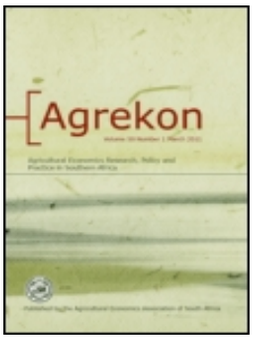

Routledge
Taylor & Francis Group

Agrekon

Agricultural Economics Research, Policy and Practice in Southern Africa

ISSN: 0303-1853 (Print) 2078-0400 (Online) Journal homepage: www.tandfonline.com/journals/ragr20

The standard error of regressions: a note on new evidence of significance misuse

C.N. Mbatha & M.A. Gustafsson

To cite this article: C.N. Mbatha & M.A. Gustafsson (2013) The standard error of regressions: a note on new evidence of significance misuse, Agrekon, 52:1, 28-39, DOI: 10.1080/03031853.2013.778463

To link to this article: https://doi.org/10.1080/03031853.2013.778463

THE STANDARD ERROR OF REGRESSIONS: A NOTE ON NEW EVIDENCE OF SIGNIFICANCE MISUSE

C.N. Mbatha*and M.A. Gustafsson**

ABSTRACT

There is a body of literature dealing with the improper use of statistical significance within economic analysis. Amongthe problematic usages that have been identified are fundamental misunderstandings about the influence of sample design and size on statistical significance, an excessive focus on statistical significance to the exclusion of economic and policy significance, and a harmful conflation of these two very different types of significance. An analysis of 51 agricultural economics papers reviewed and presented at an African conference in 2010 finds improper usage of statistical significance that is comparable or worse in nature and extent to that found in a previous meta analysis focusing on published articles in the *American Economic Review* in the 1980s and 1990s: well over half of the papers employed what is termed "sign" and "asterisks" econometrics. Overall, the findings underline the need for clearly stated and consistent analytical methods in producing papers as well as for careful review and selection of papers that employ regression analysis.

Keywords: significance, coefficients, sizes, signs

JEL Classification: F14

1. INTRODUCTION

McCloskey and Ziliak (1996) provided evidence about the nature and prevalence of unsound statistical practices in papers published in the *American Economic Review* (AER) in the 1980s. In 2004 they repeated the survey for AER articles published in the 1990s and found some improvements, similar results in some aspects, and worse results in other aspects. Their methodology was also replicated and applied by Parcell *et al*. (2000) to papers published between 1994 and 1998

* Corresponding author, Microeconomics lecturer at the University of South Africa, Department of Economics: email nhlanhla.mbatha@gmail.com (+27836704287)

** Researcher at the University of Stellenbosch, Department of Economics: email mgustafsson@sun.ac.za

Agrekon
Vol. 52 (1) 2013 • ISSN Print 0303-1853/Online 2078-0400

DOI: 10.1080/03031853.2013.778463

in the *Agricultural and Resource Economics Review* journal. Among other things found by McCloskey and Ziliak (1996) was that statistical significance was often interpreted by analysts as identical with economic or policy significance. The majority of papers focused only on the signs of variable coefficients presented and their associated t-statistic values, without discussing the magnitudes of such coefficients. In this paper, we argue that statistical tests should be motivated by the focus of the research: unnecessary routine t-tests are subject to misinterpretation. We then adapt elements of the McCloskey and Ziliak (1996) survey to papers presented at an African agricultural economics conference[1] and show that researchers still fail to interpret statistical significance tests soundly and procedurally as predetermined by theory. Our results are comparable in some aspects with McCloskey and Ziliak (1996). We find that while a majority (82%) of African papers reflect sound sampling practices, which is a similar figure to those reported in AER papers, only 24% of African papers avoid ordering coefficients based on the size of their associated t-statistic values for policy conclusions. Many of the papers only discuss the signs of the coefficients without discussing their magnitudes with respect to responsiveness between variables.

In section 2 of the paper the use and misuse of statistical significance are explained and the framework for evaluating papers is presented. Section 3 presents evidence from the AAAE conference papers and comparative discussion with AER papers. Concluding notes and some recommendations for future research and reviews of papers are presented in section 4.

2. THEORETICAL PROBLEM

Arrow (1959), McCloskey (1985), McCloskey and Ziliak (1996), and Parcell *et al*. (2000) have all illustrated and cautioned against the misinterpretation of tests of statistical significance – both the t-tests and p-values. From the evidence in the introduction above and in the following sections it appears that the same mistakes are being repeated over time in different geographical areas and fields of study.

2.1. The misuse of statistical significance tests

In regression analysis the t-tests are meant to establish whether an estimated coefficient of an independent variable is statistically reliable over "an infinite number of random samples" (Nattrass, 2005). We accept as reliable the direction and size of the estimate – normally the estimated elasticity in the relationship between the dependent and independent variables – if the t-test score > 1.96[2] or p score < 0.05. Venn (1988), cited in Hogben (1968:325), said that the values "inform us which of the differences (i.e. estimated coefficients) ... are permanent and significant, in the sense that we may be tolerably confident that if we took

another similar batch (sample) we should find a similar difference". It is therefore a misinterpretation of the test to claim that an estimated coefficient of a variable is significant for the determination of policy when the coefficient is based solely on the t-ratio being greater than 1.96, since the ratio tests only the sample for reliability in its estimation of parameters in the population over a number of experiments. A useful example of how an estimate could be statistically significant but not significant for policy purposes is presented by McCloskey and Ziliak (1996) regarding the Purchasing Price Parity (PPP) across two countries[3].

If on the other hand the sample is not reliable in its estimations of population parameters, we should expect the t-test values to increase every time the size of the sample is increased. We illustrate this using the Gustafsson and Patel (2009) specification of an earnings function[4]. When the average of the p-values associated with each of the coefficients is measured as the sample size of the labour market data is increased, the average p-value drops. Figure 1 presents the relationship between the average of p-values associated with the Beta estimates in the equation (see footnote 4) against increasing sample sizes on the x-axis.

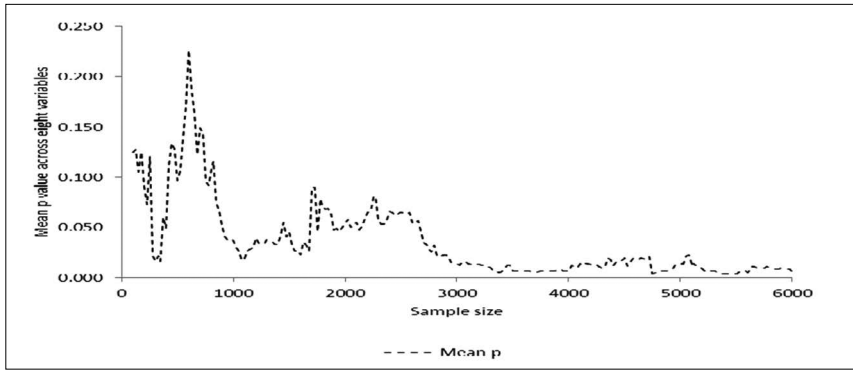

Figure 1: Statistical significance and sample size

Figure 1 illustrates, firstly, one of the ways in which, by relying on p-values from poorly designed samples, hypotheses may be wrongly rejected or accepted. Secondly, Figure 1 shows that p-values are not measures of substantive significance but of sample reliability. In Figure 1, below 1 000 observations, the sample is generally unreliable and its estimates are statistically insignificant, but when the sample contains more than 3 000 observations, on average the estimates are stable or statistically significant $p < 0.05$ because the sample can better predict true parameters. The PPP and sample size examples also illustrate that researchers should *think* more carefully about the context in which to interpret the magnitudes of the estimators – for example, with respect to social policy issues. The questions

that should be asked are: How large are the coefficients or differences, and do these sizes matter enough for policy purposes (e.g. β=0.999 vs. β=1.0)? Such questions and their answers cannot be relegated to t-tests or p-values, because the tests provide information mainly about sampling procedures as opposed to information about the importance of the specified parameters for policy purposes. Moreover, the estimators should be presented and interpreted in a way that considers the degree of responsiveness or elasticity between two variables (i.e. magnitude and direction), and not conversely[5]. Otherwise researchers would be committing what McCloskey and Ziliak (1996) refer to as "asterisks econometrics", where parameters are highlighted for discussion and as important for policy merely because of the order derived from p-values (with *** symbols to mark the most important). Associated with this practice is what they term "sign (+ or -) econometrics", where researchers report only the signs of the coefficient estimators alongside the p-values but do not present or discuss the magnitudes of coefficients.

In addition, since the t-distribution depends on the assumptions of the classical linear regression model, it is symmetrical and bell-shaped with a mean and mode of zero. In cases where the null hypothesis is zero, which is often the case with regression models, the test ratio is given by:

$$t_{calc} = \text{sample estimate/standard error}^{6}.$$

For a regression: t_{calc} = regression coefficient/standard error of the regression coefficient.

The t-ratio for a particular parameter can be large, but wrong, if some key variable is left out of the equation or specification. The implication is that pre-testing, by dropping and including variables to achieve high t-test values[7], also increases the chances of committing a Type II error[8]. The t-test should be performed for a pre-specified parameter in the hypothesis (see Wallace, 1977 and Parcell *et al*., 2000). It is therefore essential that the research aims guide the hypothesis, that model specification should lead to the statistical analysis and that t-tests be performed on variables that have been already identified as important in the hypothesis, not randomly. Even though pre-testing is common, especially with the use of new computer software that makes it easier to perform tests, it is still difficult (if not impossible) to detect if pre-testing is not explicitly self reported[9]. As pointed out by Parcell *et al*. (2000:175), "stating that statistically insignificant variables were dropped is a dilemma for the author because it is admitting to pre-testing". Hence, the investigation of this paper is limited to investigating what we also call "asterisks" and "sign" econometrics, because these are readily observable. For purposes of the evaluation, eight key questions are selected from McCloskey and Ziliak (1996).

2.2. The framework for analysis and AER findings

McCloskey and Ziliak (1996) reviewed papers that used regression analysis in the AER journal in the 1980s and repeated the review for the 1990s (McCloskey and Ziliak, 2004). They initially posed 19 questions and graded their findings a "yes" (for a sound practice), a "no" (for unsound practice), and a "not applicable". The results for eight of their key questions, which are also adopted in this paper, follow:

i) Does the paper use a small enough sample, such that statistical significance does not increase to acceptable levels by merely increasing the sample size?

The power of the test is high when "similar" and acceptable statistical significance levels are carried through from small to large samples. In that case, the chances of finding statistically meaningful but economically meaningless significance values are low. 87% of the papers scored a "yes" for sound sampling in the 1980s AER papers, but in the 1990s the figure decreased to 67.9%.

ii) Are the units and descriptive statistics for all regression variables included?

This is an elementary requirement in the empirical work of measurement. Only 32% of the AER papers scored a "yes" for performing this exercise in the 1980s. In the 1990s the practice improved to 66.4%.

iii) Are coefficients reported in elasticity form, or in some interpretable form relevant to the research question and consistent with economic theory?

Sound practice should not only report on the size of the coefficients, it should also interpret their economic effect. Around 66% of the papers scored a "yes" for sound practices in the 1980s. In the 1990s there was a marked improvement when 87% of papers reported on coefficients in elasticity form.

iv) Does the paper avoid presenting statistical significance as the scientific crescendo of the paper, or as the only serious criterion of importance?

Often papers identify for discussion the coefficients whose statistical significance values are acceptable. Along these lines, asterisks alongside statistical significance values are used often to rank the coefficients on their so-called "importance". 47% of the papers scored a "yes" for avoiding this unsound practice in the 1980s. In the 1990s the figure decreased to 36.5%.

v) Does the paper avoid ranking the coefficients according to their associated statistical significance values (asterisks econometrics)?

Authors who do this often sort the coefficients according to their number of asterisks. Close to 75% of the papers scored a "yes" for avoiding this practice in the 1980s, but in the 1990s the situation deteriorated further with only 32.8% of papers avoiding the practice.

vi) Does the paper avoid discussing only the signs of the coefficients, without discussing their sizes (sign econometrics)?

Some papers do discuss the coefficients but only with respect to their signs, while remaining mute on their sizes. Less than half (46.7%) of the papers score a "yes" for the practice in the 1980s. In this regard as well the 1990s were worse, with only 19% of papers avoiding sign econometrics.

vii) Does the paper clearly distinguish between statistical significance and substantive or economic significance?

Some papers interpret statistical significance to mean substantive or economic significance, or they do not specify that they referred only to statistical significance and not to substantive significance. Only 41.2% of the papers scored a "yes" for the sound practice in the 1980s. At 37.2%, the result was similar for the 1990s.

viii) Does the paper discuss explicitly the sizes of the coefficients?

The papers with a sound understanding of the difference between economic versus statistical significance discuss the coefficients in elasticity form, including their signs and size (answering the question: How large is the effect?). 80.2% of the papers scored a "yes" for this practice. At 78.1%, the result for the 1990s was comparable.

The AER results indicate that while in some areas there were improvements in 1990s, in other areas things got really worse, for example with respect to asterisks and sign econometrics.

In providing one of better examples of good practice in the 1980s about an interpretation of the coefficient in terms of responsiveness, McCloskey and Ziliak (1996:103) cite Christina Romer (1986:327) saying that "indeed, correcting for inventory movements reduces discrepancy ... by approximately half. This suggests that inventory movements are [economically] important." On the other hand, Boissiere, Knight and Sabot (1985:1026) do not perform as well by stating that "in both countries, cognitive achievement bears a highly significant relationship to educational level .... In Kenya, secondary education raises H by 11.75 points, or by 35% of the mean." Firstly, they rely on the p-value to conclude that cognitive achievement is significant, and secondly they do not interpret the coefficient in the way that is most useful for education policy. They do not attend to the more

policy-oriented question which asks: By what percentage does, say, four years of secondary schooling improve performance in cognitive tests? And a related question: Is this above or below social expectation?

Other poor interpretations that McCloskey and Ziliak (1996) cite from various researchers include phrases such as "significantly positive" and "almost significantly positive". Without a defined economic context for interpreting responses between variables of interest, a combination of asterisks and sign econometrics must remain suspect. The following section presents similar but even less reliable findings from papers presented at the 2010 AAAE conference.

3. NEW FINDINGS FROM THE AAAE 2010 CONFERENCE

For the 2010 AAAE conference more than 200 papers were received and reviewed for acceptance by at least two peers. Of the accepted papers, 104 were handed out in electronic format as part of the conference proceedings. Fifty-one of these were found to have used some kind of regression analysis and were therefore selected for evaluation using eight questions selected from the McCloskey and Ziliak (1996) list. Each paper was coded according to the area of origin of the first author: continent and country.[10] A "yes=1" or "no=2" or "3=not applicable" was assigned as practices performed in each of the papers to each question. An overall impression of how the papers performed per question was tabulated.

The findings were mixed. For some questions, the AAAE papers performed better than the AER papers, but in many more cases the performances were worse than those found in the AER papers. For example, while only 32.4% of AER presented units and descriptive statistics, 78% of AAAE papers did this correctly. On the other hand, while 41% of AER papers distinguished clearly between statistical and economic significance, only 10% of the AAAE papers did this. In Table 1 the performances per question for the AAAE papers are presented.

Table 1: Performance of AAAE papers

Selected survey question	No. to which question is applicable	% "yes"	AER 1980s
Does the paper use a small sample, such that statistical significance does not increase to acceptable levels by merely increasing the sample size?	45	82	85.7
Are the units and descriptive statistics for all regression variables included?	49	78	32.4

Are coefficients reported in elasticity form, or in some interpretable form relevant to the research question and consistent with economic theory?	48	43	66.5
Does the paper avoid presenting statistical significance as the scientific crescendo of the paper, or the only important criterion of importance?	45	29	47.3
Does the paper avoid ranking the coefficients according to their associated statistical significance values (asterisks econometrics)?	48	21	74.7
Does the paper avoid discussing only the signs of the coefficients, without discussing their sizes (sign econometrics)?	46	31	46.7
Does the paper clearly distinguish between statistical significance and substantive or economic significance?	46	10	41.2
Does the paper discuss explicitly the sizes of the coefficients?	48	27	80.2

The table clearly illustrates the poor performance of AAAE papers. It is also clear that the performance there was less credible than for the AER papers. In the evaluation of the AAAE papers, only two of the eight questions performed more than 50% of what would be considered sound practice, while the figure for the AER papers was four out of eight. 90% of the AAAE papers did not distinguish between statistical and economic significance. This could be the basis for why many of them also practised asterisks (79%) and sign (69%) econometrics. These unsound practices were identifiable even though 82% of the papers reported good sampling procedures.

One of the examples of sound practice by AAAE papers included authors A1 who, when exploring the effects of innovation on efficiency in Nigeria's maize sector, also explored the question of "How big is big?" by discussing elasticity values but apparently did not confuse the sizes of impacts with statistical significance values. They discussed all the coefficients in some elasticity format that would be easy to use in policy questions. They reported, for example, that "the cost elasticity of 0.74 … implies that a 10% increase in maize output results in a 7.4% increase in total cost ... The estimate of the variance parameter γ is 0.83 and significant at 1%, implying that 83% of the total variation in output is due to inefficiency." This was sound reporting and sound interpretation of statistical results compared, for instance, with papers by A2 authors.

A2 authors provided a clear example of how statistical significance was often confused with impact of one variable over another (i.e. substantive significance). Using a model exploring the effects of age, household and farm sizes, livestock, on the quantity of organic materials used (as opposed to chemical fertilizers), they

concluded that "age is significantly associated with the adoption of certain farm innovations", instead of reporting that the age coefficient as an estimator was only statistically significant as indicated by the t-value. Nowhere in their discussion was there any distinction made between statistical and substantive significance. The importance of parameters was ordered using asterisk symbols associated with p-values but without ever commenting on the responsiveness information contained in the sizes of coefficients: a clear case of asterisks econometrics.

Another case of sign and asterisks econometrics was performed by authors A3, who when discussing the effects of adding value to honey, reported that "the results show that the decision to add value is *positively and significantly* influenced by the amount of honey harvested … while it is *negatively influenced* (by) the age of the farmers and the education level of the household head". These findings on significantly positive and negative influences of estimators were asserted without reporting on the magnitudes of coefficients.

A4 authors not only performed asterisks econometrics but, surprisingly, also self reported as having pre-tested to drop or include variables[11] in their models based on p-values. They reported that "all the squared variables and interaction terms were used, but most of them were dropped because they were not statistically significant at the 10% level and did not improve the model".

These reports show that after almost thirty years of cautioning against the misuse of the tests of significance, many researchers continue to perform unsound practices, with all the implied economic and social cost from inaccurate and often wrong conclusions recommended to policy makers. While in certain aspects the AER practices improved across the two decades, they also deteriorated in other instances, especially in asterisks and sign econometrics. With the exceptions of sound sampling procedures and presentation of descriptive statistics for all variables in equations, in the AAAE papers the practices were poorer with respect to almost all the eight questions discussed.

4. FUTURE RESEARCH

Here, though, a caveat is necessary: the assumption made in the paper is that the evaluation criteria used by McCloskey and Ziliak (1996 and 2004) were identical to the evaluation criteria used in this note. This assumption is debatable. While the same evaluation questions were used, the application of each question within the evaluation is naturally susceptible to some degree of subjectivity, hence examples on final judgment have been provided. The comparisons of the results across the evaluations should be interpreted with this caution in mind. Nonetheless, what the present evaluation illustrates is that broadly similar interpretative problems to those found in the AER persist among agricultural economists in Africa, and many

calls over the years to check and correct these practices, it seems, have also been disregarded.

In order to create awareness and reduce these problems in future research, we propose three simple interventions.

(a) Reviewers should first check that researchers are consistent about the variable relationships that are being explored; that these should be based on stated theory or research aims; and that for those relationships, researchers should focus on the t-tests and not on other auxiliary variables, unless those were specified beforehand as variables of interest.
(b) Variables should not be ordered as important for policy merely as a result of their being based on t-tests, unless the research question is explicitly about exploring statistically significant variables. The aims must determine the statistical analysis and conclusions must flow from these.
(c) To eliminate asterisk and sign econometrics – which stem often from conflating statistical and economic significance – an adapted list of questions similar to those used in this paper should be provided to reviewers for use in evaluating papers on these issues.

NOTES

1 Fifty papers that used regression models and were presented at the 2010 conference of the African Association of Agricultural Economists (AAAE) and Agricultural Economists Association of South Africa (AEASA).

2 Absolute value.

3 For PPP tests the null hypothesis is usually that $\beta = 1.0$ (i.e. prices at home rise at exactly the same rate as abroad). However, if the tests yield an unbiased/reliable estimator $\beta = 0.999$, this means that prices at home rise at nearly the same rate as abroad. One cannot reject the null for this small difference even if the difference is permanent or statistically significant based on $t > 1.96$. That would be a Type I error – rejecting a true hypothesis.

4 $\mathbf{h}(E) = \beta_0 + \beta_1 S + \beta_2 X + \beta_3 X^2 + \beta_4 R_1 + \beta_5 R_2 + \beta_6 R_3 + \beta_7 G + \beta_8 T$ where ß is years of schooling, X is years of experience, R is race dummies, G is a gender dummy and T is a rank dummy.

5 i.e. Presentation of significance first and then of responsiveness.

6 For this equation, when the sample mean and sample size increase, the numerator increases, which increases the t-value and level of significance.

7 See a discussion on data mining by Du Plessis (2005). He also argued, like McCloskey and Ziliak (1996), that the prevalence of both data mining and misinterpretation of statistical significance tests has increased because of the ease with which computer software allows pre-testing and exclusion of variables that do not have high p-values.

8 Failing to reject a false hypothesis.

9 Less than 5% of AAAE papers reported anything that could be interpreted as pre-testing.

10 For example, a code A-Nig-01 was assigned to the first paper found that was written by a primary author affiliated to an organisation based in Africa and in Nigeria, but here we report evidence from papers only as A01 or A02.

11 Even though this practice is common, it is still rare and almost naïve to find researchers reporting on it.

REFERENCES

Arrow, K. 1959. Decision theory and the choice of a level of significance for the t-Test. In Olkiin, I. (ed.), *Contributions to probability of statistics: Essays in honor of Harold Hotelling*. Stanford: Stanford University Press.

Aye, G.C. and Mungatana, E.D. 2010. Technological innovation and efficiency in the Nigerian maize sector: Parametric stochastic and non-parametric distance function approaches. *AAAE/AEASA 2010 Conference Proceedings*. Cape Town.

Berem, R.M. and Obare, G.O. 2010. Is value addition in honey a panacea for poverty reduction in the Asal in Africa? Empirical evidence from Baringo District, Kenya. *AAAE/AEASA 2010 Conference Proceedings*. Cape Town.

Boissiere, J.R., Knight, J.B and Sabot, R.H. 1985. Earnings, schooling, ability, and cognitive Skills. *American Economic Review* 75(5):1016-1030.

DeGroot, M.H. (1975). *Probability and statistics*. Reading MA: Addison-Wesley.

Du Plessis, S. (2006). The miracle of the Septuagint and the promise of data mining in economics. *Stellenbosch Economic Working Papers: 15/06*. Stellenbosch: Stellenbosch University.

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

Gosset, W. S. (1907*). The Probable Error of a Mean (by a student). *Biometrika* 5(93): 351–360.

Gustafsson, M. & Patel, F. 2009. Managing the teacher pay system: What the local and international data are telling us. Stellenbosch: Stellenbosch University. Available from http://ideas.repec.org/p/sza/wpaper/wpapers99.html (Accessed June 2010).

Hogben, L.T. 1968. *Statistical theory: The relationship probability, credibility, and error*. New York: Norton.

McCloskey, D.N. 1985. The loss function has been mislaid: The rhetoric of significance tests. *American Economic Review* 75(2):201–205.

McCloskey, D.N. 1992. Other things equal: The bankruptcy of statistical significance. *Eastern Economic Journal*18(3):359–361.

McCloskey, D.N. 1995. The analytical economist: The insignificance of statistical significance. *The Scientific American* (April):32–33.

McCloskey, D.N. 2009. Two vices: Proof and significance. Paper presented at the American Economic Association (AEA) meeting. Chicago 3 January.

McCloskey, D.N. and Zecher, J.R. 1984. The success of purchasing power parity. In Bordo, M and Schwarz, J. (eds.), *A retrospective on the classical gold standard (1821-1931)*. Chicago: Chicago University Press.

McCloskey, D.N. and Ziliak, S.T. 1996. The standard error of regressions. *Journal of Economic Literature* XXXIV (March):97–114.

McCloskey, D.N. and Ziliak, S.T. 2004. Size matters: the standard error of regressions in the *American Economic Review. The Journal of Socio-Economics* 33 (2004):527–546.

Natrass, N. 2005. A practical guide to regression analysis for cross-sectional surveys using Stata. Centre for Social Science Research (CSSR). Cape Town: University of Cape Town (UCT). Cape Town.

Parcel, J.L., Kastens, T.L., Dhuyvetter, K.C. and Schroeder, T.C. (2000). Agricultural economists' effectiveness in reporting and conveying research procedures and results. *Agricultural and Resource Economics Review* 29(2):173–182.

Romer, C. D. 1986. Is the stabilisation of the postwar economy a figment of the data? *American Economic Review* 76(3):314–334.

Wallace, T.D. (1977). Pretest estimation in regression: A survey. *American Journal of Agricultural Economics* 59:431–443.