twin effects may cancel out. That is why the Euler equation is not rejected.

Cochrane (2005) also acknowledges the role of volatility. He rewrites the continuous-time version of Eq. (11) in terms of the Sharpe ratio:

$$\frac{|E\left(R^s - R^b\right)|}{SD\left(R^s - R^b\right)} \frac{1}{SD\,(consumption\ growth)} < \sigma \tag{12}$$

In his annual dataset of aggregate nondurable consumption, σ must be at least 33 — an absurdly high value. If we switch to quarterly non-seasonally adjusted data, the results improve during my sample period of 1989–2014. The Sharpe ratio is about 0.25 and the standard deviation of aggregate nondurable consumption growth rises to 0.03. Thus, σ has to be greater than eight, but this is still too high. In the household data, the standard deviation jumps to 0.3, so $\sigma > 5/6$. Of course, measurement error inflates the standard deviation and this result is merely a lower bound. "Cochrane's inequality Eq. (12)" is a test with low power, but nevertheless, models based on aggregate data failed to pass it. This shows the robustness of the puzzle in the aggregate data and highlights the importance of volatility in the CEX.

4.2. Measurement error

One concern is that measurement error may be driving the results. Additional noise in consumption could inflate the standard errors. Thus, the null might not be rejected even if it is false. However, the results are robust if the variance of the measurement error is sufficiently low. As before, c_{it} denotes consumption per capita by household i at time t. Let $\hat{c}_{it}$ denote the amount of c_{it} reported to the CEX. There is a multiplicative error ε_{it} shown in Eq. (13).

$$c_{it} = \hat{c}_{it}\varepsilon_{it} \tag{13}$$

One reason for using the growth rate rather than the level of consumption is that the errors might cancel out in Eq. (14).

$$\frac{c_{i,t+1}}{c_{it}} = \frac{\hat{c}_{i,t+1}\varepsilon_{i,t+1}}{\hat{c}_{it}\varepsilon_{it}} \tag{14}$$

For instance, if the CEX consistently misses 5% of total consumption, then $\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}} = 1$ and the growth rate is completely unaffected.

Eq. (10) can be rewritten as follows in Eq. (15).

$$E\left(\left(\frac{\hat{c}_{i,t+1}}{\hat{c}_{it}}\right)^{-\sigma}\left(R^s_{t+1} - R^b_{t+1}\right)\right)E\left(\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right)^{-\sigma}\right) + Cov\left(\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right)^{-\sigma}, \left(\frac{\hat{c}_{i,t+1}}{\hat{c}_{it}}\right)^{-\sigma}\left(R^s_{t+1} - R^b_{t+1}\right)\right) = 0 \tag{15}$$

Following Kocherlakota (1996), I test Eq. (16).

$$E\left(\left(\frac{\hat{c}_{i,t+1}}{\hat{c}_{it}}\right)^{-\sigma}\left(R^s_{t+1} - R^b_{t+1}\right)\right) = E\,(m) = 0 \tag{16}$$

These tests will be valid if the covariance term is zero and $E\left(\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right)^{-\sigma}\right)$ is one. It is difficult to see why the covariance would *not* be zero; why should the errors correlate with $\left(\frac{\hat{c}_{i,t+1}}{\hat{c}_{it}}\right)^{-\sigma}\left(R^s_{t+1} - R^b_{t+1}\right)$? This would suggest that people get better or worse at reporting their consumption on the CEX survey depending on their marginal utility of future consumption growth. It seems highly implausible that such a dependence would exist.

Earlier, I showed that the growth rate of consumption in the CEX is about the same as in the PCE. Though this is evidence that $E\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right) = 1$, it does *not* necessarily follow that $E\left(\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right)^{-\sigma}\right) = 1$. Kocherlakota and Pistaferri (2009) overlook this. In fact, Jensen's inequality implies $E\left(\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right)^{-\sigma}\right) > 1$ when $E\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right) = 1$ and $\sigma > 0$. Brav et al. (2002) note that the expected value of m will still be zero if the null is true. However, the point estimates of m will be biased away from zero. Thus, the test will be more likely to reject the Euler equation.

The other reason for the higher variance is measurement error in the CEX. By itself, it would make the tests less likely to reject the Euler equation. On the other hand, the mean estimates are biased away from zero, increasing the chance that the model is wrongly rejected. Which effect is stronger? I ran 10,000 simulations; since the CEX captures 95% of total consumption (Branch, 1994), I set the mean of ε_{it} equal to 0.95. However, I could not find a single paper that estimated the variance. Perhaps this should not be a surprise. By comparing consumption per capita in the CEX and the PCE, you can easily find the average measurement error. However, since each household's true consumption is unknown, the standard deviation remains a mystery. Geisen et al. (2011) found a solution. They re-interviewed a sub-sample of CEX respondents and asked for receipts. The receipts revealed true consumption, which usually differed from what the subjects reported on the survey. This information could be used to find the standard deviation of the measurement error, but the authors did not do that. Instead, they reported an unusual statistic that is similar to the average percent deviation (the details are in Appendix C). However, if I assume a distribution for the measurement error, I can calibrate the standard deviation to match the statistic in Geisen et al. (2011). I started by assuming that ε_{it} is normal and i.i.d. In that case, the standard deviation is 0.053. However, Toda and Walsh (2015) suggest a different distribution: the double Pareto. It has fatter tails than the normal distribution does. Eq. (17) shows the PDF when the mean is normalized to one.

$$f\left(\varepsilon_{it}\right) = \begin{cases} \frac{b}{2}\varepsilon_{it}^{-1-b} if \varepsilon_{it} \geq 1 \\ \frac{b}{2}\left(2 - \varepsilon_{it}\right)^{-1-b} if \varepsilon_{it} \leq 1 \end{cases} \tag{17}$$

Fig. 3 displays the PDFs. The first graph is zoomed out, but it is not entirely clear that the double Pareto distribution has fatter tails. The second graph zooms in to show this. For comparability, both distributions below have a standard deviation of 0.053.

To match Geisen et al., 2011, I start by calibrating the standard deviation of the double Pareto distribution to 0.057. In all of the simulations, the CRRA parameter is set to values between 1 and 5 since that range is used later in my results. Each simulation is for a dataset of 10,000 observations. The true mean of m is set to zero. I test whether the mean is zero for m with measurement error (i.e., $m\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right)^{-\sigma}$). Ideally, the null would be rejected 5% of the time; that would indicate that the hypothesis tests were still valid.

The first case to consider is normally distributed measurement error. Fig. 4 shows the results. The null is still rejected about 5% of the time even though measurement error affects both the mean and variance. Thus, the hypothesis tests are still reliable.

How severely does the measurement error affect the mean? In this case, the impact is quite modest. We would like to have data on m, but instead we observe $m\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right)^{-\sigma}$. The "inflation" or "bias" term is $\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right)^{-\sigma}$. Fig. 5 shows the inflation term for different values of the CRRA parameter.

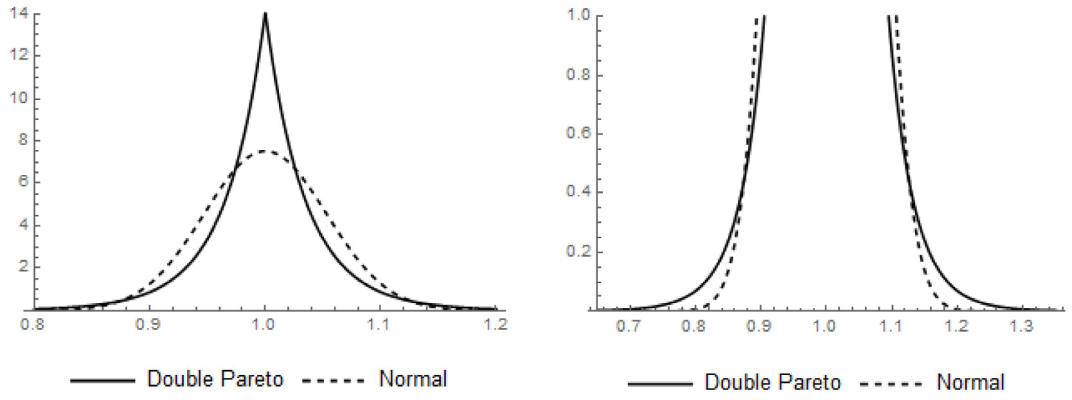


Fig. 3. PDFs of a normal distribution and a double Pareto distribution.

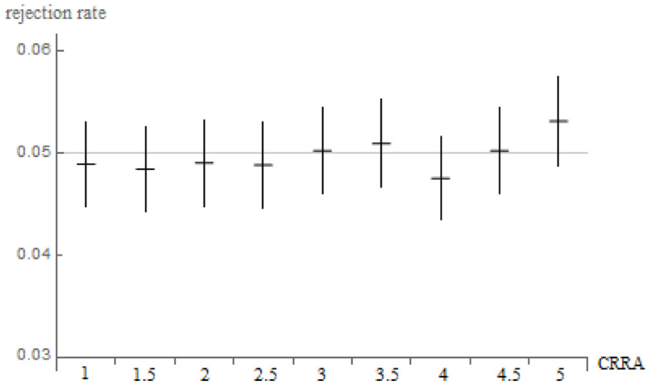


Fig. 4. 95% confidence intervals for the rejection rate of the null hypothesis in the simulations, assuming that the measurement error is normally distributed.

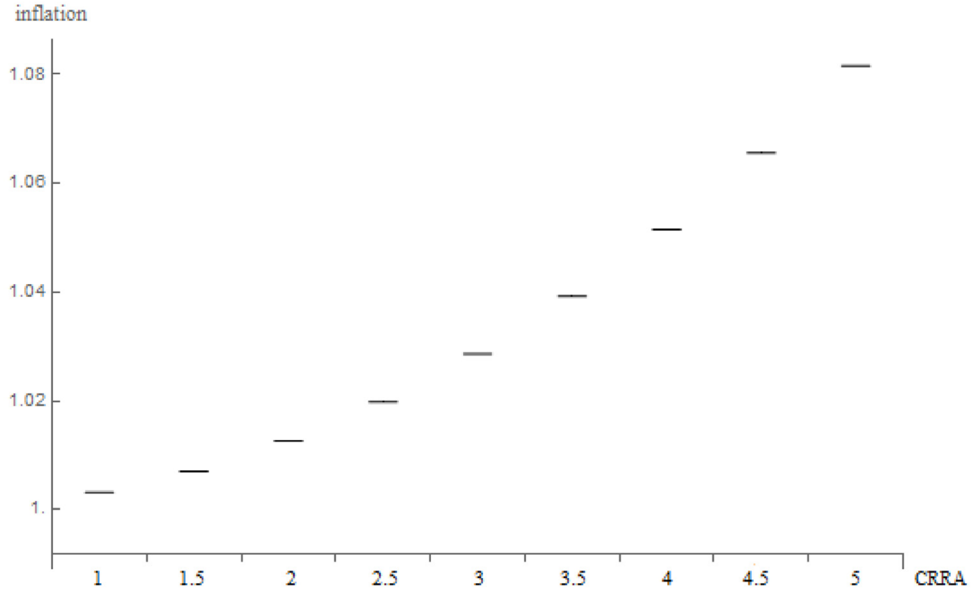


Fig. 5. 95% confidence intervals for the inflation term $\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right)^{-\sigma}$.

The bias is larger as the CRRA parameter increases. This is not surprising because now the noise is being raised to a higher power. Nevertheless, at most the bias is only about 8%.

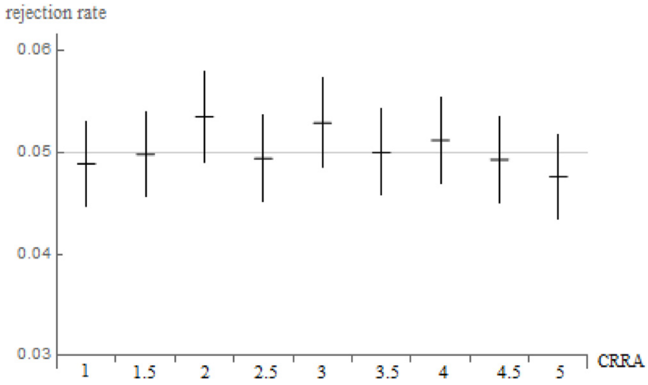


Fig. 6. 95% confidence intervals for the rejection rate of the null hypothesis in the simulations, assuming that the measurement error is normally distributed.

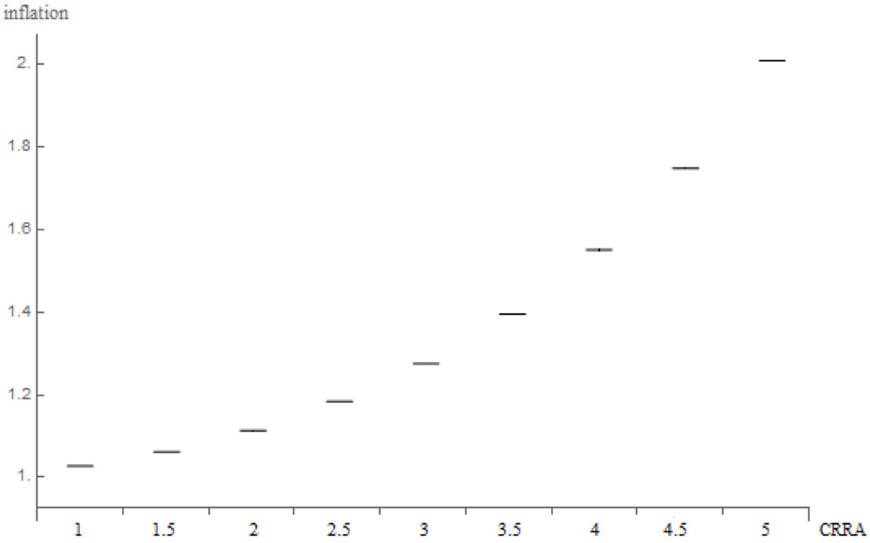


Fig. 7. 95% confidence intervals for the inflation term $\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right)^{-\sigma}$, assuming that the measurement error is normally distributed.

However, Appendix C shows that the calibration of the measurement error depends on a lot of assumptions. For robustness, I also explored how much measurement error can be tolerated before the hypothesis tests become unreliable. I tried increasing the standard deviation of ε_{it} to 0.15. Fig. 6 displays the implications for the rejection rate.

When the standard deviation is 0.15, the null is still rejected about 5% of the time, so all the tests remain valid.

When there is a lot of noise, then it is hard to reject the null. The noise also increases the bias. When the standard deviation increases, there is a larger chance that $\varepsilon_{i,t+1}$ and ε_{it} are far apart. Thus, the bias term $\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right)^{-\sigma}$ can be a very large number. Fig. 7 shows how severe the inflation can be.

The next case to consider is measurement error that follows a double Pareto distribution. Initially, I set the standard deviation to 0.057 to match (Geisen et al., 2011). Figs. 8 and 9 show the results.

All of the hypothesis tests remain valid. However, the inflation factor is noticeably more severe than when the measurement error was normally distributed. This is because the double Pareto distribution generates more outliers. As a result, the hypothesis tests are less robust to measurement error. If the standard deviation is increased to 0.07, then some of the tests start to reject the null too rarely. Fig. 10 shows the rejection rate and Fig. 11 displays the inflation term.

The outliers had a massive impact on the average bias. In the most severe case (CRRA parameter = 5), the median for the bias term was less than two, though the mean is off the charts.

In summary, the hypothesis tests remain reliable as long as the standard deviation of the measurement error is low. If the measurement error follows a normal distribution, the standard deviation can be as high as 0.15. However, if the distribution is double Pareto, then the standard deviation has to be below 0.07. The measurement error increases both the mean and the standard deviation of m. The higher average causes the model to be rejected more often, but the higher standard deviation makes it harder to reject. These effects approximately cancel out when the measurement error is small. However, if there is a lot of noise or if it is

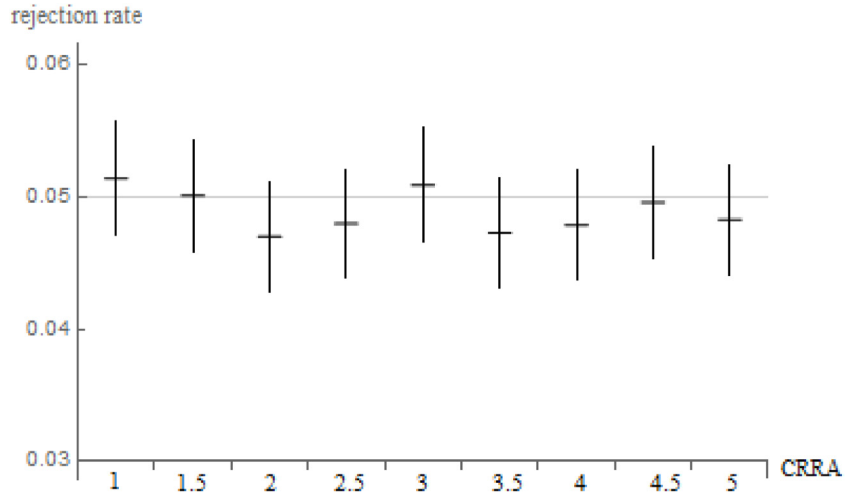


Fig. 8. 95% confidence intervals for the rejection rate of the null hypothesis in the simulations, assuming that the measurement error follows a double Pareto distribution.

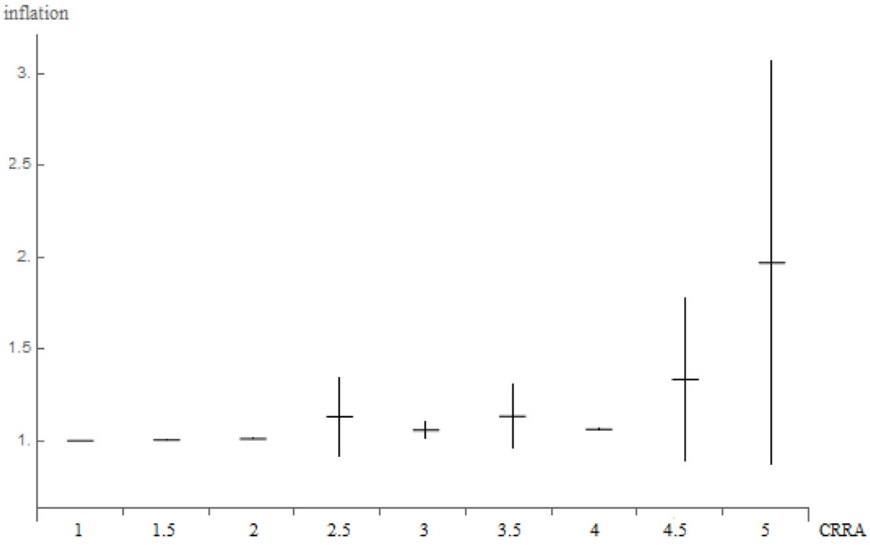


Fig. 9. 95% confidence intervals for the inflation term $\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right)^{-\sigma}$, assuming that the measurement error follows a double Pareto distribution.

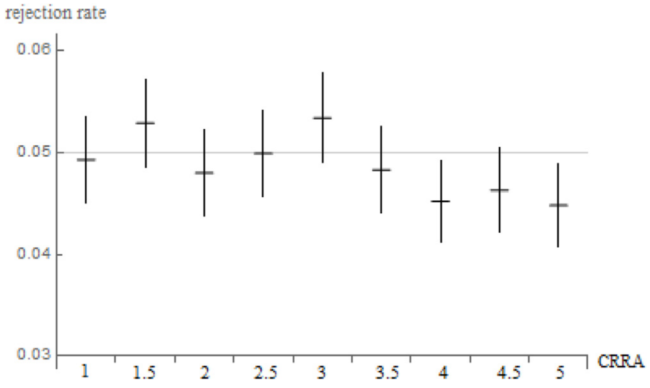


Fig. 10. 95% confidence intervals for the rejection rate of the null hypothesis in the simulations, assuming that the measurement error follows a double Pareto distribution.

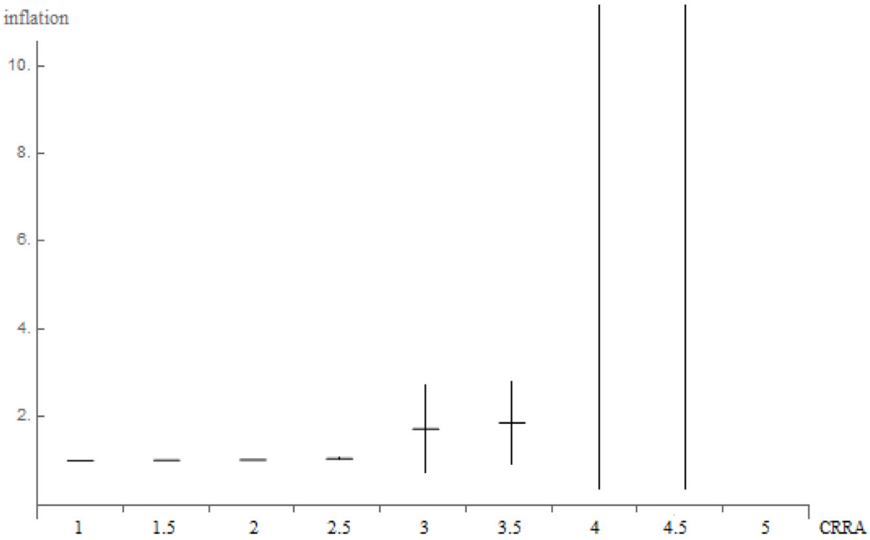


Fig. 11. 95% confidence intervals for the inflation term $\left(\frac{\varepsilon_{i,t+1}}{\varepsilon_{it}}\right)^{-\sigma}$, assuming that the measurement error follows a double Pareto distribution.

Table 7
Sample estimates of m adjusted for measurement error. The model predicts that the mean of m is zero.

CRRA term(σ)	Normal distribution				Double Pareto distribution			
	$SD\left(\varepsilon_{it}\right)=0.053$		$SD\left(\varepsilon_{it}\right)=0.15$		$SD\left(\varepsilon_{it}\right)=0.057$		$SD\left(\varepsilon_{it}\right)=0.07$	
	Mean m	Std. error	Mean m	Std. error	Mean m	Std. error	Mean m	Std. error
1	0.0200	0.0094	0.0196	0.0092	0.0200	0.0094	0.0200	0.0093
2	0.0273	0.0125	0.0248	0.0114	0.0272	0.0125	0.0270	0.0124
3	0.0609	0.0319	0.0492	0.0257	0.0591	0.0306	0.0364	0.0164
4	0.3234	0.3481	0.2187	0.2360	0.3179	0.3428	0.0008	0.0006
5	4.3473	7.3700	2.3416	3.9671	2.3834	3.4166	0.0000	0.0000

raised to a high power, then the tests are no longer trustworthy. The impact of the larger standard errors outweighs the bias in the mean, and as a result the null is rejected too rarely.

The inflation terms can be used to de-bias the sample mean of m. Appendix C explains how the standard errors are corrected. Table 7 considers four cases. First, there are the baseline calibrations for the normal distribution and the double Pareto distribution. Then there are the extreme cases: the maximum amount of measurement error that can be tolerated.

For $\sigma = 1$, the point estimate is nearly identical to its value in the aggregate data (Table 1). After that, the point estimates rise as σ rises. The exception is when σ is four or higher in the last two columns. In those cases, the average of the inflation term was more than four hundred(!), but that was driven by outliers and there was an extremely wide confidence interval.

Thus, while the model is not rejected when $\sigma = 3$, no GMM test would produce that estimate – even if it accounted for measurement error. However, as stated above, my goal was not to find the σ that best fits the equity premium. Instead, I wanted to find which values of σ currently in use were compatible with the data. This addresses the larger issue of whether the standard framework in macro is defensible.

4.3. Other tests used in the literature

In Appendix A, I prove that many of the tests used in the literature are biased. If I had used those tests, I would have wrongly concluded that the null is rejected. Cogley (2002) estimated the model's equity premium (Eq. (18)) with first-, second-, and third-order approximations.

$$E\left[R^s_{t+1}-R^b_{t+1}\right] = -\frac{Cov\left(R^s_{t+1}-R^b_{t+1}, \left(\frac{c_{t+1}}{c_t}\right)^{-\sigma}\right)}{E\left[\left(\frac{c_{t+1}}{c_t}\right)^{-\sigma}\right]} \tag{18}$$

Then he checks if a 7% annual equity premium is in the confidence interval. The model almost invariably fails this test, so he concludes that the puzzle is unresolved. The problem is that while the sample mean and covariance are unbiased, the *ratio* of a sample covariance to a sample mean is not.

In Table 5, an unbiased test demonstrated that the model is not rejected if the CRRA parameter is at least 2.9. With a Cogley test, I would have reached a different conclusion. Fig. 12 shows the results for a third-order Taylor series approximation. The graph

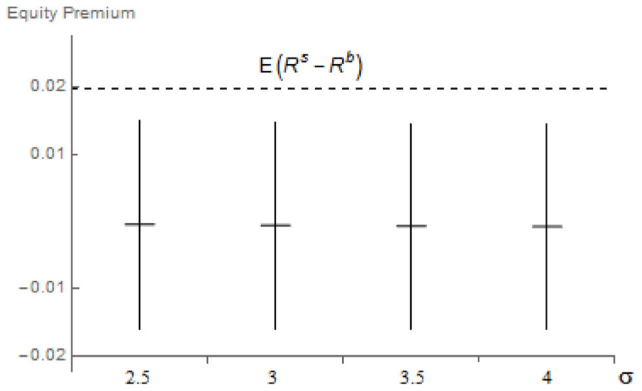


Fig. 12. Cogley test of the model's equity premium and 95% confidence interval, based on a third-order Taylor series approximation.

Table 8
Second-and third-order Brav tests. The model predicts that the mean of m is zero.

Coeff. of relative risk aversion (σ)	Second-order			Third-order		
	Mean m	Std. error	P-value	Mean m	Std. error	P-value
1.0	0.0228	0.0090	0.013	0.0031	0.0056	0.586
1.5	0.0262	0.0101	0.011	−0.0153	0.0062	0.015
2.0	0.0303	0.0115	0.009	−0.0424	0.0114	0.000
2.5	0.0353	0.0131	0.008	−0.0793	0.0203	0.000
3.0	0.0408	0.0150	0.008	−0.1269	0.0323	0.000
3.5	0.0470	0.0170	0.007	−0.1856	0.0473	0.000
4.0	0.0535	0.0192	0.006	−0.2561	0.0656	0.000

for the first- and second-order approximations is very similar. As before, since my data is quarterly, the empirical equity premium is about 2% rather than 7%. This is consistently outside the model's confidence intervals. My sample period is different from Cogley's. However, that is not driving the results, since the premium that we calculate for the model is very similar.

Brav et al. (2002) took a different approach. Instead of calculating Taylor series approximations of the equity premium, they approximated the stochastic discount factor S_t. Then they tested if $E\left[S_t\left(R_t^s - R_t^b\right)\right] = 0$ using Eqs. (19) and (20).

$$g_{it} = \frac{c_{it}}{c_{i,t-1}} \tag{19}$$

$$E\left[S_t\left(R_t^s - R_t^b\right)\right] = \frac{\beta\left(R_t^s - R_t^b\right)}{\left(I^{-1}\sum_i g_{it}\right)^{\sigma}}\left(1 + \frac{\sigma(\sigma+1)}{2} I^{-1}\sum_{i=1}^{I}\left(\frac{g_{it}}{I^{-1}\sum_i g_{it}} - 1\right)^2\right) \equiv E(m) = 0 \tag{20}$$

This test is also biased because it takes the inverse of a sample mean in the $\left(\frac{g_{it}}{I^{-1}\sum_i g_{it}} - 1\right)^2$ term. The *inverse* of a sample mean is biased, since $\frac{1}{E[X]} \neq E\left[\frac{1}{X}\right]$. The details are in Appendix A. However, the main results in their paper were based upon the third-order approximation (Eq. (21)).

$$E\left[S_t\left(R_t^s - R_t^b\right)\right] = \frac{\beta\left(R_t^s - R_t^b\right)}{\left(I^{-1}\sum_i g_{it}\right)^{\sigma}}\left(1 + \frac{\sigma(\sigma+1)}{2} I^{-1}\sum_{i=1}^{I}\left(\frac{g_{it}}{I^{-1}\sum_i g_{it}} - 1\right)^2 + \frac{\sigma(\sigma+1)(\sigma+2)}{6} I^{-1}\sum_{i=1}^{I}\left(\frac{g_{it}}{I^{-1}\sum_i g_{it}} - 1\right)^3\right) \equiv E(m) = 0 \tag{21}$$

Here is what happens when I apply their tests to my dataset.

In the second-order test, the null is consistently rejected. A striking reversal occurs in the third-order test: with σ between 1.5 and 4.0, I would have found that m is significant and *negative* – people should be buying fewer stocks and more bonds! However, the popular choice of $\sigma = 1$ (log preferences) would have passed the test.

These are very different from the results in Table 5, but there are two possible explanations. First, their test is biased. Second, they had a different sample. If sampling were not a factor, then the results in Table 8 should match what they found in their paper. However, they conclude that the sample mean is exactly zero when σ is between 3 and 4, not 1.0 and 1.5. The time period and the criteria for excluding outliers are different in their paper. They seasonally adjust their data, which is not appropriate in this context (see the discussion in Section 3). In addition, they drop the rural households from the sample. That is because the CEX only surveyed urban households in the early 1980s. Thus, in order to have comparable samples across time, you can either drop the years that were urban-only or drop the rural households in all years. They picked the latter option; I chose the former.

When I use their sampling criteria and their test in my time period of 1989–2013, the results are still different, as shown in Table 9. The time period is the only remaining difference; their paper is based on 1982–1996 data.

Table 9
Third-order Brav test for seasonally adjusted nondurable consumption by urban households. The model predicts that the mean of m is zero.

Coeff. of relative risk aversion (σ)	Mean m	Std. error	P-value
1.0	0.0181	0.0075	0.017
1.5	0.0174	0.0072	0.017
2.0	0.0165	0.0068	0.017
2.5	0.0153	0.0063	0.017
3.0	0.0137	0.0057	0.017
3.5	0.0117	0.0050	0.020
4.0	0.0093	0.0044	0.037
4.5	0.0063	0.0042	0.134
5.0	0.0028	0.0047	0.548
5.5	−0.0012	0.0061	0.850

Table 10
Euler equation test for nondurable consumption by urban households with clustered standard errors. The model predicts that the mean of m is zero.

Coeff. of relative risk aversion (σ)	Mean m	Std. error	P-value
1	0.0191	0.0090	0.037
2	0.0209	0.0100	0.040
2.6	0.0234	0.0117	0.048
2.7	0.0241	0.0122	0.051
3	0.0274	0.0148	0.066
4	0.1106	0.0992	0.268
5	1.7874	1.7492	0.309

Now the Euler equation is not rejected when σ is as low as 4.5. The sample mean is exactly zero for a value of σ that is between 5.0 and 5.5. In their paper, σ could be as low as 3–4, so the puzzle is slightly harder to solve in my sample period. Thus, the choice of sample period is not driving my results.

As an additional robustness check, I apply their sampling criteria to my data and use the unbiased Euler equation test. However, for reasons explained in Section 3, I do not make any seasonal adjustments. Specifically, this is a test of Eq. (10) for non-durable, non-seasonally adjusted consumption by urban households. Their criteria for outliers: "First, we delete from the sample households with consumption reported in fewer than three consecutive quarters. Second, we delete the consumption growth $\frac{c_{it}}{c_{i,t-1}}$ if $\frac{c_{it}}{c_{i,t-1}} < \frac{1}{2}$ *and* $\frac{c_{i,t+1}}{c_{it}} > 2$ [emphasis in the original]. Third, we delete the consumption growth $\frac{c_{it}}{c_{i,t-1}}$ if it is greater than five".

Table 10 shows that, with these criteria, the coefficient of relative risk aversion can be as low as 2.7 – a result that is almost the same as the original Euler equation test (see Table 5). Thus, my results are robust to different criteria for excluding outliers.

5. Additional tests

Initially, the results based on the CEX seem to vastly improve upon the results from the aggregate data. However, the reader may have noticed a potentially important substitution: the aggregate data relied on consumption *per capita*, while the CEX data used consumption *per household*. If households added or lost members during the survey, this could affect the results. Fortunately, the CEX includes data on how many people live in the household, so it is easy to derive the growth rate of consumption per capita within a household. This makes the CEX and aggregate results comparable.

Taxes are another issue to consider. Though the CEX gathers data on taxes, it is notoriously poor (BLS, 2014b), so many papers use the NBER's TAXSIM program. It is designed to calculate total tax liabilities and marginal rates from survey data (Feenberg et al., 1993). The tax code is complex enough that we cannot expect TAXSIM to be perfect, but it is an excellent approximation. It uses information on many major deductions, accounts for the Alternative Minimum Tax, and even computes state taxes.

I used TAXSIM to investigate whether taxes impact the puzzle. For stocks, returns can come in the form of capital gains or dividends. Let τ^k_{it} and τ^d_{it} denote household i's marginal tax rates at time t on capital gains and dividends, respectively. Taxes are levied on nominal returns, dividends are d_t, $r^s_t \equiv R^s_t - d_t - 1$, and $r^b_t \equiv R^b_t - 1$. The Euler equation becomes Eq. (22)

$$E\left(\left(\left(1 + r^s_{t+1}\left(1 - \tau^k_{i,t+1}\right) + d_{t+1}(1 - \tau^d_{i,t+1})\right) - (1 + r^b_{t+1}(1 - \tau^k_{i,t+1}))\right)\left(\frac{p_t}{p_{t+1}}\right)\left(\frac{c_{i,t+1}}{c_{it}}\right)^{-\sigma}\right) \equiv E\left(m_\tau\right) = 0. \tag{22}$$

One shortcoming in the data is that until 1993, the CEX did not record where respondents live, which is a problem when calculating state taxes. Another limitation is that the respondent's state is not recorded in some cases if the state has a low population. This is designed to protect confidentiality (BLS, 2014a). Clearly, TAXSIM cannot calculate state taxes if it does not know which state the respondent is from; in this case it can only compute federal taxes. Thus, the marginal rates are biased downwards.

Table 11
The equity premium puzzle and taxes, based on quarterly CEX data and clustered standard errors. The model predicts that the mean of m_τ is zero.

CRRA (σ)	Household consumption			Per capita			Per capita without taxes		
	Mean m_τ	Std. error	P-value	Mean m_τ	Std. error	P-value	Mean m	Std. error	P-value
1	0.0181	0.0085	0.035	0.0182	0.0085	0.035	0.0201	0.0094	0.035
2	0.0249	0.0115	0.033	0.0254	0.0119	0.036	0.0281	0.0131	0.035
2.6	0.0379	0.0180	0.038	0.0390	0.0193	0.046	0.0422	0.0209	0.046
2.7	0.0417	0.0202	0.042	0.0429	0.0217	0.051	0.0461	0.0234	0.051
2.8	0.0463	0.0230	0.047	0.0476	0.0248	0.058	0.0508	0.0266	0.059
2.9	0.0518	0.0266	0.054	0.0532	0.0288	0.067	0.0564	0.0306	0.068
3	0.0586	0.0313	0.064	0.0600	0.0339	0.080	0.0630	0.0358	0.082

Table 12
The equity premium puzzle and taxes, based on quarterly CEX data for non-durables and clustered standard errors. The model predicts that the mean of m_τ is zero.

Coeff. of relative risk aversion (σ)	Household consumption			Per capita		
	Mean m_τ	Std. error	P-value	Mean m_τ	Std. error	P-value
1	0.0174	0.0081	0.036	0.0174	0.0082	0.036
2	0.0202	0.0093	0.033	0.0206	0.0096	0.035
3	0.0370	0.0155	0.019	0.0380	0.0166	0.024
3.7	0.1325	0.0643	0.042	0.1335	0.0658	0.045
3.8	0.1701	0.0856	0.050	0.1709	0.0871	0.053

Table 11 shows that these factors do not make much of a difference. In some specifications, the coefficient of relative risk aversion may be as low as 2.7. Previously, it had to be at least 2.9.

All of the previous results have depended on total consumption expenditures; however, perhaps non-durable expenditures are a better metric. The concern is that consumption of durables (e.g., a new car) lasts far longer than three months. The CEX only tracks the one-time expenditure, which in this example is not a good measure of consumption. If you paid for the car in the first quarter of the year, it would appear that your consumption surged in that period and then dropped back to normal levels, when in reality consumption of the car persists through several quarters.

This noticeably affects the Euler equation test; $\sigma = 3$ no longer suffices. However, only a slight increase in risk aversion is necessary. For both household and per capita consumption, the Euler equation is not rejected when the CRRA parameter is 3.8. Table 12 presents the results.

My main results have focused on the unconditional mean. However, the model makes an even stronger claim: the *conditional* mean is zero.

$$E_t\left(\left(\frac{c_{i,t+1}}{c_{it}}\right)^{-\sigma}(R^s_{t+1}-R^b_{t+1})\right)=0 \tag{23}$$

If this is true, then there should be no autocorrelation. Recall that the dataset provides, at most, two consecutive observations for Eq. (23). Thus, there will be one lag in the Breusch–Godfrey LM test for autocorrelation.

Eq. (23) has a second implication. If I regress $\left(\frac{c_{i,t+1}}{c_{it}}\right)^{-\sigma}(R^s_{t+1}-R^b_{t+1})$ on lags of R^s_t and R^b_t, then the lagged variables should be insignificant in an F test. The results of these two tests are in Table 13.

The null of no autocorrelation is soundly rejected for the popular calibration $\sigma = 1$ (log preferences). However, for slightly higher levels of relative risk aversion, there are no problems. In particular, there is no evidence of autocorrelation when σ is between 2.7 and 3.8 – the range in which the Euler equation was not rejected in the earlier tests. In all of the calibrations, the lags are far from significant. The conditional Euler equation is not rejected.

6. Conclusion

The equity premium is about more than explaining rates of return. It calls into question the foundations of modern macroeconomics. The assumptions that cause the puzzle to emerge – complete markets, CRRA utility, little or no frictions – are nearly ubiquitous in macro models. Though few would claim that their models are a complete description of how the economy works, the gap between the predicted and observed equity premium is so large that our models appear indefensible.

However, there is no puzzle if we look at the household data instead. When an unbiased test is applied, the model is not rejected; the vast literature attempting to reconcile the equity premium with standard models would never have gotten started. The justification for using household data comes from a straightforward assumption: there is no market for insurance against idiosyncratic income shocks. This is quite realistic. When markets are incomplete, the existence of a representative agent is no longer assured. Thus, we cannot substitute aggregate consumption into the Euler equation; instead, we turn to the household data. In both datasets, there is a negative covariance between the equity premium and the stochastic discount factor. In the aggregate data, the covariance effect is swamped by the large premium on stocks, creating a puzzle. However, in the household data, the covariance effect can

Table 13
Conditional Euler equation tests. The model predicts that there is no autocorrelation. It also predicts that lags of R_t^s and R_t^b will be insignificant in the F test.

Coeff. of relative risk aversion (σ)	Autocorrelation test		F test	
	χ^2	P-value	F	P-value
1	372.484	0.000	0.27	0.762
2	1.383	0.240	0.26	0.774
2.7	0.001	0.976	0.30	0.738
2.8	0.001	0.974	0.33	0.722
2.9	0.001	0.975	0.36	0.702
3	0.001	0.977	0.39	0.678
3.1	0.001	0.980	0.43	0.651
3.2	0.000	0.982	0.48	0.622
3.3	0.000	0.985	0.52	0.595
3.4	0.000	0.987	0.57	0.569
3.5	0.000	0.988	0.61	0.547
3.6	0.000	0.990	0.64	0.528
3.7	0.000	0.991	0.67	0.512
3.8	0.000	0.992	0.70	0.499

offset the equity premium. Furthermore, this can be achieved with a reasonable amount of risk aversion. The CRRA parameter can be as low as 2.7–3.8.

However, it is important to keep in mind that the puzzle is about macro models in general — not just rates of return. If there is no puzzle once we dispense with the representative agent, what about all the models that retain that assumption? Fortunately, these models may still be defensible. Any macroeconomic model is a simplification of reality. Some features of the real world are necessarily excluded. The question should not be, ‘‘does the model exclude any variables that affect the real-world macroeconomy?’’ but rather, ‘‘does the model exclude any variables that are *important* to the macroeconomy?’’ If we are studying the equity premium, then market incompleteness is an important feature and should be included. However, if the model is primarily trying to explain features of the business cycle, then perhaps market incompleteness might not be so critical. Though assuming the existence of a representative agent creates large problems when focusing on the equity premium, it does not necessarily follow that the same assumption will cause issues when researching other macroeconomic questions.

Overall, these results show that the standard macro framework *can* explain the equity premium. The only modification required is that there is no insurance market for idiosyncratic income shocks. However, I cannot claim a complete victory for the standard macro framework. The popular calibration $\sigma = 1$ (log preferences) is rejected in nearly every test. Can this calibration be justified, or do we need to set the CRRA parameter to a slightly higher value? That is a question for future research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jempfin.2020.05.002.

References

Balduzzi, Pierluigi, Yao, Tong, 2007. Testing heterogeneous-agent models: an alternative aggregation approach. J. Monetary Econ. 54 (2), 369–412.

Barro, Robert J., 2006. Rare disasters and asset markets in the twentieth century. Q. J. Econ. 823–866.

Basu, Parantap, Semenov, Andrei, Wada, Kenji, 2011. Uninsurable risk and financial market puzzles. J. Int. Money Finance 30 (6), 1055–1089.

BLS (Bureau of Labor Statistics), 1997. Consumer Expenditure Interview Survey, 1996: Interview Survey and Detailed Expenditure Files. United States Department of Labor [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], Washington, DC, https://www.icpsr.umich.edu/icpsrweb/ICPSR/series/20/studies/2794?archive=ICPSR&amp;sortBy=7&amp;paging.startRow=26.

BLS (Bureau of Labor Statistics), 2013. Consumer Expenditure Interview Survey Public Use Microdata: 2012 Users' Documentation. United States Department of Labor, Washington, DC.

BLS (Bureau of Labor Statistics), 2014a. Consumer Expenditure Interview Survey Public Use Microdata: 2013 Users' Documentation. United States Department of Labor, Washington, DC.

BLS (Bureau of Labor Statistics), 2014b. Consumer Expenditure Survey: Frequently Asked Questions. United States Department of Labor, Washington, DC, http://www.bls.gov/cex/csxfaqs.htm#q23.

Branch, E. Raphael, 1994. The consumer expenditure survey: A comparative analysis. Mon. Labor Rev.

Brav, Alon, Constantinides, George M., Geczy, Christopher C., 2002. Asset pricing with heterogeneous consumers and limited participation: Empirical evidence. J. Polit. Econ. 110 (4).

Chang, Eric C., Huang, Roger D., 1990. Time-varying return and risk in the corporate bond market. J. Financ. Quant. Anal. 25 (3), 323–340.

Clive, Gaunt, Gray, Philip, McIvor, Julie, 2000. The impact of share price on seasonality and size anomalies in Australian equity returns. Account. Finance 40 (1), 33–50.

Cochrane, John H., 2005. Financial markets and the real economy. Found. Trends Finance 1 (1), 1–101.

Cogley, Timothy, 2002. Idiosyncratic risk and the equity premium: Evidence from the consumer expenditure survey. J. Monetary Econ. 309–334.

Constantinides, George M., 2008. Comment on barro and Ursùa. Brookings Pap. Econ. Activity 341–350.

Constantinides, G.M., Duffie, D., 1996. Asset pricing with heterogeneous consumers. J. Polit. Econ. 219–240.

Constantinides, George M., Ghosh, Anisha, 2017. Asset pricing with countercyclical household consumption risk. J. Finance 72 (1), 415–460.

Feenberg, Daniel, Richard, Coutts, Elizabeth, 1993. An introduction to the TAXSIM model. J. Policy Anal. Manag. 12 (1), 189–194, http://www.nber.org/taxsim/, Winter 1993.

Ferson, Wayne E., Harvey, Campbell R., 1992. Seasonality and consumption-based asset pricing. J. Finance 47 (2), 511–552.

Garrett, Ian, Kamstra, Mark J., Kramer, Lisa A., 2005. Winter blues and time variation in the price of risk. J. Empir. Financ. 12 (2), 291–316.

Geisen, Emily, Richards, Ashley, Strohm, Charles, Wang, Joan, 2011. US Consumer Expenditure Records Study. US Census Bureau.

Heaton, John, 1995. An empirical investigation of asset pricing with temporally dependent preference specifications. J. Econom. Soc. 681–717.

Heaton, John, Lucas, Deborah J., 1996. Evaluating the effects of incomplete markets on risk sharing and asset pricing. J. Polit. Econ. 104 (3), 443–487.

Jacobs, Kris, 1999. Incomplete markets and security prices: Do asset-pricing puzzles result from aggregation problems? J. Finance 54 (1), 123–163.

Julliard, Christian, Ghosh, Anisha, 2012. Can rare events explain the equity premium puzzle? Rev. Financ. Stud. 25 (10), 3037–3076.

Kocherlakota, Narayana R., 1996. The equity premium: It's still a puzzle. J. Econ. Literature 42–71.

Kocherlakota, Narayana, Pistaferri, Luigi, 2009. Asset pricing implications of pareto optimality with private information. J. Polit. Econ. 117 (3), 555–590.

Mankiw, N. Gregory, 1986. The equity premium and the concentration of aggregate shocks. J. Financ. Econ. 17 (1), 211–219.

Mankiw, N. Gregory, Zeldes, Stephen P., 1991. The consumption of stockholders and nonstockholders. J. Financ. Econ. 29 (1), 97–112.

Martin, Gervais, Klein, Paul, 2010. Measuring consumption smoothing in CEX data. J. Monetary Econ. 57 (8), 988–999.

Mehra, Rajnish, Prescott, Edward C., 1985. The equity premium: A puzzle. J. Monetary Econ. 15 (2), 145–161.

Rietz, Thomas A., 1988. The equity risk premium: A solution. J. Monetary Econ. 22 (1), 117–131.

Toda, Alexis Akira, Walsh, Kieran, 2015. The double power law in consumption and implications for testing Euler equations. J. Polit. Econ. 123 (5), 1177–1200.

Toda, Alexis Akira, Walsh, Kieran James, 2017. Fat tails and spurious estimation of consumption-based asset pricing models. J. Appl. Econometrics 32 (6), 1156–1177.

Vissing-Jørgensen, Annette, 2002. Limited asset market participation and the elasticity of intertemporal substitution. J. Polit. Econ. 110 (4), 825–853.