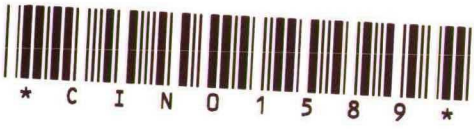
entER for omic Research

Discussion paper



TILBURG UNIVERSITY

Center
for
Economic Research

No. 9431

ON TESTS AND SIGNIFICANCE IN ECONOMETRICS

by Hugo Keuzenkamp and
Jan Magnus

March 1994

ISSN 0924-7815

ON TESTS AND SIGNIFICANCE IN ECONOMETRICS

Hugo A. Keuzenkamp
Department of Economics, Tilburg University, The Netherlands

Jan R. Magnus
London School of Economics, London, UK and
CentER for Economic Research, Tilburg University, The Netherlands

Abstract:
We discuss different aims of testing: theory testing, validity testing, simplification testing and decision making. Different testing methodologies may serve these aims. In particular, the approaches of Fisher and Neyman-Pearson are considered. We discuss the meaning of statistical significance. Significance tests in the *Journal of Econometrics* are evaluated. The paper concludes with a challenge to ascertain the impact of statistical testing on economic thought.

Classification code: B23, B40, C12

Key words:
Statistical tests; inference; significance level

Proofs: Send proofs to:

Hugo Keuzenkamp
Department of Economics
Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands

1 Introduction[1]

In a provocative paper, McCloskey (1985, p. 182) contends that 'no proposition about economic behaviour has yet been overturned by econometrics'. McCloskey is not a lonely skeptic. Many outsiders are doubtful of the value added of econometric testing (e.g. Hahn, 1992). But also many econometricians are increasingly worried about the credibility gap between econometric theory and applied economics (for example, Spanos, 1986, p. 660). Whether the skeptics are right or wrong, we must face the question: What is the significance of testing econometric hypotheses?

Testing hypotheses belongs to the basic pastimes of econometricians. It is a compulsory topic in any course in introductory statistics and econometrics. In such a course, students are made familiar with notions like Type I and Type II errors, significance level, and power. This is firmly in the tradition of statistical testing along the lines proposed by Jerzy Neyman and Egon Pearson (1928, 1933). However, econometric practice seems closer to the approach of Sir R.A. Fisher, although he is rarely mentioned (apart from references to the *F*-test). We will clarify the differences between both approaches below.

At first sight, the lessons of an introductory econometrics course seem most useful, if one judges the amount of papers in economic journals that contain statistical tests. A casual investigation of titles of papers shows that there is a lot of 'testing' in the literature. Less comforting is the amount of 'evidence' that is found. What proportion of the results of tests, or of the evidence, is regarded to be powerful by a significant part of the audience? If the value added of testing is low, some reflections on the merits of testing in econometrics are due. It must be admitted that it is hard (but perhaps not impossible) to find a convincing example of a meaningful economic proposition, that has been rejected (or definitively supported) by econometric tests. Many statistical hypotheses have been tested and rejected. But in how many cases did the result remain unchallenged by a respectable colleague, or how often was a statistical rejection rather than common sense responsible for exorcizing a

[1] We are grateful to Michael McAleer and Mark Steel for their helpful suggestions.

defective economic argument? If the value added of testing is low, some reflections on the merits of testing in econometrics are due.

In Section 2, we discuss aims of testing, relating them to popular views in the philosophy of science. In Section 3, some statistical methods of testing are discussed. Statistical significance is analyzed in Section 4, while testing in the Journal of Econometrics is the topic of Section 5. We conclude the paper with a challenge to the readers.

2 Aims of testing

Why test? Sometimes one wonders about the abundance of tests reported in empirical papers, as the purpose of many of these tests is not always communicated to the reader. Occasionally, the number of test statistics reported in a paper exceeds the number of observations used in calculating them! In many cases, the implications of a positive or negative result are not made clear. If a null hypothesis that apes behave perfectly rationally is rejected at the 5% significance level, do we care? And should we be interested in the normality of the residuals, or would it be more useful to put the tests aside and read Darwin's *Origin of Species* instead? But perhaps it is inherent to our occupation as econometricians that we stick to providing statistical inferences.

An important reason for the popularity of testing is that it is often thought to be a major if not the main ingredient to scientific progress (Popper, 1968; Stigler, 1965, p. 12; Blaug, 1980) and the best way to move from alchemy to science (remember Hendry's three golden rules of econometrics: test, test and test; *cf.* Hendry, 1980). According to Popper's demarcation criterion, scientific hypotheses are falsifiable ones. Unfalsifiable propositions belong to the domain of metaphysics, not science. You want to be scientific? Then test your hypotheses! And one of the founders of statistical testing writes 'Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies

may be raised to the rank of sciences' (Fisher, 1973a, p. 2).[2] Hence, if we wish to be scientific, then let us test hypotheses--statistical hypotheses, that is.

Occasionally, econometricians reach out to the scientific ideal of testing economic hypotheses, confronting theory (more precisely, a particular specification of the theory) with facts. We will call this theory testing. It is the most ambitious of the aims of testing. Examples are testing monetarism, real business cycle theory, the efficient market hypothesis, hysteresis, or properties of consumer demand. Ideally, tests in this category deal with efforts to test one theory against a rival one, that is, to discriminate (monetarism versus Keynesianism, hysteresis versus heterogeneity). Scientific progress, it is often argued, consists of replacing a defective theory by a better one. Non-nested hypotheses tests, encompassing tests, but also specification tests and, occasionally, model selection tests, belong to the category of theory testing.

Theory testing is closely related to a once popular approach in the philosophy of science, the hypothetico-deductive (HD) method.[3] This method consists of formulating sets of hypotheses, from which predictions of novel facts can be deduced: the consequences. These are the testable implications. The empirical scientist either should try to measure the degree of confirmation (according to logical positivists of the Vienna Circle, like Carnap) or try to falsify these testable implications (according to falsificationists like Popper). Prominent members of the Cowles Commission, in particular Haavelmo (1944) and Koopmans (1947), advocated an HD-approach to econometrics which resulted in a formalistic methodology of economic inference. More recently, HD-econometrics can be found in the writings of new classical economists, in particular by those who search for 'deep' (structural) parameters. Another recent publication in the tradition of the HD-approach is Stigum (1990). But Summers (1990) forcefully argues that formalistic empirical econometrics has not yielded

---

[2] Ironically, the quote continuous as follows: 'This particular dependence of social studies upon statistical methods has led to the unfortunate misapprehension that statistics is to be regarded as a branch of economics, whereas in truth methods adequate to the treatment of economic data, in so far as they exist, have mostly been developed in the study of biology and the other sciences.'

[3] See Chapter 3 in Earman (1992) for discussion and references.

interesting insights in macroeconomics: this approach to inference leads merely to a 'scientific illusion'.

Popper's falsificationism has had a strong impact on the minds of economists. Popper is about the only philosopher of science occasionally quoted in *Econometrica*. In the philosophy of science literature, however, falsificationism has become increasingly unpopular. Not in the least because actual science rarely follows the Popperian maxims. As Hacking (1983, p. 15) notes, 'accepting and rejecting is a rather minor part of science' (see also the contributions in Earman, 1983, and those in De Marchi, 1988). Theory testing is an aim that, in practice, is less important than some would like to think.

An alternative to hypothetico-deductivism is Bayesian inductive inference. Carnap (1952) also contributed to this approach, but Jeffreys (1961) had a stronger impact on the Bayesian minority in econometrics.[4] This alternative approach shares with the HD-method a belief in growth of knowledge (a feature that has been attacked by so-called post-modernist philosophy; see Mirowski, 1994). However, the aim of theory testing is less important in the Bayesian inductive tradition than within Popperian hypothetico-deductivism (see e.g. Leamer, 1978, p. 9). Some Bayesians do not see merit in hypothesis testing, they hold measurement as the more interesting aim of inference. If rival hypotheses exist, and e.g. prediction is the purpose of inference, the best one can do is to weigh the alternative hypotheses and use a basket of weighed predictions. Other Bayesians use Jeffreys' Posterior Odds Ratios as a test statistic. If decision making is the purpose of the test, then the behaviouristic approach of Savage (1972) is advocated by some Bayesians (below, we will discuss decision making as a distinct aim of testing).

If theory-testing is an interesting aim at all, it is not yet clear that econometrics is the best tool for this purpose. Identifying informative historical episodes (see e.g. Summers, 1991) or devising laboratory experiments (increasingly popular among game theorists, who rarely supplement their experiments with statistical analysis, as casual reading of such experimental

---

[4] See Howson and Urbach (1989) and Earman (1992) for philosophical backgrounds of Bayesian confirmation theory.

reports in *Econometrica* reveals) may generate more effective tests than many Uniformly Most Powerful (UMP) tests. Consider Science with capital S: physics. Here, sophisticated statistical considerations play a minor role in appraising theories. Giere (1988, p. 190) discusses the different attitudes towards data appraisal in nuclear physics and the social sciences. Nuclear physicists tend to judge the fit between empirical and theoretical models primarily on qualitative arguments. Test statistics such as χ^2 are rarely reported in nuclear physics papers contained in e.g. the *Physical Review*.[5] Theory (or hypothesis) testing does not necessarily depend upon the tools we learned in our statistics courses.

We will now turn to other aims of testing, less prominent in philosophical writings, but dominant in practical research. Most tests are not as ambitious as the theory tests discussed above. An important case is the class of the (statistical) validity tests (mis-specification tests or diagnostic checks). Validity tests are performed in order to find out whether the statistical assumptions underlying some model are credible. Spanos (1994) is an example of extensive validity testing. He follows the argument that in order to pursue a theory test, one first has to be sure of the validity of the statistical assumptions that are made. According to this view, validity testing is a prerequisite to theory testing (note that Granger *et al.*, 1994, advocate the reverse ordering). If theory testing is not the ultimate aim, validity testing still may be important. Much empirical work aims to show that a particular model (formally or informally related to some theory) is able to represent the data. If much information in the data remains unexploited (for example, revealed by non-white-noise residuals), this representation will be suspect or unconvincing to a large part of the audience.

Sometimes, however, it is argued that the merits of validity tests should not be over-emphasized. One may obtain a very neat 'valid' statistical model of some economic phenomenon, after extensive torturing of the data. Such a specification suggests much more precise knowledge than the data actually contain. Sensitivity analysis, either along the lines of Leamer (1978) or Friedman (see for example the discussion in Summers, 1991), is at least as important as validity testing in order to make credible inferences. Illuminating in this context is the exchange between Hendry and Ericsson (1991) and Friedman and Schwartz

[5] Baird (1988) makes a similar observation.

(1991).

A third important aim of testing is simplification testing. Simple models that do not perform notably worse than more complex ones are typically preferred to the complex ones. Inference conditional on exogeneity assumptions is often preferred to full information estimation. Still, it is regularly argued that, apart from convenience, there are no clear formal reasons why simple models deserve special credit (but see Keuzenkamp and McAleer, 1994, for discussion and further references). A popular view on simplification testing is that the researcher should start with a very general model, and perform a downward test strategy in which uninformative elements of a model are deleted (Hendry and Ericsson, 1991). In practice, many researchers feel that simplicity matters, but rather than testing from general to simple, they perform iterative simplification searches.

Finally, a frequently expressed goal of testing is decision making (e.g. Granger *et al.*, 1994). This view on testing, and its implementation to statistics, is primarily due to the Neyman-Pearson theory of inductive behaviour (Neyman and Pearson, 1928, 1933). The decision-theoretic approach to testing has been further elaborated by Wald and, from a Bayesian perspective, by Savage (1972). Lehmann (1986) is the authorative reference for the frequentist approach, while Berger (1985) provides the Bayesian arguments.

Decision making, based on statistical acceptance rules, can be important for process quality control, but may even be extended to the appraisal of theories. This brings us back to theory testing. Lakatos, the neo-Popperian philosopher, claims that the Neyman-Pearson version of theory testing 'rests completely on methodological falsificationism' (Lakatos, 1978, p. 25n). Apart from the fact that this reverts historical priority (the first German edition of Popper (1968) appeared in 1934), it is also at odds with Popper's own rejection of behaviourism (see Keuzenkamp, 1994, Chapter 3.4.4, for further discussion). Still, it may be argued that the Neyman-Pearson approach to theory testing (popularized in econometrics by Haavelmo, 1944) fits in the broader hypothetico-deductive approach, of which Popper's version is only

one brand.[6]

At this point, one of the most bitter disputes in science deserves special mention: the Fisher versus Neyman-Pearson controversy. One of the sources of their dispute was the aim of testing. While Neyman-Pearson acceptance rules can be placed in the hypothetico-deductive camp, the views of Fisher are closer to a Bayesian inductive approach.[7] Fisher's theory of estimation and testing is a theory of learning, meant for inductive *inference* from small samples. Neyman and Pearson opposed aiming at inductive inference. They interpret tests along behaviouristic lines, as acceptance rules in the context of repeated sampling. At best, Fisher was willing to support such an interpretation for problems in commerce or technology, but not for appraising scientific hypotheses. The reason is that in such cases, repeated sampling is a misleading fiction, and there is no well defined decision problem. Many advances made in science do not serve a well specified purpose, moreover, 'they may be put sooner or later to the service of a number of purposes, of which we can know nothing' (Fisher 1973b, pp. 106-7). Even if there would be a well specified decision problem, estimation was of more interest to Fisher than devising UMP tests.[8] Indeed, for many econometric papers that appear in the *Journal of Econometrics* and *Econometrica* among others, it is hard to define the decision problem and loss functions that should figure in the background if a Neyman-Pearson approach were followed.

Such doubts are shared by Savage (1972, p. 254) who writes that, although having tested many sharp null hypotheses, he is unable to give a satisfactory analysis of testing such hypotheses. To him, the role of extreme null hypotheses in science is 'obscure'. A problem with such hypotheses is that in many cases the loss associated with the alternative is zero, only

---

[6] Giere (1983) is a philosopher's view on theory testing, which is an augmented version of the Neyman-Pearson theory (without mentioning Neyman-Pearson).

[7] Although Fisher rejected Bayesianism in cases where there is no informative prior probability, he had an alternative: so-called fiducial inference (see Fisher, 1973b). This has been characterized as 'a bold attempt to make the Bayesian omelette without breaking the Bayesian eggs' (Savage, 1961, p. 578).

[8] For Fisher's views on the Neyman-Pearson methodology, see Fisher (1973b, pp. 42, 80, 103-107), and Section 3 below.

a loss (or gain) exists if the null is exactly satisfied. The behaviouristic theory of inference is difficult to apply in such circumstances. Still, many econometricians do test sharp null hypotheses, and think that these tests are straightforward applications of testing in the Neyman-Pearson tradition.

Many such sharp null hypotheses are of little scientific interest anyway. Still, even the best journals, such as the Journal of Econometrics, report tests of purchasing power parity or perfectly efficient markets, even if we are all aware that these theories are not literary true. Would it not be more interesting, in such cases, to measure how close the real world is to the ideal world of the theories? According to Leamer (1978, p. 9), hypothesis testing searches are rare, while Jeffreys (1961, p. 389) remarks that 'what are called significance tests in agricultural experiments seem to me to be very largely problems of pure estimation.' Jeffreys' argument, if applied to economics, would run like this. A labour economist has a very good idea of what to expect when estimating a model that analyzes the returns to schooling. His problem is to choose the variables, and obtain a sample of sufficient size, such that the effect of education and other variables of interest become detectable. It is the magnitude of the effects that is of primary interest. Any level of significance can be obtained by making the sample size large enough, unless the null hypothesis is exactly true (Berkson, 1938).

This concludes our discussion of four distinct aims of testing: theory testing, validity testing, simplification testing and testing for making decisions. We now turn to a number of statistical methods that serve these aims of testing.

3 Methods of statistical testing

Informal statistical testing of hypotheses has a long history (frequently cited examples of significance testing avant la lettre are Arbuthnot on male vs. female births in 1710, Mitchell on the distribution of stars in 1767, and Laplace in 1812; see e.g. Hacking, 1975, p. 168; Baird, 1988). In 1885, Edgeworth introduced the term 'significant' in statistics (Baird, 1988).

The modern approach to significance testing starts with Karl Pearson's goodness of fit test for large samples (Pearson, 1900). The basic philosophy of his testing procedures is as follows. A sample is used to estimate a test parameter of interest. The distribution under the null is known, and if the estimate falls too far into the tail of the distribution, one of the following conclusions must be drawn: either something very uncommon has happened, or the null hypothesis is wrong. The *P*-value (tail area integral) thus obtained is compared with a benchmark, like 0.01 or 0.05 (see Section 4, below). Subsequently, in statistical inference, the option that the null is wrong is chosen if the *P*-value falls beyond the benchmark level.

Henry L. Moore belonged to the first economists who applied Pearson's methods to economics (see also Stigler, 1965). W.S. Gosset, better known by his pseudonym Student, introduced the small sample *t*-test for the equality of means in 1908. But Fisher greatly extended the scope of testing, he also derived the correct degrees of freedom that belong to different applications of the tests.[9] Fisher also invented analysis of variance and the *F*-test (originally labelled as *z*-test).[10] His method of maximum likelihood remains widely used, but his reliance on the likelihood principle and conditional inference is not generally accepted.[11] In his doctoral thesis, Koopmans (1937) built on Fisher's methods of estimation and testing. It is notable that Tinbergen (1939), who supervised Koopmans' thesis, does not make use of significance testing. Instead, Tinbergen's approach is better characterized as importance testing. (Jeffreys would probably argue that this again is a problem of estimation, rather than testing.)

To sum up the Fisherian theory of significance testing, it contains the following characteristics:

i) reliance on tail areas (*P*-values);

---

[9] In 1922, Gosset sent his tables of the *t*-test to Fisher, writing 'you are the only man that's ever likely to use them!' See Joan Fisher Box (1978, pp. 116, 451).

[10] z is a transformation of F which was easier to interpolate ($z = 1/2 \ln F$).

[11] Lehmann (1993) discusses the issue of conditional inference and compares Fisher's perspective with that of Neyman.

ii) intended for small samples;

iii) instruments for inductive scientific inference.

This approach has received two kinds of criticism. The first comes from Jeffreys, who rejects i). The second criticism is given by Neyman and Egon Pearson (the son of Karl Pearson), who argue that i) alone is not sufficient to select a test procedure; ii) should be replaced by repeated sampling; and who also disagree with iii). To start with Jeffreys (1961, p. 385), he argues that any particular set of observations has a low probability to obtain, hence, 'If mere improbability of the observations, given the hypothesis, was the criterion, any hypothesis whatever would be rejected.' In the posterior odds approach, advocated by Jeffreys, this problem vanishes since the ratio of probability values for two distinct hypotheses will be informative; the small factors cancel. The *P*-integral methodology instead does not appraise the probability of the actual observations, in view of a hypothesis, but takes the observations that would generate *P*-values beyond the benchmark level. 'The latter gives the probability of departures, measured in a particular way, equal to *or greater than* the observed set, and the contribution from the actual value is nearly always negligible. *What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure.' (Jeffreys, 1961, p. 385, his italics). In other words, this method of inference violates the likelihood principle. Another criticism of Jeffreys (1961, p. 390) is that it is not very useful to reject a hypothesis without having some idea of what to put in its place (see also Keuzenkamp and Barten, 1994).

We already noted that the test approach advanced in Neyman and Pearson (1928) and further explored in their later writings (in particular, Neyman and Pearson, 1933) diverges from Fisher's in some important respects (but note that Neyman and Pearson adopted many of Fisher's insights, and at first were even convinced that their work was just an improvement of Fisher's; see Fisher Box, 1978, for details, and Reid, 1982, for a perspective that sides with Neyman and Pearson).

First, there is a philosophical distinction between Neyman-Pearson methods and Fisherian hypothesis testing. The Neyman-Pearson approach is not based on inductive aspirations (such

as Fisher's), but is directed to behaviour, following the then fashionable behaviouristic school of thought in psychology and other disciplines (J.B. Watson's classic on behaviourism appeared in 1930, and papers on that topic have appeared since 1913). The Neyman-Pearson tests are acceptance procedures, decision rules (see above), not methods of inference.

Secondly, Neyman and Pearson were dissatisfied with the existence of a wide range of tests while no one knew which one was the 'best'. According to Fisher, the research worker normally knows what alternatives are relevant (without specifying them) and, therefore, what test is to be selected. However, Neyman and Pearson tried to define general optimality conditions for tests, in a context of repeated sampling. This can only be done after the un-specified alternative hypotheses in the Fisher-approach are replaced by specific alternatives. Once this step is made, the notion of errors of the first kind (rejecting a correct null hypothesis) and the second kind (accepting a false null hypothesis) can be introduced, the power of a test is defined, and UMP tests can be obtained in a number of cases. The likelihood ratio (LR) test (first proposed on intuitive grounds in 1928, then justified on theoretical grounds in 1933), stands out as their principal contribution to the theory of hypotheses testing.

Summarizing, to contrast the Neyman-Pearson approach to Fisher's, the following points characterize the Neyman-Pearson methodology:

i) emphasis on size and power, leading to UMP tests;
ii) applications to contexts of repeated sampling;
iii) instruments for inductive behaviour and making decisions.

At the formal level Neyman and Pearson seem to have won the battle with Fisher. Many economists have learnt Neyman-Pearson methods of hypothesis testing in their first introductory course in statistics. For example, the treatment of hypothesis testing in the popular statistics textbook of Wonnacott and Wonnacott (1985) is based on a simplified version of Neyman-Pearson testing (an explicit reference to Neyman-Pearson is given on p.

257).[12] Another popular textbook used in econometrics is Judge *et al.* (1988). Hypothesis testing is interpreted as a decision problem in the light of the costs of making an incorrect decision (p. 93), the discussion is entirely in the spirit of Neyman-Pearson procedures. The encyclopaedic nature of this book is reflected in alternative discussions of hypothesis testing, in particular posterior odds (p. 131), but Fisher's approach is not discussed. Goldberger's (1991) textbook deals with hypothesis testing from a Neyman-Pearson perspective (Chapter 20) and even explains the Neyman-Pearson implication of a rejection of a null at a 5% significance level: 'Loosely speaking, when the null is true, in 5% of the samples drawn from the population, the decision will be "reject the null"' (Goldberger, 1991, p. 215). Finally, the survey paper of Engle (1984) gives an overview of test procedures (Lagrange Multiplier tests, starting at the null and testing whether movements to the alternative lead to an improvement; Wald tests, starting at the alternative; and LR tests that may proceed symmetrically), all based on Neyman-Pearson principles.

The implementation of Neyman-Pearson methods at the practical level is not easy, though. There is a wide divergence between empirical econometrics and the maxims of a 'celibate priesthood of statistical theorists', as Leamer (1978, p. vi) observes. One reason for the dominance of the Neyman-Pearson approach among this priesthood might be that it lends itself to mathematical recreation. Another non-substantive reason is the attraction that the words 'best' and 'powerful' exert. But it is more interesting to evaluate the substantive features of Neyman-Pearson testing. They have several drawbacks.

First, consider the notion of power. According to Fisher, emphasis on power is in many cases hardly relevant. To a practical researcher, 'it is, of course, a matter of indifference with what probability he might be led to accept the hypothesis falsely, for in his case he is not accepting it' (Fisher, 1973b, p. 42). Another problem with the power of a test is that it may be low when the model is mis-specified (i.e. the maintained hypothesis is wrong). On the other hand, some tests (such as the Durbin-Watson test) happen to be rather powerful against mis-specifications for which they are not intended. The Neyman-Pearson approach

---

[12] In a subtle deviation from the Neyman-Pearson interpretation of testing, Wonnacott and Wonnacott (1985, p. 259) argue that a statistical test is a device to judge the acceptability or plausibility of the hypothesis.

hinges on the 'axiom of correct specification' (Leamer, 1978, p. 4). Recently, efforts have been made to extend the scope of Neyman-Pearson methods to mis-specified models (an example is Vuong, 1989). An alternative, becoming increasingly popular, is to use non-parametric methods of inference. Some investigators who support this approach believe that they can avoid making the specification errors that afflict parametric inference (see also Härdle and Kirman, 1994).

Secondly, the fiction of repeated sampling is questionable. One of the first critics was Fisher. He states that 'if we possess a unique sample in Student's sense on which significance tests are to be performed, there is always, as Venn (1876) in particular has shown, a multiplicity of populations to each of which we can legitimately regard our sample as belonging: so that the phrase "repeated sampling from the same population" does not enable us to determine which population is to be used to define the probability level, for no one of them has objective reality, all being products of the statistician's imagination.' (Fisher, 1955, p. 71).[13]

Thirdly, although some argue that the decision-theoretical approach should be natural to economists, in many cases it is very difficult to determine what decision really inspires a particular test, and what loss is involved (see Section 2 above). Although the decision-theoretical approach to theory testing is obscure, it may be helpful in cases of validity testing, which has some resemblance to process quality control (if we are willing to ignore the Neyman-Pearson emphasis on repeated sampling). The loss, e.g. involved with serial correlation, might be that readers who stick to the 5% convention will stop reading a research report if they suspect that serial correlation is not properly taken care of. It still is not a formal loss, expressed in dollars, but loss resulting from loss of readership driven by (bad or good) conventions (see the blunt comments in Friedman, 1988, footnote 11). Although such a justification of Neyman-Pearson methods for validity testing could be sustained, one of the proponents of Neyman-Pearson methods makes a distinction between 'model design criteria (exhausting the available data evidence) and genuine tests in the Neyman-Pearson sense (based on previously unavailable evidence)' (Hendry, 1992, p. 366).

---

[13] The reference to Venn relates to the second edition of *The Logic of Chance*.

He adds that 'only information that arrived after a model is in the public domain can be deemed an adequate basis for a test' (Hendry 1992, p. 374). If we understand Hendry correctly, he argues that validity testing (model design) is not genuinely Neyman-Pearson, but theory testing is a kind of Neyman-Pearson quality control test. We agree with the first statement, as genuine Neyman-Pearson testing requires repeated sampling. For the same reason, the second statement seems less convincing (the accumulation of a handful of extra quarterly observations can hardly count as an instance of repeated sampling). Moreover, it is not clear how to interpret scientific inference as a genuine decision problem, to be solved with behaviouristic arguments.

A fourth problem with Neyman-Pearson testing is that, if we have two explicitly specified alternatives to choose from, it is more natural to choose the one with the higher likelihood without considering the power functions and without having to take one as the null and the other as the alternative (see Jeffreys, 1961, p. 396; a Bayesian would consider the posterior odds ratio.) Vuong (1989) discusses how the LR test can be used in a symmetric way for model selection and testing non-nested hypotheses in a context of independent observations.

A general problem of significance testing, whether Neyman-Pearson or Fisherian, occurs when multiple tests are carried out. Depending on how dependent these tests are, the overall significance level may be much higher than the individual significance levels. The problem was recognized by the early econometricians. Indeed, Haavelmo (1944, p. 83) already discusses, in today's parlance, pre-testing. It is valid, he argues, but not if the set of a priori admissible hypotheses is 'a function of the sample point'. This rules out to experiment with the maintained hypothesis. Naive induction, as one might call this method, cannot be totally ignored (to use an understatement) if one appraises empirical econometrics. Moreover, not only the maintained hypothesis may be the result of 'data mining', but not infrequently the alternative hypothesis is inspired by a rejection of a null rather than specified in advance, as it should in case of the Neyman-Pearson methods. The problem of interpreting the resulting test statistics remains unsolved today (see Godfrey 1988, p. 3; Leamer, 1978, p. 5). Indeed, as Hendry (1992, p. 369) notes, test statistics can frequently be made insignificant by construction, since the residuals are not autonomous but derived processes.

Many authors agree that significance tests are not the only or ultimate tests of economic hypotheses. Friedman is not alone in his verdict that 'the real test of a theory' lies in its predictive ability, a theme he has consistently repeated since his 1940 review of Tinbergen's (1939) statistical tests of business cycle theories. This ability may be evaluated quantitatively, with statistical tools, but also qualitatively. Theil (1971, p. 545) argues that statistical procedures are not sacrosanct in modelling: 'The real test is provided by prediction based on an independent set of data. It is not at all self-evident that selections that are exclusively based on the smallest residual-variance estimates lead to the best predictions.' Similar opinions are expressed by Hendry (1992, p. 374), Zellner (1988, p. 31) and numerous other econometricians.

4. What is 'significant'?

If economists have natural constants, then their values are 0.01 and 0.05. From early applications to the most recent hypothesis tests, investigators have relied on a significance level of 0.01 or 0.05. This convention owes much to Fisher's tabulation of statistical distributions in Fisher (1973a), first published in 1925.[14] Fisher and Gosset ('Student') cooperated in calculating tables for the t-distribution. Fisher also tabulated the distributions of χ^2 and the z-transformation of the F-distribution. Originally, Fisher hoped to include existing tables of χ^2, made by W.P. Elderton and published in Biometrika of 1902, in his book. However, Karl Pearson (editor of Biometrika, father of Egon Pearson) did not allow him to reprint those tables. Pearson did not approve of Fisher's refinements of interpreting the χ^2 test (in particular, the issue of degrees of freedom), and their personal relations were bad [see M.G. Kendall (1963) and Fisher Box (1978)]. Hence, Fisher was forced to make a distinct table by himself. He decided to turn the tables inside out, which seemed more convenient as well. Existing tables provided P-values (tail areas) for given values of χ^2 and t. Fisher argues that 'Instead of giving the values of P corresponding to an arbitrary series

[14] We are grateful to Jim Durbin for historical advice on this matter. See also Hall and Selinger (1986) for a discussion of the historical roots of the 5% convention.

of values of χ^2, we have given the values of χ^2 corresponding to specially selected values of P' (Fisher, 1973a, p. 79; see Fisher Box, 1978, pp. 246-7 for further background). The *P*-values for which the χ^2 distribution was tabulated (for n=1...30) are .99, .98, .90, .80, .70, .50, .30, .20, .10, .05, .02 and .01 (Fisher, 1973a, pp. 112-113). A similarly extensive tabulation is provided for the *t*-distribution (*op. cit.* p. 176). Hence, the .05 significance level is not singled out as one with special merit, although Fisher (1973a, pp. 114-5) writes that 'If the difference is many times greater than the standard error, it is certainly significant, and it is a convenient convention to take twice the standard error as the limit of significance; this is roughly equivalent to the corresponding limit P=.05 already used for the χ^2 distribution.' Finally, as the *z* (or *F*-) distribution needs separate tables for all significance levels, Fisher decided to tabulate this distribution for 'three especially important values of P' (Fisher 1973a, p. 228, pp. 244-9): .05, .01 and .001. Those are the significance levels that we observe as the few natural constants that economists rely on when they do empirical research. Still, Fisher (1973b, p. 42) warns against dogmatically applying a fixed level of significance in all circumstances.

Although Fisher was not the first statistician who tested at a 5%-significance-level, he facilitated its breakthrough by suggesting to use 'significant' as an abbreviation of 'significant at the 5%-level' and moreover by means of his convenient tabulation. His interest in small sample analysis is reflected by the fact that his tables run from n=1 to n=30 (and in some cases also include 60 and infinity). Fisher does not discuss what the appropriate significance levels are for large samples. Berkson (1938) observed that, as sample size grows to infinity, any sharp null hypothesis is likely to be rejected at a fixed significance level. This has yielded the suggestion to vary the significance level with sample size (Leamer, 1978). As we will see in the next section, this suggestion has been largely ignored in practice. A possible explanation is that statisticians try to measure parameters using some benchmark level of precision. A fixed significance level serves this purpose. If our conjecture is valid, we expect to find that models estimated with many observations to have a higher dimensional parameter vector (we did not attempt to test this hypothesis statistically).

Given the conventional significance levels, it remains to explain what they really mean. Most

textbooks ignore this issue, but there are notable exceptions. Wonnacott and Wonnacott (1985) prefer the expression 'statistically discernible at the 5% error level' to the more familiar phraseology 'statistically significant at the 5% significance level'. An explicit warning is given that statistical significance is not the same as importance (or substance, in Goldberger, 1991, p. 240). Furthermore, the discussion of the χ^2 test points to some limitations of hypotheses tests (Wonnacott and Wonnacott, 1985, pp. 488-9), in particular to the fact that such tests often give answers to the wrong question. Goldberger's (1991, p. 215) explanation of the meaning of rejection at a 5%-significance-level, quoted in Section 2 above, is the valid interpretation of Neyman-Pearson testing. But how often is the question that an econometrician has to answer a decision problem in the context of repeated sampling? Fisher's interpretation of a small P-value (which follows the tradition of Laplace to K. Pearson), that either something very unlikely has happened, or the null is false, may be more useful in econometric practice. A third alternative is to interpret P-values as odds factors. In this case, however, the Bayesian (posterior odds) perspective may be preferred, as Jeffreys already showed that posterior odds ratios often tell a different story than significance tests based on a fixed significance level (Jeffreys, 1961; Berger, 1985).

A different perspective on interpreting significance tests arises when one realizes that, at least in economics, most inferences are based on extensive data-mining. Karl Pearson objected to using arbitrary levels of significance to assess the validity of a hypothesis. Instead, statistics involves curve fitting and gradual approximation from poor fit to good fit, not from falsity to truth. Goodness of fit tests 'are used to ascertain whether a reasonable *graduation* curve has been achieved, and not to assert whether one or another hypothesis is true or false' (K. Pearson, letter to *Nature*, 1935, cited in Hall and Selinger, 1986, p. 359). This skeptical view on significance testing is not much heard today, Milton Friedman being one of the exceptions (see e.g. Friedman, 1988, p. 323, footnote 11).

5 Testing in the Journal of Econometrics

The strong emphasis in journals on significance testing not only exists in economics. The

JRSS (Journal of the Royal Statistical Society) is sometimes referred to as the JSSR (Journal of Statistically Significant Results).[15] Similarly, the economic literature abounds with significance tests. Zellner (1979) contains a small survey of 22 quantitative articles in five issues of different leading economic journals in 1978. He finds that significance testing is very popular, that 1% and 5% significance levels dominate, and power considerations are rarely discussed, despite the dominance of Neyman-Pearson methods in the training of economists. According to another survey, of Canterbery and Burkhardt (1983, p. 31), out of 542 empirical papers that appeared in the *American Economic Review*, *Journal of Political Economy*, *Economic Journal* and *Quarterly Journal of Economics* from 1973-1978, only three articles attempted to refute the hypothesis under investigation. Although this may sound unnerving, there is a reason why econometricians do not play the falsificationist game with much enthusiasm. In most cases, rejection of economic hypotheses is easy, whereas verification is hard (anyone with experience in economic modelling knows how difficult it can be to obtain models that are 'satisfactory').

For the purpose of investigating the significance of significance tests, we surveyed the papers in the Journal of Econometrics (excluding Annals), Volumes 1-46 (1973-1990). In total, 668 papers were counted. Of those papers, 17% have 'test' (or 'testing') in the title. Not all 668 papers contain data. 26% contain artificial data, used for Monte Carlo investigations. Of the papers containing empirical data, many use those data for the purpose of illustration only (this is obviously the case if, for example, 'Klein I' is estimated--one of the most popular models in this Journal). We excluded those papers from our analysis (in a few cases, the choice is somewhat arbitrary).

This left 137 papers (21%) with an empirical message that exceeds mere illustration. Among those papers, 99 made use of significance tests. The significance levels were (in increasing popularity):

.02 (1 paper),
.001 (2 papers),
.10 (2 papers),

[15] See Wonnacott and Wonnacott (1986, p. 573), who also discuss the 'editor's bias' of preferring significant test results.

.005 (5 papers),
.01 (26 papers),
.05 (63 papers).[16]

The choice of the significance level might depend on sample size, in view of Berkson's (1938) observation and similar recommendations of Jeffreys (1961, p. 435) and Leamer (1978, p. 105). Hence, we investigated the relation between significance level and sample size.[17] Indeed, a few explicit references can be found concerning sample size and the trade-off between 'expected loss from Type I and Type II errors', as in a paper in Vol. 22 where a sample of 728 observations inspires a 1% significance level. In a paper in Vol. 44, a null hypothesis is rejected at a 5% significance level which the author had rather preferred to accept. Given sample size (14,487 observations), the author argues that conventional significance levels are not appropriate. Instead, with such large samples, 'a case can be made for using a Bayesian procedure'. However, upon further analysis of the relation between significance level and sample size in empirical papers, it appears that the correlation between sample size and significance level is opposite to what might be expected. The correlation coefficient is positive and has a value of .20! Hence, in practice, the choice of significance levels seems arbitrary and depends more on convention and, occasionally, on the desire of an investigator to reject or accept a hypothesis rather than on a well defined evaluation of conceivable losses that might result from incorrect decisions.

The papers which explicitly attempt to test a theory statistically are rare (less than a dozen); the cases where a clear conclusion (acceptance or rejection of the theory) emerges, are even rarer. In cases where a decisive conclusion is obtained, the same volume may contain a test of the same hypothesis with the opposite result (e.g. tests of efficient markets in Vol. 4). If a theory is rejected (e.g. neoclassical production theory, Vol. 7, or the theory of demand, Vol. 15, both at a 1%-significance level), it often remains unclear what the implications are

---

[16] A number of papers refers to more than one significance level. In that case, the most stringent (lowest) level is reported. Where this could be verified, it turned out that papers where the significance level remains implicit ('the parameter is significant') all refer to the 5%-significance-level. Hence, in those cases where this could not be verified, we assume that the 5% level is applied as well. Papers which report *P*-values are not included in this count.

[17] As it is not our purpose to blame specific authors, we refer in the following only to volume numbers and not to specific papers.

(the 'not very constructive conclusion' is 'worth remembering', or 'rejection of the theory is not necessarily implied'). Occasionally, it is acknowledged that the implication of significance tests is often unclear: a rejection does not necessarily mean a rejection of the hypothesis of interest, as auxiliary hypotheses might be false instead (see also Keuzenkamp and Barten, 1994).

In analyzing significance tests published in the *Journal of Econometrics*, we were surprised that some elementary rules are occasionally violated. Sample sizes are not always reported. Investigators claim to test a hypothesis at an unspecified or a 95% significance level, when a 5% level is meant. The advice by Goldberger (1991, p. 217), to use correct wording, is appropriate not only to undergraduate students.

6 Theory testing and significance: a challenge

According to Engle (1984, p. 776), 'If the confrontation of economic theories with observable phenomena is the objective of empirical research, then hypothesis testing is the primary tool of analysis.' This is the view of a mainstream econometric theorist. This view puts high emphasis on testing, but many econometricians are aware of the limited impact of testing and concur in McCloskey's skepticism. Spanos (1986, p. 660) acknowledges that 'to my knowledge, no economic theory was ever abandoned because it was rejected by some empirical econometric test, nor was a clear-cut decision between competing theories made in lieu of the evidence of such a test.' The same verdict has been expressed by economic theorists like Hahn (1992): 'I know of no economic theory which all reasonable people would agree to have been falsified.' Not only theorists argue like this. Summers (1991, p. 133) writes that 'It is difficult to think today of many empirical studies from more than a decade ago whose bottom line was a parameter estimate or the acceptance or rejection of a hypothesis.' In many cases, formal econometric hypothesis testing is unpersuasive. The value

added of econometric tests may be less than desired. Some even argue that we only test what we already believe beforehand. According to Keynes (1921), 'The truth is that sensible investigators only employ the correlation coefficient to test or confirm conclusions at which they have arrived on other grounds. But that does not validate the crude way in which the argument is sometimes presented, or prevents it from misleading the unwary-since not all investigators are sensible.'

These skeptical observations on significance testing for the purpose of theory testing should challenge econometricians who think otherwise. Therefore, we invite readers to name a paper that contains significance tests which significantly changed the way you think about some economic proposition. The following rules of the game apply:

1. You may interpret the notion 'significance test' broadly, i.e. both Fisher's and Neyman-Pearson's interpretations are accepted (please indicate if one of those interpretations is most appropriate).
2. Give exact reference to author, paper, journal etc., and to the particular test(s) you think persuaded economists.
3. Summarize the test result by:
   a) what is the hypothesis tested;
   b) is the hypothesis accepted or rejected;
   c) if the hypothesis is rejected, is there a constructive message?
4. If possible, provide auxiliary evidence that the particular test has been persuasive to others.

The responses to this challenge will be processed statistically and if the results are of sufficient interest, they will be reported. Your may send us your suggestion until six months after publication of this Issue. The most convincing contribution will be awarded with an invitation for a one week visit to CentER (expenses paid).

References

Baird, Davis, 1988, Significance Tests, History and Logic, in: Samuel Kotz and Norman L. Johnson (eds.), Encyclopedia of Statistical Sciences, Volume 8 (John Wiley and Sons, New York).

Berger, James O., 1985, Statistical decision theory and Bayesian analysis, second edition (Springer Verlag, Berlin).

Berkson, J., 1938, Some difficulties of interpretation encountered in the application of the chi-squared test, Journal of the American Statistical Association 33, 526-542.

Blaug, Mark, 1980, The Methodology of Economics or How Economists Explain (Cambridge University Press, Cambridge).

Canterbery, E. Ray and Robert J. Burkhardt, 1983, What do we mean by asking whether economics is a science?, in: Alfred S. Eichner, ed., Why Economics is not yet a Science (MacMillan Press, London) 15-40.

Carnap, Rudolph, 1952, The continuum of inductive methods (University of Chicago Press, Chicago).

De Marchi, Neil, ed., 1988, The Popperian Legacy in Economics (Cambridge University Press, Cambridge).

Earman, John, 1983, Testing Scientific Theories, Minnesota Studies in the Philosophy of Science (University of Minnesota Press, Minneapolis).

Earman, John, 1992, Bayes or Bust? (MIT Press, Cambridge).

Engle, Robert F., 1984, Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics, in: Z. Griliches and M.D. Intriligator (eds.), Handbook of Econometrics, Volume II (North Holland, Amsterdam), 775-826.

Fisher, R.A., 1955, Statistical methods and scientific induction, Journal of the Royal Statistical Society B 17, 69-78.

Fisher, R.A., 1973a, Statistical Methods for Research Workers, 14th ed. (Hafner Publishing Company, New York).

Fisher, R.A., 1973b, Statistical Methods and Scientific Inference, 3rd ed. (Hafner Press, New York).

Fisher Box, Joan, 1978, R.A. Fisher, The Life of a Scientist (John Wiley and Sons, New York).

Friedman, Milton, 1940, Review of Tinbergen (1939), American Economic Review XXX, 657-661.

Friedman, Milton, 1988, Money and the stock market, Journal of Political Economy 96, 221-239.

Friedman, Milton and Anna J. Schwartz, 1991, Alternative approaches to analyzing economic data, American Economic Review 81, 39-49.

Giere, Ronald N., Testing theoretical hypotheses, in: Earman (1983, 269-298).

Giere, Ronald N., 1988, *Explaining Science, a cognitive approach* (University of Chicago Press, Chicago).

Godfrey, L.G., 1988, Misspecification tests in econometrics (Cambridge University Press, Cambridge).

Goldberger, Arthur S., 1991, A Course in Econometrics (Harvard University Press, Cambridge).

Granger, Clive, Maxwell L. King and Halbert White, 1994, Testing economic theories and the use of model selection criteria, Journal of Econometrics-Annals, this Issue.

Haavelmo, Trygve, 1944, The probability approach in econometrics, Econometrica 12, Supplement.

Hacking, Ian, 1975, The emergence of probability (Cambridge University Press, Cambridge).

Hacking, Ian, 1983, Representing and Intervening, Introductory topics in the philosophy of natural science (Cambridge University Press, Cambridge).

Hall, P. and B. Selinger, 1986, Statistical significance: balancing evidence against doubt, Australian Journal of Statistics 28, 354-370.

Hahn, Frank, 1992, Answer to Backhouse, RES Newsletter nr. 78, July.

Härdle, Wolfgang and Alan Kirman, 1994, Non classical demand: a model-free examination of price quantity relations in the Marseille fish market, Journal of Econometrics-Annals, this Issue.

Hendry, David F., 1980, Econometrics: Alchemy or Science?, Economica 47, 387-406.

Hendry, David F., 1992, Assessing Empirical Evidence in Macroeconometrics with an Application to Consumers' Expenditure in France, in: Allessandro Vercelli and Nicola Dimitri (eds.), Macroeconomics: a Survey of Research Strategies (Oxford University Press, Oxford), 363-392.

Hendry, David F. and Neil R. Ericsson, 1991, An econometric analysis of U.K. money

demand in *Monetary Trends in the United States and the United Kingdom* by Milton Friedman and Anna J. Schwartz, American Economic Review 81, 8-38.

Howson, Colin and Peter Urbach, 1989, Scientific Reasoning, The Bayesian Approach (Open Court, La Salle).

Jeffreys, Harold, 1961, Theory of Probability, 3rd ed. (Clarendon Press, Oxford).

Judge, George G., R. Carter Hill, William E. Griffiths, Helmut Lütkepohl and Tsoung-Chao Lee, 1988, Introduction to the theory and practice of econometrics, 2nd ed. (John Wiley and Sons, New York).

Keuzenkamp, Hugo A., 1994, Probability, Econometrics and Truth, A Treatise on the Foundations of Econometric Inference, unpublished Ph.D. thesis (Dept. of Economics, Tilburg).

Keuzenkamp, Hugo A. and Anton P. Barten, 1994, Rejection without falsification, on the history of testing the homogeneity condition in the theory of consumer demand, Journal of Econometrics-Annals, this Issue.

Keuzenkamp, Hugo A. and Michael McAleer, 1994, Simplicity and econometric inference (manuscript, Dept. of Economics, Tilburg).

Keynes, John Maynard, 1921, A Treatise on Probability, The Collected Writings of John Maynard Keynes, VIII (St. Martin's Press, New York).

Koopmans, Tjalling, 1937, Linear Regression Analysis of Economic Time Series (De Erven F. Bohn, Haarlem).

Koopmans, Tjalling, 1947, Measurement without theory, Review of Economic Statistics 29, 161-172.

Lakatos, Imre, 1978, Philosophical Volume Papers I: The Methodology of Scientific Research Programmes (Cambridge University Press, Cambridge).

Leamer, Edward E., 1978, Specification Searches (John Wiley and Sons, New York).

Lehmann, E.L., 1986, Testing Statistical Hypotheses, second edition (John Wiley and Sons, New York).

Lehmann, E.L., 1993, The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? Journal of the American Statistical Association 88, 1242-1249.

McCloskey, Donald, 1985, The rhetoric of economics (University of Wisconsin Press, Madison).

Mirowski, Philip, 1994, Three ways to think about testing in econometrics, Journal of Econometrics-Annals, this Issue.

Neyman, Jerzy and Egon S. Pearson, 1928, On the use and interpretation of certain test criteria for purposes of statistical inference, I & II, Biometrika 20A, 175-200; 263-294.

Neyman, Jerzy and Egon S. Pearson, 1933, On the problem of the most efficient test of statistical hypotheses, Philosophical Transactions of the Royal Society A 231, 289-337.

Pearson, Karl, 1900, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philosophical Magazine* 50, 157-172.

Popper, Karl R., 1968, The Logic of Scientific Discovery, second edition (Harper & Row, New York).

Reid, Constance, 1982, Neyman-From Life (Springer Verlag, Berlin.

Savage, Leonard J., 1961, The foundations of statistics reconsidered, in: Jerzy Neyman (ed.), Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume I (University of California Press, Berkeley).

Savage, Leonard J., 1972, The Foundations of Statistics, second revised edition (Dover, New York).

Spanos, Aris, 1986, Statistical Foundations of Econometric Modelling (University of Cambridge Press, Cambridge).

Spanos, Aris, 1994, On theory testing in econometrics: the case of the efficient market hypothesis, Journal of Econometrics-Annals, this Issue.

Stigler, George J., 1965, Essays in the history of economics (University of Chicago Press, Chicago).

Stigum, Bernt P., 1990, Toward a Formal Science of Econometrics (MIT Press, Cambridge).

Summers, Lawrence H., 1991, The Scientific Illusion in Empirical Macroeconomics, Scandinavian Journal of Economics 93, 129-148.

Theil, Henri, 1971, Principles of Econometrics (John Wiley and Sons, New York).

Tinbergen, Jan, 1939, Statistical testing of Business Cycle Theories, two volumes (League of Nations Economic Intelligence Service, Geneva).

Vuong, Quang H., 1989, Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses, Econometrica 57, 307-333.

Wonnacott, Ronald J. and Thomas H. Wonnacott, 1985, Introductory Statistics, fourth edition (John Wiley and Sons, New York).

Zellner, Arnold, 1979, Posterior Odds Ratios for regression hypotheses: general

considerations and some results. Reprinted in: Basic Issues in Econometrics (University of Chicago Press, Chicago) 275-305.

Zellner, Arnold, 1988, Bayesian Analysis in Econometrics, Journal of Econometrics 37, 27-50.

Discussion Paper Series, CentER, Tilburg University, The Netherlands:

(For previous papers please consult previous discussion papers.)

No.	Author(s)	Title
9324	H. Huizinga	The Financing and Taxation of U.S. Direct Investment Abroad
9325	S.C.W. Eijffinger and E. Schaling	Central Bank Independence: Theory and Evidence
9326	T.C. To	Infant Industry Protection with Learning-by-Doing
9327	J.P.J.F. Scheepens	Bankruptcy Litigation and Optimal Debt Contracts
9328	T.C. To	Tariffs, Rent Extraction and Manipulation of Competition
9329	F. de Jong, T. Nijman and A. Röell	A Comparison of the Cost of Trading French Shares on the Paris Bourse and on SEAQ International
9330	H. Huizinga	The Welfare Effects of Individual Retirement Accounts
9331	H. Huizinga	Time Preference and International Tax Competition
9332	V. Feltkamp, A. Koster, A. van den Nouweland, P. Borm and S. Tijs	Linear Production with Transport of Products, Resources and Technology
9333	B. Lauterbach and U. Ben-Zion	Panic Behavior and the Performance of Circuit Breakers: Empirical Evidence
9334	B. Melenberg and A. van Soest	Semi-parametric Estimation of the Sample Selection Model
9335	A.L. Bovenberg and F. van der Ploeg	Green Policies and Public Finance in a Small Open Economy
9336	E. Schaling	On the Economic Independence of the Central Bank and the Persistence of Inflation
9337	G.-J. Otten	Characterizations of a Game Theoretical Cost Allocation Method
9338	M. Gradstein	Provision of Public Goods With Incomplete Information: Decentralization vs. Central Planning
9339	W. Güth and H. Kliemt	Competition or Co-operation
9340	T.C. To	Export Subsidies and Oligopoly with Switching Costs
9341	A. Demirgüç-Kunt and H. Huizinga	Barriers to Portfolio Investments in Emerging Stock Markets
9342	G.J. Almekinders	Theories on the Scope for Foreign Exchange Market Intervention

No.	Author(s)	Title
9343	E.R. van Dam and W.H. Haemers	Eigenvalues and the Diameter of Graphs
9344	H. Carlsson and S. Dasgupta	Noise-Proof Equilibria in Signaling Games
9345	F. van der Ploeg and A.L. Bovenberg	Environmental Policy, Public Goods and the Marginal Cost of Public Funds
9346	J.P.C. Blanc and R.D. van der Mei	The Power-series Algorithm Applied to Polling Systems with a Dormant Server
9347	J.P.C. Blanc	Performance Analysis and Optimization with the Power-series Algorithm
9348	R.M.W.J. Beetsma and F. van der Ploeg	Intramarginal Interventions, Bands and the Pattern of EMS Exchange Rate Distributions
9349	A. Simonovits	Intercohort Heterogeneity and Optimal Social Insurance Systems
9350	R.C. Douven and J.C. Engwerda	Is There Room for Convergence in the E.C.?
9351	F. Vella and M. Verbeek	Estimating and Interpreting Models with Endogenous Treatment Effects: The Relationship Between Competing Estimators of the Union Impact on Wages
9352	C. Meghir and G. Weber	Intertemporal Non-separability or Borrowing Restrictions? A Disaggregate Analysis Using the US CEX Panel
9353	V. Feltkamp	Alternative Axiomatic Characterizations of the Shapley and Banzhaf Values
9354	R.J. de Groof and M.A. van Tuijl	Aspects of Goods Market Integration. A Two-Country-Two -Sector Analysis
9355	Z. Yang	A Simplicial Algorithm for Computing Robust Stationary Points of a Continuous Function on the Unit Simplex
9356	E. van Damme and S. Hurkens	Commitment Robust Equilibria and Endogenous Timing
9357	W. Güth and B. Peleg	On Ring Formation In Auctions
9358	V. Bhaskar	Neutral Stability In Asymmetric Evolutionary Games
9359	F. Vella and M. Verbeek	Estimating and Testing Simultaneous Equation Panel Data Models with Censored Endogenous Variables
9360	W.B. van den Hout and J.P.C. Blanc	The Power-Series Algorithm Extended to the *BMAP/PH*/1 Queue
9361	R. Heuts and J. de Klein	An (s,q) Inventory Model with Stochastic and Interrelated Lead Times

No.	Author(s)	Title
9362	K.-E. Wärneryd	A Closer Look at Economic Psychology
9363	P.J.-J. Herings	On the Connectedness of the Set of Constrained Equilibria
9364	P.J.-J. Herings	A Note on "Macroeconomic Policy in a Two-Party System as a Repeated Game"
9365	F. van der Ploeg and A. L. Bovenberg	Direct Crowding Out, Optimal Taxation and Pollution Abatement
9366	M. Pradhan	Sector Participation in Labour Supply Models: Preferences or Rationing?
9367	H.G. Bloemen and A. Kapteyn	The Estimation of Utility Consistent Labor Supply Models by Means of Simulated Scores
9368	M.R. Baye, D. Kovenock and C.G. de Vries	The Solution to the Tullock Rent-Seeking Game When R > 2: Mixed-Strategy Equilibria and Mean Dissipation Rates
9369	T. van de Klundert and S. Smulders	The Welfare Consequences of Different Regimes of Oligopolistic Competition in a Growing Economy with Firm-Specific Knowledge
9370	G. van der Laan and D. Talman	Intersection Theorems on the Simplotope
9371	S. Muto	Alternating-Move Preplays and *vN - M* Stable Sets in Two Person Strategic Form Games
9372	S. Muto	Voters' Power in Indirect Voting Systems with Political Parties: the Square Root Effect
9373	S. Smulders and R. Gradus	Pollution Abatement and Long-term Growth
9374	C. Fernandez, J. Osiewalski and M.F.J. Steel	Marginal Equivalence in *v*-Spherical Models
9375	E. van Damme	Evolutionary Game Theory
9376	P.M. Kort	Pollution Control and the Dynamics of the Firm: the Effects of Market Based Instruments on Optimal Firm Investments
9377	A. L. Bovenberg and F. van der Ploeg	Optimal Taxation, Public Goods and Environmental Policy with Involuntary Unemployment
9378	F. Thuijsman, B. Peleg, M. Amitai & A. Shmida	Automata, Matching and Foraging Behavior of Bees
9379	A. Lejour and H. Verbon	Capital Mobility and Social Insurance in an Integrated Market

No.	Author(s)	Title
9380	C. Fernandez, J. Osiewalski and M. Steel	The Continuous Multivariate Location-Scale Model Revisited: A Tale of Robustness
9381	F. de Jong	Specification, Solution and Estimation of a Discrete Time Target Zone Model of EMS Exchange Rates
9401	J.P.C. Kleijnen and R.Y. Rubinstein	Monte Carlo Sampling and Variance Reduction Techniques
9402	F.C. Drost and B.J.M. Werker	Closing the Garch Gap: Continuous Time Garch Modeling
9403	A. Kapteyn	The Measurement of Household Cost Functions: Revealed Preference Versus Subjective Measures
9404	H.G. Bloemen	Job Search, Search Intensity and Labour Market Transitions: An Empirical Exercise
9405	P.W.J. De Bijl	Moral Hazard and Noisy Information Disclosure
9406	A. De Waegenaere	Redistribution of Risk Through Incomplete Markets with Trading Constraints
9407	A. van den Nouweland, P. Borm, W. van Golstein Brouwers, R. Groot Bruinderink, and S. Tijs	A Game Theoretic Approach to Problems in Telecommunication
9408	A.L. Bovenberg and F. van der Ploeg	Consequences of Environmental Tax Reform for Involuntary Unemployment and Welfare
9409	P. Smit	Arnoldi Type Methods for Eigenvalue Calculation: Theory and Experiments
9410	J. Eichberger and D. Kelsey	Non-additive Beliefs and Game Theory
9411	N. Dagan, R. Serrano and O. Volij	A Non-cooperative View of Consistent Bankruptcy Rules
9412	H. Bester and E. Petrakis	Coupons and Oligopolistic Price Discrimination
9413	G. Koop, J. Osiewalski and M.F.J. Steel	Bayesian Efficiency Analysis with a Flexible Form: The AIM Cost Function
9414	C. Kilby	World Bank-Borrower Relations and Project Supervision
9415	H. Bester	A Bargaining Model of Financial Intermediation
9416	J.J.G. Lemmen and S.C.W. Eijffinger	The Price Approach to Financial Integration: Decomposing European Money Market Interest Rate Differentials

No.	Author(s)	Title
9417	J. de la Horra and C. Fernandez	Sensitivity to Prior Independence via Farlie-Gumbel-Morgenstern Model
9418	D. Tolman and Z. Yang	A Simplicial Algorithm for Computing Proper Nash Equilibria of Finite Games
9419	H.J. Bierens	Nonparametric Cointegration Tests
9420	G. van der Laan, D. Talman and Z. Yang	Intersection Theorems on Polytopes
9421	R. van den Brink and R.P. Gilles	Ranking the Nodes in Directed and Weighted Directed Graphs
9422	A. van Soest	Youth Minimum Wage Rates: The Dutch Experience
9423	N. Dagan and O. Volij	Bilateral Comparisons and Consistent Fair Division Rules in the Context of Bankruptcy Problems
9424	R. van den Brink and P. Borm	Digraph Competitions and Cooperative Games
9425	P.H.M. Ruys and R.P. Gilles	The Interdependence between Production and Allocation
9426	T. Callan and A. van Soest	Family Labour Supply and Taxes in Ireland
9427	R.M.W.J. Beetsma and F. van der Ploeg	Macroeconomic Stabilisation and Intervention Policy under an Exchange Rate Band
9428	J.P.C. Kleijnen and W. van Groenendaal	Two-stage versus Sequential Sample-size Determination in Regression Analysis of Simulation Experiments
9429	M. Pradhan and A. van Soest	Household Labour Supply in Urban Areas of a Developing Country
9430	P.J.J. Herings	Endogenously Determined Price Rigidities
9431	H.A. Keuzenkamp and J.R. Magnus	On Tests and Significance in Econometrics