UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

CENTER FOR INTERNATIONAL ECONOMICS

Working Paper Series

Working Paper No. 2013-04

Comment on Ameriks, Caplin, Leahy & Tyler (2007): Measuring Self-Control Problems

Volker Seiler

April 2013

CIE CENTER FOR INTERNATIONAL ECONOMICS

Comment on Ameriks, Caplin, Leahy & Tyler (2007): Measuring Self-Control Problems

Volker Seiler[*]

22.04.2013

Abstract

Ameriks, Caplin, Leahy & Tyler (2007) study how lack of self-control may lead to overconsumption and low wealth accumulation. For their analysis, the authors calculate difference scores from their survey questions. We review their study and highlight potential issues in the use of difference scores such as lack of reliability, validity, spurious correlation and variance restriction. The concerns we raise put into question their findings and the conclusions they draw.

Key Words: Self-control, difference scores, reliability, validity, spurious correlation, variance restriction

JEL classification codes: A12, B41, C18, C42

1

1 Introduction

Models of self-control generally argue that individuals aim to undertake some ideal action but are tempted to deviate from this ideal. In their 2007 contribution "Measuring Self-Control Problems", Ameriks, Caplin, Leahy & Tyler (2007) propose a survey instrument for measuring such self-control problems. The authors empirically investigate the influence self-control exhibits on wealth accumulation using questionnaires sent to participants of the Teachers Insurance and Annuity Association - College Retirement Equities Fund (TIAA-CREF) and find that, contrary to the widely held belief that self-control problems lead to overconsumption and thus low wealth, some study participants are able to accumulate high levels of wealth due to underconsumption. Furthermore, the authors find that self-control problems seem to be higher for younger individuals than for older persons.

Given the fact that Ameriks, Caplin, Leahy & Tyler (2007) measure self-control as a difference score between an individual's expectation and ideal, several questions arise as to the reliability and validity of their measures as well as issues concerning spurious correlations with other measures and variance restriction (Peter, Churchill & Brown (1993)). For other disciplines, these issues have already been addressed. For example, Johns (1981) investigates measures based on difference scores in studies of organizational behaviour and Peter, Churchill & Brown (1993) analyze the usage and shortcomings of difference scores in consumer research. However, as the usage of difference scores seems new in the field of economic research, we will address these questions in the following sections.

2 Review of the study by Ameriks, Caplin, Leahy & Tyler (2007)

In their attempt to measure self-control problems, Ameriks, Caplin, Leahy & Tyler (2007) ask study participants how they would allocate a prize between two periods, namely ten certificates for a dream restaurant night. Each certificate covers the complete cost for food and drinks as well as being seated at the best table in a restaurant of the participant's choice. The certificates are valid for a period of two years and participants have to decide how to allocate the certificates in each year. More specifically, their survey instrument consists of four questions, namely (see Ameriks, Caplin, Leahy & Tyler (2007, p. 967))

"(a) From your current perspective, how many of the ten certificates would you ideally like to use in year 1 as opposed to year 2?

(b) Some people might be tempted to depart from their ideal allocation in (a). Which of the following best describes you (please mark only one):

- I would be strongly/somewhat tempted to keep more certificates for use in the second year than would be ideal;

- I would have no temptation in either direction (skip to d);

- I would be somewhat/strongly tempted to use more certificates in the first year than would be ideal.

(c) If you were to give in to your temptation, how many certificates do you think you would use in year 1 as opposed to year 2?

(d) Based on your most accurate forecast of how you think you would actually behave, how many of the nights would you end up using in year 1 as opposed to year 2?"

The authors use the difference between expected consumption in the first period (d) and ideal consumption (a) as a measure of self-control problems and label this measure expected-ideal gap (EI gap), where a positive value indicates the problem of overconsumption whereas a negative value indicates underconsumption. Furthermore, the authors calculate the temptation-ideal gap (TI gap) as the difference between the most tempting choice (c) and the ideal choice (a). Additionally, the authors gave participants the opportunity to impose binding restraints on the use of the certificates for year 1 and/or year 2. Based on these restrictions, they calculate the revealed preference gap (RP gap) as the difference between the expected consumption (d) and the restricted set.

3 Issues Concerning Difference Scores

3.1 Reliability of Difference Scores

The first issue concerning the usage of difference scores refers to their reliability. Reliability is the degree to which a measure is free of random measurement error (see Kuder & Richardson (1937), Carmines & Zeller (1979), Peter (1979), and Nunnally & Bernstein (1994)). However, difference scores generally exhibit lower reliability than the measures from which they are constructed (see Lord (1958), Johns (1981), and Peter, Churchill & Brown (1993)). This is due to the fact that the reliability of difference scores does not only depend on the reliabilities of their components but also on the correlation between these components:

$$r_D = \frac{\sigma_1^2 r_{11} + \sigma_2^2 r_{22} - 2r_{12}\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2r_{12}\sigma_1\sigma_2} \tag{1}$$

where r_{11} and r_{22} represent the reliabilities of the first and second component, σ_1^2 and σ_2^2 the variances of the components, and r_{12} the correlation coefficient of the components, respectively (see Johns (1981), and Peter, Churchill & Brown (1993)).

Only in case of zero correlation between the components, the reliability of the difference score is equal to the average reliability of its components (Johns (1981)). In case of positive correlation, the reliability of the difference score decreases with decreasing reliability of either one or both of its parts, as is illustrated in figure 1. One can see, for example, that the reliability of a difference score where its components have an average reliability of 0.70 and a correlation of 0.50 is 0.40 as the correlation between the components dilutes the reliability of the difference score. In case of a negative correlation between the components, the reliability of the difference score will exceed that of its parts (Johns (1981)).
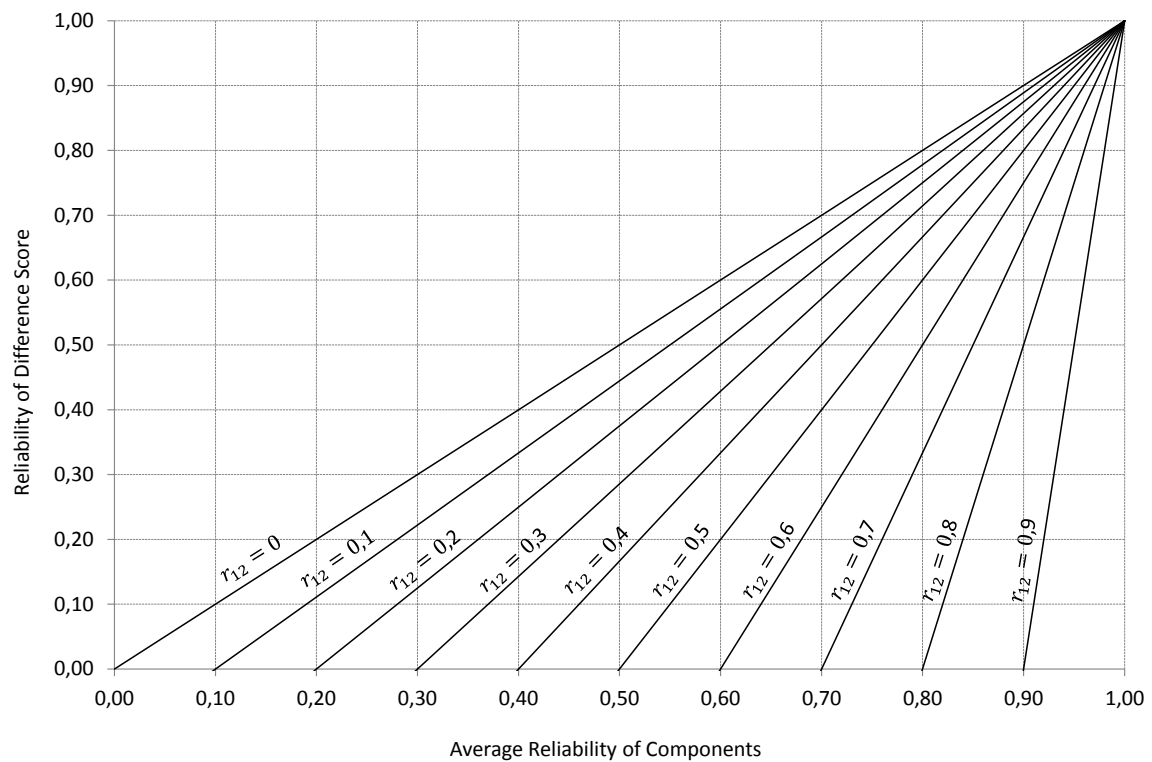

Figure 1: Reliability of Difference Scores as a Function of the Reliability of the Components and the Correlation between them

Source: Peter, Churchill & Brown (1993, p. 659).

The assumption of positive correlation between the components used to form a

5

difference score seems not unlikely, given the fact that scores on both parts are collected from the same respondents at the same time, the perspective the individual is asked to take being the only difference (see Peter, Churchill & Brown (1993)). Hence, the components of the difference score will contain common variance due to the respondent's general mood and response style (Johns (1981)). In addition to that, common variance stems from respondents using their answer to the first component of the difference score to anchor their answer to the second component (Johns (1981)).

As single items are used to measure the components, not only do difference scores lack reliability, but their components as well, as the measures are generally not grounded in theory. As Johns (1981, p. 449) puts it: "Many articles which report difference scores have obtained them from measuring instruments which were constructed ad hoc for use in the particular study in question. Since the reliabilities of the component scores of these instruments are virtually never reported, and since the instruments have no published history of reliability, one is left to deduce 'face reliability' from 'face validity'."

Ameriks, Caplin, Leahy & Tyler (2007) use single item measures as components for the calculation of the expected-ideal (EI) gap, the temptation-ideal (TI) gap and the revealed preference (RP) gap. Hence, no assessment of the reliability of these measures is possible ex post as would be for multi-item scales, e.g. the calculation of item correlations (Bohrnstedt (1969), Spector (1992), Edwards & Bagozzi (2000), and DeVellis (2003)), Cronbach's coefficient alpha (Cronbach (1951), Carmines & Zeller (1979), Peter (1979), Spector (1992), Cortina (1993), Nunnally & Bernstein (1994), DeVellis (2003), and Netemeyer, Bearden & Sharma (2003)) or Jöreskog's rho (Werts, Linn & Jöreskog (1974), and Fornell & Larcker (1981)). Furthermore, Ameriks, Caplin, Leahy & Tyler (2007) use two items from the Revised NEO Personality Inventory (NEO-PIR) to measure conscien-

tiousness (see Costa, McCrae & Dye (1991)), namely "'[s]ometimes I am not as dependable or reliable as I should be'; and 'I never seem able to get organized.'"[1] Instead of using the two items as a multi-item scale and assess its reliability, they use each item separately to measure conscientiousness. However, Johns (1981, p. 459) postulates that "potential components should be carefully justified by theory and rigorously measured, analyzed, and reported, whether or not differences are ultimately employed. Potential components should consist of internally consistent multiple-item scales, not heterogeneous collections of factorially indeterminant items." Furthermore, Costa, McCrae & Dye (1991) originally regarded conscientiousness as a construct that comprises six facets, namely competence, order, dutifulness, achievement striving, self-discipline and deliberation, each of which is measured with a multi-item scale. Hence, it is unclear which facets the two items used by Ameriks, Caplin, Leahy & Tyler (2007) cover.

3.2 Validity of Difference Scores

Whereas reliability indicates the absence of random error, validity indicates the absence of random and systematic error (see Carmines & Zeller (1979)). An instrument that is deemed to be valid is always reliable; however, the reverse argument does not hold true. Hence, reliability is a necessary but not a sufficient condition for validity (see Peter (1979), Peter (1981), Nunnally & Bernstein (1994), and Netemeyer, Bearden & Sharma (2003)).

When assessing the validity of measures, one aspect to consider is discriminant validity, i.e., whether the correlation of measures between theoretically distinct constructs is not too high (Peter, Churchill & Brown (1993), and Reichardt & Coleman (1995)). However, the low reliability of difference scores dilutes the correlation between different constructs. Hence, researchers might wrongly conclude

---

[1] See Wiggins & Pincus (1994, p. 82).

that an adequate amount of discriminant validity is given, as the difference score shows only low correlations with other constructs. This problem is the more severe, the higher the correlation between the two components of the difference score (Peter, Churchill & Brown (1993)).

Ameriks, Caplin, Leahy & Tyler (2007) report a correlation of the expected-ideal (EI) gap and the temptation-ideal (TI) gap of 0.4. Thus, one might conclude that the two constructs indeed measure different things. However, as the authors do not report the reliabilities of their difference scores and the correlation of the items used to form the difference scores, no definitive assessment of validity is possible. It might be the case that the correlation coefficient of 0.4 is not an indication of discriminant validity but due to the correlations of the components and/or the unreliability of the difference scores calculated thereof.

Furthermore, another issue that arises concerning the use of difference scores is the fact that a difference score measure will show a high correlation with either one or both of its components. Theoretically, a difference score measures a distinct construct than its components. However, if the difference score exhibits high correlations with the components it consists of, it fails to show discriminant validity. In case of extremely high correlations of the difference score with its parts it seems thus highly questionable whether the difference score measures a construct unique from its components (Peter, Churchill & Brown (1993)). To illustrate this point, we conducted a monte carlo simulation (mc) and simulated three independent and identically distributed random variables: $a, b, c \sim IID\ N(0,1)$. We generated 1.000 observations for each variable and repeated this procedure 100 times. We used these three random variables to calculate two difference score measures d_1 and d_2 where $d_1 = a - b$ and $d_2 = c - b$, respectively. Table 1 shows the average correlation coefficients of our 100 simulations. The difference scores show correlations of $|0.71|$ with their measures. Hence, it seems doubtful whether the difference scores

really measure something different from their components. As Ameriks, Caplin, Leahy & Tyler (2007) do not report the correlations of the components and the difference score measures, no evaluation of this issue is possible.

3.3 Spurious Correlation

Given the high correlations between difference scores and the measures from which they are calculated, correlations between difference scores and other constructs are very likely to be spurious. This stems from the correlation between some construct and a difference score reflecting nothing else but the correlation of this construct with the components from which the difference score is calculated (Wall & Payne (1973), and Peter, Churchill & Brown (1993)). Thus, Peter, Churchill & Brown (1993, p. 660) emphasize that "the difference between two variables provides no additional information for predicting or explaining a criterion beyond that held in the components themselves." Hence, as difference scores tend to be correlated with one or both of its components, the chances to observe correlations with other constructs that are connected to the components are amplified (Johns (1981)). However, these correlations are nothing else but methodological artifacts (Johns (1981), and Peter, Churchill & Brown (1993)).

	a	b	c	d_1	d_2
a	1.00				
b	0.00	1.00			
c	0.00	0.00	1.00		
d_1	0.71	-0.71	0.00	1.00	
d_2	0.00	-0.71	0.71	0.50	1.00

Table 1: Correlation Matrix

Ameriks, Caplin, Leahy & Tyler (2007) report a correlation between the expected-ideal (EI) gap and the temptation-ideal (TI) gap of 0.4. However, both, EI gap and TI gap have one component in common, i.e., question (a) on the ideal choice. Hence, the observed value of a correlation of 0.4 between the difference scores is nothing else but spurious correlation. The same holds for the correlation between the expected-ideal (EI) gap and the revealed preference (RP) gap for which the authors report a correlation of 0.5. In this case, question (d) on expected consumption is a component of both, EI gap and RP gap. To illustrate this point, we refer to the correlation matrix of our monte carlo simulation: Although a, b and c are completely uncorrelated, the correlation coefficient of the difference scores d_1 and d_2 is 0.50, as both share one component, i.e., b.

Furthermore, in case the difference score and one or both of its components are used in the same analysis, the high multi-collinearity produces misleading results due to unstable parameter estimates (Peter, Churchill & Brown (1993)). In their regression analysis, Ameriks, Caplin, Leahy & Tyler (2007) include the EI gap and ideal choice (question (a)) which is one of the components of the EI gap both at the same time as explanatory variables to estimate their impact on net worth. They do not find a significant impact for the ideal consumption level (question (a)) but a negative impact of the EI gap which is significant at the 1 percent level of confidence. Accordingly, Ameriks, Caplin, Leahy & Tyler (2007, pp. 968-969) state that "[t]he regression identifies a clear relationship between self-control problems and wealth accumulation. Note that we also include the answer to question (a) on the ideal level of consumption and find it to have no explanatory power whatsoever." In light of the previous arguments, it is doubtful that these results really are what they seem to be. Rather, it is more likely that the parameter estimates on the EI gap and the ideal level of consumption are biased due to a non-negligible amount of multi-collinearity. However, Ameriks, Caplin, Leahy &

Tyler (2007) do not provide any assessment of potential multi-collinearity, e.g., a correlation matrix of the explanatory variables or the variance inflation factor (VIF) (see Mason & Perrault (1991), Cohen, Cohen, West & Aiken (2003), and Hair, Black, Babin, Anderson & Tatham (2006)).

Concerning the TI gap, the authors find a positive correlation with their wealth measures. If both, the EI gap and the TI gap are included in the analysis, only the EI gap exhibits a significant effect. From this finding, Ameriks, Caplin, Leahy & Tyler (2007, p. 971) conclude that "[t]he TI gap appears to work through the EI gap." However, the authors do not undertake any attempt to investigate this proposition, for example using hierarchical regressions or structural equation modelling. Moreover, this effect again might be due to problems of multi-collinearity, as both, the EI gap and the TI gap contain question (a) concerning the ideal choice.

3.4 Variance Restriction

Next to issues of reliability, validity, and spurious correlations, difference scores suffer from possible restriction of their variance (Peter, Churchill & Brown (1993)). This is the case if one of the components from which the difference score is calculated is consistently higher than the other component. If for one of the measures from which the difference score is calculated more is universally better, this problem of variance restriction is very likely to occur (see Wall & Payne (1973), and Peter, Churchill & Brown (1993)).

Ameriks, Caplin, Leahy & Tyler (2007, p. 968) report that "[t]he EI gap is typically small: 95 percent of responses are less than two in absolute value. Roughly two in every three respondents have EI gaps of zero, corresponding to their having no self-control problem according to our measure." This finding might be due to variance restriction.

4 Concluding Remarks and Recommendation

Ameriks, Caplin, Leahy & Tyler (2007) shed further light on how overconsumption might lead to low levels of wealth accumulation. While research in this field is important as it helps to understand why people tend to prefer consumption today over saving for retirement and designing policies to overcome this myopia, we raise some questions concerning the methodological approach taken by the authors. More precisely, we illustrate how the usage of difference scores might lead to questionable reliability and validity, issues concerning spurious correlation, and variance restriction. We agree with Peter, Churchill & Brown (1993, p. 661) who conclude that "difference scores should not be used simply because they are intuitively appealing or computationally convenient." Instead a single measure that directly captures the EI gap should be employed (see Wall & Payne (1973), Johns (1981), and Peter, Churchill & Brown (1993)). As Johns (1981, p. 454-455) puts it: "[T]he use of difference scores implies that they should do something 'more' or 'better' than other measures, especially their own components, which are already available, usually more reliable, and more likely to be grounded in theory. Difference scores may 'steal' variance from their components, carry this variance less reliably, and then do a poor job of explaining other variables. At best, this job may be poor because it lacks parsimony - one is required to measure two variables directly and calculate a third variable to do what a single measure might have done. At worst, the difference score is more weakly related to some other variable than one or both components and fails to supplement their singular effects. Of course, it might be argued that parsimony and efficiency are not the only criteria for scientific research, and that we may be interested in exploring 'how something works' rather than 'accounting for a lot of variance.' While this may be true, there is seldom theoretical reason for an a priori belief that 'things work' through subtraction rather than some other combination strategy."

Furthermore, in case the difference score and one or both of its components are used in the same analysis, the high multi-collinearity produces misleading results due to unstable parameter estimates. Hence, Peter, Churchill & Brown (1993, p. 662) conclude that "given that an arithmetic difference score adds no additional information for predicting a criterion beyond that held in the components themselves, there is at least some question as to the value of conceptualizing constructs as differences between two other constructs."

While the usage of questionnaires to gather data in order to measure abstract constructs as latent variables is common in areas such as psychology and marketing, the methodologies to analyze such data are a relative new approach in economic research. We believe that the economics profession can learn from the other fields. For example, in the study by Ameriks, Caplin, Leahy & Tyler (2007), the research question can be examined with the help of alternative techniques which do not require the calculation of difference scores such as hierarchical regression or structural equation modeling (SEM). Alternatively, it would be interesting to see experimental studies that analyze the questions put forward by the authors.

References

Ameriks, J., Caplin, A., Leahy, J. & Tyler, T. (2007), 'Measuring self-control problems', *American Economic Review* **97**(3), 966–972.

Bohrnstedt, G. W. (1969), 'A quick method for determining reliability and validity of multiple-item scales', *American Sociological Review* **34**(4), 542–548.

Carmines, E. G. & Zeller, R. A. (1979), *Reliability and Validity Assessment*, Quantiative Applications in the Social Sciences, Sage Publications, Newbury Park, London, New Delhi.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003), *Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd edn, Lawrence Earlbaum Associates, Mahwah (New Jersey), London.

Cortina, J. M. (1993), 'What is coefficient alpha? an examination of theory and applications', *Journal of Applied Psychology* **78**(1), 98–104.

Costa, P. T. J., McCrae, R. R. & Dye, D. A. (1991), 'Facet scales for agreeableness and conscientiousness: A revision of the neo personality inventory', *Personality and Individual Differences* **12**(9), 887–898.

Cronbach, L. J. (1951), 'Coefficient alpha and the internal structure of tests', *Psychometrika* **16**(3), 297–334.

DeVellis, R. (2003), *Scale Development: Theory and Applications*, Applied Social Research Methods Series, Sage Publications, Thousand Oaks, London, New Delhi.

Edwards, J. R. & Bagozzi, R. P. (2000), 'On the nature and direction of relationships between constructs and measures', *Psychological Methods* **5**(2), 155–174.

Fornell, C. & Larcker, D. F. (1981), 'Evaluating structural equation models with unobservable variables and measurement error', *Journal of Marketing Research* **18**(1), 39–50.

Hair, J. F. J., Black, W. C., Babin, B. J., Anderson, R. E. & Tatham, R. L. (2006), *Multivariate Data Analysis*, 6th edn, Pearson Prentice Hall, Upper Saddle River (New Jersey).

Johns, G. (1981), 'Difference score measures of organizational behavior variables: A ciritque', *Organizational Behavior and Human Performance* **27**(3), 443–463.

Kuder, G. & Richardson, M. (1937), 'The theory of the estimation of test reliability', *Psychometrika* **2**(3), 151–160.

Lord, F. M. (1958), 'The utilization of unreliable difference scores', *Journal of Educational Psychology* **49**(3), 150–152.

Mason, C. H. & Perrault, W. D. J. (1991), 'Collinearity, power, and interpretation of multiple regression analysis', *Journal of Marketing Research* **28**(3), 268–280.

Netemeyer, R. G., Bearden, W. O. & Sharma, S. (2003), *Scaling Procedures: Issues and Applications*, Sage Publications, Thousand Oaks, London, New Delhi.

Nunnally, J. C. & Bernstein, I. H. (1994), *Psychometric Theory*, 3rd edn, McGraw-Hill, New York, St. Louis, San Francisco, Auckland, Bogotá, Caracas, Lisbon, London, Madrid, Mexico City, Milan, Montreal, New Delhi, San Juan, Singapore, Sydney, Tokyo, Toronto.

Peter, J. P., Churchill, G. A. J. & Brown, T. J. (1993), 'Caution in the use of difference scores in consumer research', *Journal of Consumer Research* **19**(4), 655–662.

Peter, P. J. (1979), 'Reliability: A review of psychometric basics and recent marketing practices', *Journal of Marketing Research* **16**(1), 6–17.

Peter, P. J. (1981), 'Construct validity: A review of basic issues and marketing practices', *Journal of Marketing Research* **18**(2), 133–145.

Reichardt, C. S. & Coleman, S. (1995), 'The criteria for convergent and discriminant validity in a multitrait-multimethod matrix', *Multivariate Behavioral Research* **30**(4), 513–538.

Spector, P. E. (1992), *Summated Rating Scale Construction: An Introduction*, Quantitative Applications in the Social Sciences, Sage Publications, Newbury Park, London, New Delhi.

Wall, T. D. & Payne, R. (1973), 'Are deficiency scores deficient?', *Journal of Applied Psychology* **58**(3), 322–326.

Werts, C., Linn, R. & Jöreskog, K. (1974), 'Intraclass reliability estimates: Testing structural assumptions', *Educational and Psychological Measurement* **34**(1), 25–33.

Wiggins, J. S. & Pincus, A. L. (1994), Personality structure and the structure of personality disorders, *in* P. T. J. Costa & T. A. Widiger, eds, 'Personality Disorders and the Five-Factor Model of Personality', American Psychological Association, Washington DC, pp. 73–93.

Recent discussion papers

2013-04	Volker Seiler	Comment on Ameriks, Caplin, Leahy & Tyler (2007): Measuring Self-Control Problems
2013-03	B. Michael Gilroy Anastasia Heiman Mark Schopf	Gibt es eine optimale Frauenquote?
2013-02	Yuanhua Feng Chen Zhou	Forecasting financial market activity using a semiparametric fractionally integrated Log-ACD
2013-01	Berno Buechel Nils Röhl	Robust Equilibria in Location Games
2012-11	B. Michael Gilroy Mark Schopf Anastasia Semenova	Grundeinkommen und Arbeitsangebot: Die Perspektive Deutschlands
2012-10	N. Bilkic B. Carreras Painter T. Gries	Unsustainable Sovereign Debt - is the Euro Crisis only the Tip of the Iceberg?
2012-09	B. Michael Gilroy Heike Schreckenberg Volker Seiler	Water as an Asset Class
2012-08	B. Michael Gilroy Mark Schopf Anastasia Semenova	Basic Income and Labor Supply: The German Case
2012-07	Tim Krieger	Es geht nicht ohne internationalen Ordnungsrahmen: Zum Verhältnis von Markt und Staat im Angesicht der Finanzkrise *[forthcoming in: Amos international: Internationale Zeitschrift für christliche Sozialethik (2012), 3]*
2012-06	Thomas Gries Margarete Redlin	Trade Openness and Economic Growth: A Panel Causality Analysis
2012-05	Mark Schopf Hendrik Ritter	Reassessing the Green Paradox
2012-04	Yuanhua Feng David Hand Keming Yu	A Multivariate Random Walk Model with Slowly Changing Drift and Cross-correlation Applied to Finance
2012-03	Sarah Brockhoff Tim Krieger Daniel Meierriecks	Looking Back on Anger: Explaining the Social Origins of Left-Wing and Nationalist-Separatist Terrorism in Western Europe, 1970-2007
2012-02	Sonja Brangewitz Sarah Brockhoff	Stability of Coalitional Equilibria within Repeated Tax Competition
2012-01	Sarah Brockhoff Tim Krieger Daniel Meierriecks	Great Expectations and Hard Times - The (Nontrivial) Impact of Education on Domestic Terrorism