INFLAÇÃO ÓPTIMA

Isabel Correia **
*Pedro Teles***

Como deve ser conduzida a política monetária no longo prazo? Literatura recente mostra que a regra de Friedman é óptima, o que significa que a taxa de juro nominal deveria em média ser igual a zero, pelo que os preços deveriam baixar ao longo do tempo, dado a taxa de juro real ser positiva. Seguindo a regra de Friedman, o estado abstém-se de tributar a moeda, o que é desejável apesar da necessidade de recorrer a impostos distorcionários para financiar as despesas públicas.

1. INTRODUÇÃO

Como deve ser conduzida a política monetária no longo prazo, e também no curto prazo, em resposta a flutuações na economia? O estudo da política monetária desejável requer a identificação prévia dos efeitos de longo e de curto prazos da moeda. Só depois é possível determinar qual é a melhor estratégia de política, aquela que proporciona uma melhor afectação de recursos na economia. Para este fim, será também necessário utilizar modelos que reproduzam os factos relevantes e em que se possa colocar a questão de optimalidade.

Os efeitos de longo prazo da política monetária são bem conhecidos. Também estão disponíveis modelos adequados para avaliar esses efeitos. Pelo contrário, o estudo dos efeitos de curto prazo da moeda gera ainda alguma controvérsia, tanto na modelização teórica, como na própria identificação dos factos. As respostas à questão de como deve ser conduzida a política monetária de curto prazo estão portanto longe de ser definitivas. Em parte por esta razão, neste texto concentramo-nos em apresentar os resultados conhecidos sobre a política monetária óptima de longo prazo.

Na identificação dos efeitos de longo prazo da moeda, a teoria quantitativa é consensual. Economias com taxas de crescimento da moeda mais altas são economias com taxas de inflação mais altas e com taxas de juro nominais mais altas, também. Os efeitos sobre as taxas de juro reais e sobre as taxas de crescimento não têm significado (ver figuras 1 e 2[(1)]). Mesmo não havendo efeitos sobre o crescimento, há efeitos significativos da inflação na afectação de recursos, e por isso se justifica colocar a questão de optimalidade da política monetária de longo prazo. Sendo a política de longo prazo, as decisões incidem sobre as taxas médias de crescimento da moeda e dos preços, e sobre as médias seculares das taxas de juro nominais. As distorções provocadas por uma alta taxa média de inflação, ou por altas taxas de juro nominais, são como as de qualquer imposto. A alta inflação tributa as transacções que usam moeda, tornando o consumo e o investimento mais caros e desviando recursos para lazer ou para processos alternativos de realizar transacções. Como outro imposto qualquer, uma inflação média alta também permite ao Estado arrecadar mais receitas, por poder financiar défices com emissão de moeda, em vez de pagar

* As opiniões expressas no artigo são da inteira responsabilidade dos autores e não coincidem necessariamente com a posição do Banco de Portugal.

** Departamento de Estudos Económicos.

(1) Lucas (1996).

Figura 1

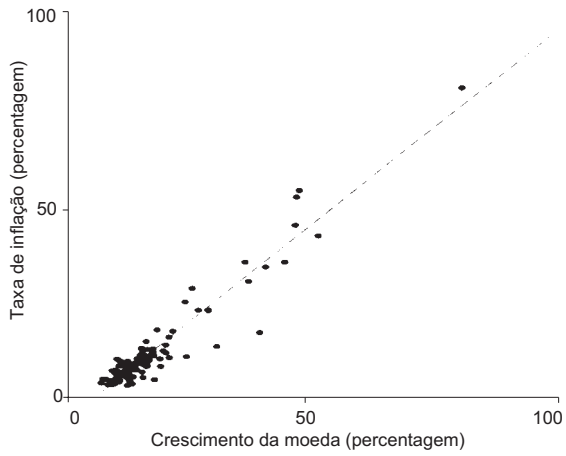
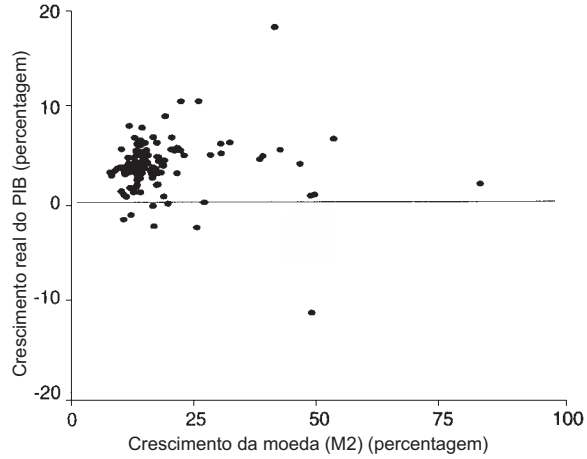


Figura 2



taxas de juro altas sobre dívida pública. O objectivo da política monetária de longo prazo é minorar o efeito das distorções provocadas pelo imposto inflação, tendo em atenção que se esse imposto for reduzido, outros impostos, também distorcionários, terão que ser aumentados de forma a financiar as despesas públicas necessárias. O conjunto de questões a que responderemos inclui: qual é a inflação média óptima, quando o Estado necessita de cobrar impostos distorcionários para financiar as despesas públicas necessárias? Qual deve ser a taxa média de crescimento da massa monetária? Qual é a taxa de juro nominal em obrigações de maturidades longas, que resulta da política óptima de longo prazo?

Milton Friedman propôs em 1969, em *The Optimum Quantity of Money*, uma regra de política monetária que pudesse dar origem a taxas de juro nominais tão baixas quanto possível: "a regra para a quantidade óptima da moeda é atingida por uma taxa de inflação que torne a taxa de juro nominal igual a zero". Os argumentos defendidos por Friedman são argumentos simples de optimalidade de Pareto, válidos apenas se fosse possível tributar sem provocar distorções. Um bem que tem um custo de produção zero, e de facto a moeda tem custos marginais de produção muito baixos, deve ter um preço também igual a zero. Como a taxa de juro nominal é o preço de deter moeda, por ser aquilo que os agentes privados deixam de receber por decidirem deter esse activo de maior liquidez, então a taxa de juro nominal de longo prazo deverá ser, segundo Friedman, igual a zero. Esta regra para a taxa de juro nominal significa que os preços deverão baixar em média a uma taxa igual à taxa de juro real de longo prazo: A quantidade de moeda deverá baixar a uma taxa consistente com a deflação necessária.

A crítica predominante à regra de Friedman é atribuída a Phelps (1973) que usou os princípios de tributação óptima de Ramsey (1927): na ausência de tributação não distorcionária, o problema de tributação óptima é financiar uma sequência exógena de despesas públicas da forma menos distorcionária possível. Neste contexto, a distorção marginal causada por uma unidade de receita proveniente de um imposto, deveria ser igual para todos os impostos. A implicação pareceria ser que também a moeda deveria ser tributada, como outro bem qualquer, e por isso o preço da moeda deveria ser superior ao seu custo de produção. A taxa de juro nominal de longo prazo deveria, portanto, ser maior que zero.

Desenvolvimentos recentes na teoria monetária de equilíbrio geral vieram questionar a intuição de Phelps e recuperar a bondade da regra de Friedman. Apesar da necessidade de recorrer a impostos distorcionários, a moeda não deve ser tributada. Este resultado é verdadeiro num ambiente económico em que a moeda é necessária para realizar transacções, explicitamente através de uma função de transacções em que a moeda pode ser substituí-

da por outros factores produtivos[2]. Porque neste ambiente, a moeda é um bem intermédio, aparentemente poder-se-ia invocar os resultados de tributação óptima de bens intermédios de Diamond e Mirrlees (1971), segundo os quais, em determinadas condições, esses bens não devem ser tributados. Acontece que as condições do teorema de Diamond e Mirrlees, em particular a condição de linearidade da estrutura de produção, não se verificam necessariamente nos modelos monetários. Por exemplo, se pensarmos, como é razoável, que a tecnologia de transacções propostas por Baumol (1952) e Tobin (1956) é uma boa descrição do processo de transacções, então a estrutura produtiva deixa de ser de rendimentos constantes à escala, pelo que já seria desejável tributar os bens intermédios.

Num ambiente mais próximo daquele usado por Phelps (1973), em que a moeda é usada para fornecer serviços de liquidez, modelizados como um bem final, Correia e Teles (1999) derivaram regras de tributação óptimas e concluíram, também nesse contexto, que a regra de Friedman é a regra geral de (não) tributação óptima da moeda. Dessa forma mostraram que a intuição de Phelps não pode ser aplicada à moeda. A intuição de Phelps não pode ser aplicada porque a moeda é um bem de custo zero, tributado através de um imposto específico, a taxa de juro nominal. Ora os resultados de tributação óptima de Ramsey (1927), ou de Diamond e Mirrlees (1971), referem-se a taxas de impostos *ad-valorem* sobre bens com um custo positivo de produção. Acontece que o resultado geral de que essas taxas de imposto devem ser positivas, não implica que o imposto específico deva ser positivo, quando o custo de produzir o bem se aproxima de zero. De facto, em termos gerais, esse imposto aproxima-se também de zero.

(2) Kimbrough (1986), Faig (1986, 1988), Guidotti e Végh (1983), Chari, Christiano e Kehoe (1983) demostraram o resultado, impondo condições restritivas. Correia e Teles (1996) generalizaram essas condições.

2. A INFLAÇÃO ÓPTIMA

Nesta secção, descrevemos em pormenor o resultado de tributação óptima da moeda em Correia e Teles (1996). Para responderem à questão de qual deve ser a inflação óptima de longo prazo, quando todos os impostos são distorcionários, Correia e Teles (1996) usam um modelo monetário em que a moeda é utilizada para transacções de tal forma que o tempo gasto em transacções é uma função do volume de transacções e do *stock* de moeda. Nesta forma de modelizar a moeda, a moeda é um bem intermédio necessário para a produção de transacções. Uma possível explicação da função de transacções, e a única com fundamentação microeconómica, é a tecnologia de transacções proposta por Baumol (1952) e Tobin (1956), segundo a qual o tempo gasto em transacções é uma função do rácio do volume de transacções por unidade de moeda, o número de visitas ao banco. Esta função de transações é homogénea de grau zero.

No modelo, há um número grande de famílias que têm uma dotação de tempo que podem usar para lazer, para a produção de um bem agregado, para produção de transacções, ou para a produção da própria moeda. As transacções têm um custo que é medido em termos de tempo dedicado a essa actividade. A moeda pode reduzir esse custo. É esta fricção que permite que a moeda tenha valor. As famílias têm preferências sobre bens de consumo e lazer. Em cada período há mercados de bens e trabalho e mercados de activos, moeda e obrigações nominais. Um governo benevolente escolhe a combinação óptima de impostos sobre o rendimento e do imposto inflação que financiam uma sequência exógena de despesas públicas.

Neste ambiente económico em que a moeda é um bem intermédio, Correia e Teles (1996) concluem que, quando a moeda tem um custo de produção negligenciável, é desejável que o governo se abstenha de tributar a moeda, qualquer que seja o grau de homogeneidade da função de transacções. Se pelo contrário a moeda requeresse custos significativos de produção, então já seria óptimo tributar a moeda, e essa taxa dependeria do grau de homogeneidade da função de transacções.

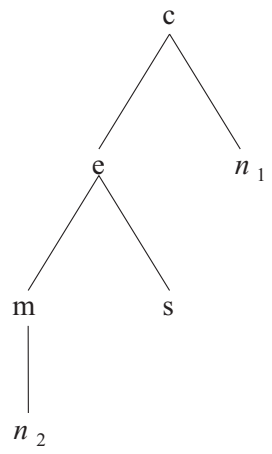
O resultado de que bens intermédios não devem ser tributados num ambiente de segundo óp-

timo, quando a tecnologia é de rendimentos constantes à escala, é bem conhecido desde o trabalho de Diamond e Mirrlees (1971). Nesse trabalho é demonstrado que eficiência na produção é uma característica da solução de segundo óptimo quando impostos sobre o consumo estão disponíveis. Como corolário desse resultado, bens intermédios não devem ser tributados. Eficiência na produção significa que o trabalho é afectado optimamente, como no primeiro óptimo, entre diferentes utilizações. Isso significa que a produtividade marginal do trabalho usado na produção de um determinado bem é igual ao produto da produtividade marginal do bem intermédio na produção desse bem e da produtividade marginal do trabalho na produção do bem intermédio. No modelo monetário em que há um bem agregado e não há capital, o imposto sobre o consumo proposto por Diamond e Mirrlees (1971) é equivalente a um imposto único sobre o trabalho, sem que os bens intermédios sejam tributados. Estas regras de imposto óptimo são regras sobre valores de impostos *ad-valorem*.

As regras de tributação de Diamond e Mirrlees (1971) não se aplicam directamente à economia monetária por duas razões. Porque a estrutura de produção é específica e porque há restrições naturais aos impostos que podem ser cobrados. Os aspectos distintivos da estrutura de produção no modelo monetário são, primeiro, que o bem de consumo requer trabalho e transacções de acordo com uma estrutura de produção Leontief, e, segundo, que funções de transacções interessantes, como a Baumol-Tobin, não são de rendimentos constantes à escala. O outro aspecto distintivo é que o tempo dedicado à produção de transacções não pode ser tributado, visto a actividade de transacções não passar pelo mercado.

No contexto do modelo monetário, a eficiência na produção é atingida quando a moeda e o tempo dedicado à sua produção não são tributados e apenas o tempo dedicado à produção do bem é tributado. Se a função de transacções for de rendimentos constantes à escala, é desejável atingir eficiência na produção, e portanto a moeda não deve ser tributada. No entanto se a função de transacções não for de rendimentos constantes à escala, como é o caso da função Baumol-Tobin, então já seria óptimo distorcer a produção e o imposto *ad-valorem* óptimo sobre a moeda já não seria zero. Acontece

Figura 3
ESTRUTURA PRODUTIVA NUMA ECONOMIA REAL EQUIVALENTE



que um imposto *ad-valorem* positivo corresponde a um imposto unitário igual a zero, quando o custo de produção da moeda tende para zero. Como o imposto de inflação é um imposto unitário o resultado da optimalidade da regra de Friedman tem a sua explicação, em última análise, na característica de bem livre da moeda.

De forma a compreender profundamente estes resultados da tributação óptima da moeda, é conveniente construir economias reais fictícias, mas equivalentes ao modelo monetário, onde são determinadas as regras óptimas de tributação de bens intermédios e de bens livres.

A economia real fictícia equivalente à economia monetária está representada na Figura 3. Nesta economia, os agentes têm preferências sobre consumo, *c, e lazer h. c* é produzido usando transacções, *e*, e trabalho n_1, de acordo com uma função de produção Leontief, $c=\min(e,n_1)$. A produção de *e* requer tempo, *s*, e *m*. O bem intermédio *m* é produzido com trabalho n_2, a uma taxa marginal constante $(m=\alpha n_2)$. O total de tempo na economia é normalizado para uma unidade. A estrutura de tributação é que *c*, n_1,n_2 e *m* podem ser tributados, mas *e* e *s* não o podem ser. Estas restrições sobre a capacidade tributária são restrições naturais no modelo monetário equivalente porque as transacções não passam pelo mercado.

Assumindo que a função $s=l(e,m)$ é homogénea de grau *k*, então a solução de tributação óptima é caracterizada pelas seguintes taxas de tributação *ad-valorem* da moeda

$$\tau_m = 0, \quad \text{quando } k = 1$$
$$\tau_m > 0 \quad \text{quando } k < 1$$
$$\tau_m < 0 \quad \text{quando } k > 1$$

quando a taxa de imposto sobre o trabalho usado na produção de moeda é zero, $\tau_2 = 0$.

Neste caso só é óptimo manter eficiência na produção de transacções *e*, quando a função de produção de *e* é de rendimentos constantes à escala. Quando há lucros, ou seja, quando a função transacções não é de rendimentos constantes à escala, o efeito dos impostos nos lucros, explica os desvios da eficiência na produção na solução de segundo óptimo. Quando $k \neq 1$, a possibilidade de lucros diferentes de zero, e a ausência de um imposto sobre estes lucros, justifica regras de tributação óptimas que induzam uma redução nos lucros. A redução nos lucros, mesmo negativos, é equivalente a um imposto *lump-sum*. Por isso, a solução de segundo óptimo permite uma distorção na produção, através da tributação de bens intermédios, de forma a reduzir os lucros implícitos na produção de transacções.

A razão pela qual a eficiência na produção de *e* é obtida quando τ_2 e τ_m são iguais a zero é o facto de τ_s ser por natureza igual a zero. Então, eliminando a tributação do trabalho utilizado na produção de *m*, e do próprio *m*, consegue-se manter a eficiência neste ramo da produção de transacções.

Visto transacções e horas de trabalho serem utilizadas em proporções fixas na produção do bem de consumo, a produção não é distorcida pela tributação de n_1. Por esta razão a solução de Ramsey, mesmo com as restrições particulares do sistema fiscal descritas (*s* e *e* não podem ser tributados) é um segundo óptimo, e não um terceiro ou quarto óptimo. Se se impusesse τ_2 igual a τ_1, eficiência na produção implicaria que τ_m fosse negativo. Assim podemos dizer que o resultado obtido para tecnologias de rendimentos constantes à escala que *m* não deve ser tributado, garante a eficiência na produção mas, devido às restrições dos instrumentos fiscais impostas, não é uma extensão natural do resultado de Diamond e Mirrlees. Neste caso o bem intermédio não é tributado mas o rendimento do trabalho é tributado a taxas muito diferenciadas dependendo do sector onde tem origem:

O rendimento do trabalho na produção de moeda e na produção de transacções não é tributado e o rendimento do trabalho na produção do bem de consumo é tributado a uma taxa positiva.

Quando *m* é um bem livre, se a taxa de juro nominal for zero, e portanto a moeda estiver a ser utilizada plenamente ($l_m = 0$), o efeito marginal de *m* sobre os lucros é zero. Apesar do facto de, para funções de transacções homogéneas de grau $k \neq 1$, o nível de lucros implícitos ser diferente de zero, e de em geral existir um efeito marginal de *m* sobre os lucros, no ponto de saciedade em moeda real, ponto em que o bem livre tem produtividade marginal zero, este efeito é nulo e por isso o ponto de saciedade define a quantidade óptima de moeda. Este resultado pode ser interpretado como o resultado limite do imposto unitário óptimo que incide sobre um bem intermédio que utiliza recursos, quando o custo de produzir o bem se torna arbitrariamente pequeno. A intuição é que o imposto unitário equivalente a um imposto finito *ad-valorem* sobre um bem com custo de produção arbitrariamente baixo, é arbitrariamente baixo.

Em qualquer caso os custos variáveis de produção de moeda serem zero é a hipótese essencial para a optimalidade da regra de Friedman. Tomamos esta hipótese como certa apesar de ser evidente que existem custos fixos importantes na criação monetária. Assim a moeda como um *input* primário livre, e não como um bem intermédio, é a hipótese quantitativamente razoável relevante assim como a justificação teórica fundamental da robustez da optimalidade da regra de Friedman.

3. CONCLUSÕES

A inflação média, de longo prazo, tem efeitos reais sobre o nível da actividade económica. Para minorar estes efeitos, a literatura sobre regras de política monetária de longo prazo recomenda uma política consistente com taxas de juro nominais perto de zero, o que corresponde a deflação, de acordo com a regra de Friedman de 1969. Este resultado é surpreendente por ser válido mesmo quando se entra em consideração com a necessidade que o Estado tem de recorrer a impostos distorcionários para financiar os gastos públicos (Correia e Teles, 1996 e 1999). A intuição básica do resultado é que a taxa de juro nominal é uma taxa de

imposto unitária sobre um bem (moeda) com um custo de produção muito baixo e por isso, mesmo que em termos proporcionais seja óptimo tributar a moeda a uma taxa elevada, o imposto específico equivalente é muito baixo. Uma vez determinada a política óptima, põe-se a questão quantitativa de quais são os ganhos de bem-estar de reduzir a taxa de juro nominal para taxas perto do zero. Correia e Teles (1994) calculam que o ganho de reduzir a taxa de juro nominal de 5 por cento para a regra de Friedman, estará perto de 1 por cento do PIB(3).

Este resultado limite, de que não é desejável tributar a moeda, pode, no entanto, ser corrigido de acordo com considerações várias, como sejam a tributação da economia subterrânea, custos altos de administração do sistema fiscal, ou custos de variação de preços. Dado que a economia subterrânea é precisamente um sector que não pode ser tributado através do sistema fiscal, justifica-se por razões de eficiência e equidade, que o imposto de inflação seja usado para esse fim. Em termos quantitativos, os valores óptimos da inflação de longo prazo são ligeiramente superiores a zero(4). Custos altos de cobrança de impostos sobre o consumo ou rendimento poderão também justificar um desvio da regra de Friedman, que é um desvio menor, na ordem de grandeza de um ponto percentual para a taxa de juro nominal(5). Custos de variações de preços podem também justificar desvios da regra de Friedman no sentido do objectivo de estabilidade de preços.

(3) Ver também Lucas (1994).
(4) Nicolini (1998).
(5) De Fiore (1998).

REFERÊNCIAS

Baumol, W.J., 1952, The Transactions Demand for Cash: An Inventory Theoretic Approach, *Quarterly Journal of Economics* 66, 545-556.

Chari, V.V., L. J. Christiano e P.J. Kehoe, 1996, Optimality of the Friedman Rule in Economies with Distorting Taxes, *Journal of Monetary Economics* 37, 203-223.

Correia, I. e P. Teles, 1994, Money as an Intermediate Good and the Welfare Cost of the inflation Tax, Banco de Portugal *Working Paper* nº 10 de 1994.

Correia, I. e P. Teles, 1996, Is the Friedman Rule Optimal when money is an Intermediate Good?, *Journal of Monetary Economics*, 38, 223-244.

Correia, I. e P. Teles 1999, The Optimal Inflation Tax, *Review of Economic Dynamics* 2, 325-246.

De Fiore, F., 1998, The Optimal Inflation Tax with Costs of Collecting Taxes, mimeo, European University Institute.

Diamond, P.A. e J.A. Mirrlees, 1971 Optimal Taxation and Public Production, *American Economic Review* 63, 8-27, 261-268.

Faig, M., 1986, Optimal Taxation of Money Balances, Ph. D. Dissertation (Stanford University, Stanford, CA).

Faig, M., 1988, Characterization of the Optimal tax on Money when it Functions as a Medium of Exchange, *Journal of Monetary Economics* 22, 137-148.

Friedman, M., 1969, The Optimum Quantity of Money, em M. Friedman, ed., The Optimum Quantity of Money and other Essays (Aldine Chicago, II.).

Guidotti, P.E. e C.A. Végh, 1993, The Optimal Inflation Tax when Money Reduces Transactions Costs, *Journal of Monetary Economics* 31, 189-205.

Kimbrough, K.P., 1986, The Optimum Quantity of Money Rule in the Theory of Public Finance, *Journal of Monetary Economics* 18, 277-284.

Lucas, Robert E., Jr., 1994, On the Welfare Cost of Inflation, mimeo, The University of Chicago.

Lucas, R. 1996, Monetary Neutrality, *Journal of Political Economy*, 4, 104, 661-682.

Nicolini, J. P., 1998, Tax Evasion and the Optimal Inflation Tax, *Journal of Development Economics*, 55 (1) 215-232.

Phelps, E. S., 1973, Inflation in the Theory of Public Finance, *Swedish Journal of Economics* 75, 37-54.

Ramsey, F.P., 1927, A Contribution to the Theory of Taxation, *Economic Journal* 37, 47-61.

Tobin, J., 1956, The Interest Elasticity of the Transaction Demand for Cash, *Review of Economics and Statistics* 38, 241-247.