Preliminary (01/2016)

Assessing Measurement Errors in the R&D-Innovation-Productivity Relationship

Jacques Mairesse[a], Stéphane Robin[b]

Abstract :

The CDM econometric model (after Crépon, Duguet and Mairesse, 1998) provides a structural framework to estimate the relationship between productivity, innovation and R&D. This successful multiple-equation model is generally estimated on cross-sectional Innovation Survey-type data. Some recent contributions to the CDM literature suggest that, in this type of data, the variables measuring innovation output (namely, the share of innovative sales over the last three years) and innovation input (namely, R&D expenditures at the time of the survey) are imprecisely measured. This leads to a classical error in variables (CEV) problem that may downward bias the estimates of the key parameters of the model.

This paper proposes an assessment of the magnitude of these measurement errors, in three equations that lie at the root of the CDM model: (i) the extended production function à la *Griliches, (ii) the CDM-type productivity equation and (iii) the innovation production function. We perform our assessment on a panel of three waves of the French Innovation Survey (CIS3, CIS4 and CIS 2008), using the share of innovative sales to measure innovation output and R&D expenditures to measure input.*

For each equation and innovation variable involved, we derive an estimate of the attenuation bias from a comparison of four panel estimators. In the productivity equations, we find the attenuation bias entailed by measurement errors to be more important in the innovation output variable than in the innovation input variable. This finding is consistent with the concerns expressed in the recent literature. Further analyses suggest that instrumenting the innovation output variable with the R&D variable in a simplified CDM framework yield a significantly positive estimate of the effect of innovation on productivity. Thus, instrumenting the innovation variable in the usual CDM framework may partially correct for measurement errors, in addition to controlling for endogeneity. Further research is needed to better disentangle these potential effects of instrumentation.

JEL Codes:

[a] CREST-ENSAE and UNU-MERIT
[b] PRISM-Sorbonne – University of Paris 1 Panthéon-Sorbonne and BETA – University of Strasbourg

1. Introduction

The econometric model known as CDM (after Crépon, Duguet and Mairesse, 1998) provides a structural framework to estimate the relationships between productivity and innovation output on the one hand, and innovation output and R&D on the other. This multiple-equation model is generally estimated on cross-sectional Innovation Survey-type data, i.e. on survey data that periodically collects information on innovation-related activities following the guidelines of the OECD "Oslo Manual". Some recent contributions to the CDM literature suggest that, in this type of data, the variables measuring innovation output (namely, the share of innovative sales over the last three years) and innovation input (namely, R&D expenditures at the time of the survey) appear to be imprecisely measured. This leads to a classical error in variables (CEV) problem that may downward bias the estimates of the key parameters of the model, i.e. those associated with the innovation variables. In principle, a proper control for selectivity and endogeneity takes care also of this attenuation bias, but assessing to what extent it actually does so is an interesting issue.

This paper proposes an assessment of the magnitude of the above-mentioned measurement errors, in three equations that lie at the root of the CDM model: (i) the "extended production function" à la Griliches (linking productivity to R&D), (ii) a productivity equation linking productivity to innovation output and (iii) the "innovation production function", linking innovation output to R&D. To perform our assessment, we use a panel of three waves of the French Community Innovation Survey (CIS3, CIS4 and CIS 2008) and focus on the share of innovative sales as our measure of innovation output and on R&D expenditures as our measure of R&D.

For each equation and each innovation variable involved, we derive a synthetic measure of the attenuation bias from a comparison of four panel estimators: Pooled OLS, Between, First Differences and Between in Differences. We find that, in the productivity equations, the attenuation bias entailed by measurement errors seems to be more important (especially in terms of overall significance) in the innovation output variable (share of innovative sales) than in the innovation input variable (R&D intensity). This finding is consistent with the intuitions expressed in some recent contributions to the CDM literature (e.g., Mairesse, Mohnen and Kremp, 2005). Further analyses suggest that instrumenting the innovation output variable with the R&D variable in a very simplified CDM-type framework yield a significantly positive estimate of the effect of innovation output on labour productivity. This suggests that instrumenting the innovation variable may indeed lessen measurement errors, without necessarily cancelling them totally. Further research is thus needed to better disentangle the effect of measurement errors in a CDM-type framework that simultaneously corrects for endogeneity and selectivity.

The paper is organized as follows: based on evidence drawn from the literature on CDM and on Innovation Surveys, Section 2 presents the issue of measurement errors in innovation variables and provides a reminder of the classical errors in variables problem. Section 3 briefly describes the data we use to address this problem in the context of innovation variables. We detail our methodology in Section 4, and present our results in Section 5. We conclude in a final section.

2. Measurement errors in CIS data and the CEV problem

2.1. A short review of the related literature

This paper is part of the rather literature that explores and extends the approach initially proposed by Crépon *et al.* (1998), who developed a 'structural' model linking R&D, innovation and productivity. This model, usually referred to as the CDM model, has enjoyed a certain amount of success in the literature over the recent years. CDM-type models are generally built as three-stage econometric models that relate productivity to new knowledge, which depends on firms' R&D effort, which is in turn determined by a number of firm- and environment-specific factors. CDM-type models

are generally estimated using sequential or simultaneous estimators (such as Asymptotic Least Squares or Likelihood-based estimators) offering controls for both selectivity and endogeneity biases.

Recent applications, extensions and re-examinations of CDM include Benavente (2006), who applies the original framework to original Chilean data and Griffith *et al.* (2006), who attempt to extend the CDM framework by taking into account process as well as product innovation. Griffith *et al.* (2006) estimate their variant of the CDM Model in four EU countries (France, Germany, Spain and the U.K.). Chudnosky *et al.* (2006) estimate a similar model on Argentinian firm-level data. In addition, they propose a short summary of CDM-type studies implemented between 1998 and 2006 (Chudnosky *et al.*, 2006, Table A.1, page 284). Raffo *et al.* (2008) compare three European countries and three Latin American countries (France, Spain and Switzerland versus Argentina, Brazil and Mexico). Mairesse and Robin (2012) examine how relying on different estimators of the innovation production function affect the results of a CDM-type model estimated on two samples of French manufacturing and services firms.

An important aspect of the analysis conducted in Mairesse and Robin (2012) is that it points out to some difficulties pertaining to the innovation output variables in the French CIS data (and, quite possibly, in other CIS-type data). While these variables certainly do provide a proxy of innovation output, what they exactly measure remains to a certain extent imprecise, or rather vague. Indeed, the authors state in their conclusion: *"*[A] *careful examination, through sensitivity analyses, suggests that our indicators of product and process innovation both account for "overall" innovation, especially in the services industry. When a given type of innovation is singled out, it has a positive effect on labour productivity in all periods and in all samples. Similarly, when an indicator of overall innovation is built, its effect on labour productivity is always positive, in all industries. Thus, disentangling the effects of product and process innovation may sometimes be more difficult than is usually thought" (p. 154).*

This imprecision in measurement can be linked to the findings of a previous contribution by Mairesse *et al.* (2005). This contribution suggests that, in the French CIS data, the variables measuring innovation output (either the share of innovative sales over the last three years or an indicator of product innovation) and innovation input (either R&D intensity at the time of the survey or an R&D-doing indicator) all appear to be imprecisely measured. According to this study, the necessity to instrument innovation and R&D variables in CDM-type models reveals *"important measurement errors in the innovation intensity variables, and to a lesser extent in the innovation binary indicators as well as in the R&D intensity* [variable] *and the R&D-doing indicator"* (p. 521). The study definitely acknowledges that CDM-type models are an important step on the path to understanding how firms R&D and innovation activities relate to economic performance, and it does not deny that innovation surveys do contain interesting information to apply the CDM framework. However, the authors state in the final line of their conclusion that *"much more remains to be done to improve the quality of the data and relevance of the analysis"*.

What is at stake here is the accuracy and relevance of the econometric analysis itself. Mairesse *et al.* (2005) observe that "*in nearly all cases, (...) failure to account for the endogeneity of product innovation output leads to strong downward biases in the productivity estimates"* (p. 516). The authors argue, quite reasonably, that if the endogeneity was caused by a simultaneity problem, *"one should observe an upward bias on the expectation that the errors in the innovation intensity and productivity equations would be positively correlated (i.e. that more innovative firms would be more productive, "other things equal")"* (Ibid.). The authors interpret the observed downward biases as the result of a Classical Error in Variables (CEV) problem pertaining to measurement errors in the product innovation variables: "*What exactly is a product innovation, when is a product improvement sufficient to qualify as an innovation, what exactly is the proportion in turnover due to products introduced in the last three years that are new to the firm, new to the market or protected by patents, all these questions leave a room for different evaluations depending on the perceptions of the individual answering the survey questions. It is thus likely that there is fuzziness and noise in the*

innovation data, and thus an errors-in-variable problem in the innovation intensity and occurrence variables, which yields downwards biased coefficients" (Ibid.). The case for CEV also extends to R&D variables: *"(...) [There is] a similar, though weaker, evidence of a small downward bias in the estimates of the R&D direct elasticities (...) when we fail to account for endogeneity. We can thus infer that (...) R&D figures are not free of some random measurement errors but that they are of much better quality than (...) self-reported innovation output indicators"* (Ibid.).

Taking the CEV problem seriously, our paper proposes an assessment of the magnitude of the above-mentioned measurement errors, in three equations that lie at the root of the CDM model: (i) the "extended production function" à la Griliches (linking productivity to R&D), (ii) a productivity equation linking productivity to innovation output and (iii) the "innovation production function", linking innovation output to R&D. In particular, we want to examine whether these measurement errors are more important in innovation output variables than in R&D variables, as suggested in *Mairesse et al.* (2005). Before presenting our data and methodology, we briefly remind the reader of the CEV problem, and of how one can make an assessment of its importance.

2.2. The CEV problem

Although it is a staple of most econometric textbooks, the CEV problem is largely overlooked in many contemporary applied econometric studies. A CEV problem occurs when at least one of the regressors (or "explanatory variables") of an econometric model is affected by measurement errors. Let us call x such an explanatory variable. Since x is affected by measurement errors, then it can be rewritten as:

(1) $\quad x = x^* + e,$

where x^* denotes the true value of the variable and e denotes the measurement error. In the standard linear regression model estimated by OLS (to clarify subsequent notations, let us write it $y = \beta_0 + \beta_1 x + \varepsilon$), the CEV results in an attenuation bias measured by parameter λ. This parameter is actually a synthetic indicator of the importance of measurement errors in an explanatory variable. In the standard linear regression model, it is given by:

(2.a) $\lambda = \dfrac{\sigma_e^2}{\sigma_x^2}$ in a simple regression where x is affected by a CEV

(2.b) $\lambda = \dfrac{\sigma_e^2}{\sigma_r^2}$ in a multiple regression where a regressor is affected by a CEV

In Equations (2.a) and (2.b), σ^2_e denotes the variance of measurement error e, σ^2_x denotes the variance of the imprecisely measured variable x and σ^2_r is the variance of r (the error in the linear projection of all regressors on x^*, the unobserved 'true' regressor).

If we keep to the notation given in Equation (1), where x^* denotes the true value of x, the CEV is often presented in econometric manuals (see Wooldridge, 2002, for instance) as:

(3) $\quad \hat{\beta}_1 \xrightarrow{p} \dfrac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_e} \beta_1$

with the ^ denoting an estimated parameter, and with $\sigma^2_{x^*}$ denoting the (unobserved) variance of the 'true' variable x^*). However, for practical implementation, we have to rewrite Equation (3) as follows. First, under the usual assumption that $corr(x^*, e) = 0$, we have:

(4) $\quad \sigma^2_x = \sigma^2_{x^*} + \sigma^2_e \quad \Leftrightarrow \quad \sigma^2_{x^*} = \sigma^2_x - \sigma^2_e$

Replacing the second equality of Equation (4) in Equation (3) leads to:

$$(5)\ \hat{\beta}_1 \xrightarrow{p} \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2}\beta_1 = \left(1 - \frac{\sigma_e^2}{\sigma_x^2}\right)\beta_1 = (1-\lambda)\beta_1$$

It is the latter expression of the CEV, in terms of the λ parameter, that we will use in our empirical implementation (see Section 4 and in particular Sub-Section 4.1). This empirical implementation consists in finding an estimate of the λ parameter.

Because measurement error e is unknown, it is generally impossible to compute λ when estimating, say, $y = \beta_0 + \beta_1 x + \varepsilon = \beta_0 + \beta_1 (x^* + e) + \varepsilon$ on cross-sectional data. Mairesse (1990) explains how this becomes possible when conducting estimations on panel data. Because several (related) estimators are available for panel data, it is possible to derive a value of λ from a comparison of a pair of these estimators. This is exactly what we do in the present paper. To obtain the panel data we need, we merge several waves of the French CIS survey. This survey and the merging process are described in the next section.

3. Data: building a panel using the French Community Innovation Survey

The present study uses firm-level data from the third, fourth and fifth waves (CIS3, CIS4 and CIS 2008, respectively) of the French component of the Community Innovation Survey (hereafter, CIS). The CIS is a harmonized survey that is carried out by national statistical agencies in all EU Member States under the co-ordination of Eurostat. The survey has been conducted every four year since 1992[1] and contains cross-sectional information pertaining to the year of the survey, with some retrospective information concerning the previous two years. There also exists, at least in France, a prototype CIS conducted in 1990 and known as CIS 0, which we will use the present study to examine innovation persistence and initial conditions as part of additional analyses.

CIS3 was conducted in 2001, CIS4 in 2005 and CIS 2008 in 2008 as the name suggests (from CIS 2008 onwards, CIS are referenced to by the year in which they were conducted). CIS 4 provides information for the period 1998-2000, CIS 4 covers 2002-2004 and CIS 2008 covers 2006-2008. All three waves of the survey provide information on firms' R&D activities, sources of knowledge, intellectual property protection, product and process innovations, other forms of innovation (e.g., organizational changes) and abandoned innovations. Since all three waves share the same core questionnaire with the same coding, it is comparatively easy to build a panel containing at least the most important variables (e.g., productivity, firm size, and measures of innovation input and output).

However, there are some key differences across waves. First, CIS4 samples firms with 10 employees or more, whereas CIS3 only included firms with 20 employees or more. Second, CIS3 focused mostly on manufacturing firms, whereas CIS4 covers the services industry quite extensively. Moreover, some information about firms' investment in physical capital is available in CIS 3, but not in CIS 4 and CIS 2008, which implies that we will ultimately have to obtain a measure of physical capital from administrative sources. Similarly, CIS 2008 contains some additional questions specifically dedicated to environmental innovation, which we will not use in the present study.

To conduct our empirical analysis, we merged the prototype CIS 0 with CIS 3, CIS4 and CIS 2008. Table 1 present summary statistics on key variables for each wave. The main rationale for including CIS 0 as part of the merging process is that it can provide useful information on initial conditions for some innovation variables. Although one may fear that including CIS 0 leads to severe attrition in the resulting panel, further examination of the data suggests that it this not necessarily the

[1] In addition, "intermediate" CIS (CIS 2006 and CIS 2010) have been conducted, but they generally involve a smaller number of firms and/or a lighter questionnaire than the "regular" CIS.

case. In fact, the sheer number of observations in CIS 0 (24340 firms) considerably lessens the attrition problem, and it appears that our panel loses more firms from the inclusion of CIS 3 (which has only 7016 observations) than from the inclusion of CIS 0. Moreover, we observe surprisingly strong innovation persistence from CIS 0 to the other three waves that we consider in our study, as evidenced in Table 2.

TABLE 1 ABOUT HERE

TABLE 2 ABOUT HERE

To build our panel, we apply a simple selection rule. We keep firms that are observed in at least two subsequent waves of the CIS and in CIS 0. By doing so, we can obtain various balanced and unbalanced panels. Focusing on balanced panels (centered, by construction, on surviving firms that are strongly inclined to innovate), there are (*i*) 1059 firms observed in 1990, 2000 *and* 2004; (*ii*) 1027 firms observed in 1990, 2004 *and* 2008; (*iii*) 1171 firms observed in 2000, 2004 and 2008; (*iv*) 579 firms observed in 1990, 2000, 2004 *and* 2008. For the analysis conducted in this paper, we focused on the balanced panel of 1171 firms observed in 2000, 2004 and 2008 according to our selection rule.

4. Methodology

4.1. Estimation of the attenuation bias

Our empirical analysis consists in estimating the attenuation bias λ (see Sub-Section 2.2.) to assess the extent to which key innovation and R&D variables in the CDM model suffer from measurement error. This econometric analysis is primarily exploratory and relies on the estimation of a few simple models on the French CIS panel described in Section 3. We first perform three simple regressions using only variables taken from the CIS, matched with R&D series taken from French yearly R&D surveys. To perform our analysis, we use the balanced panel of 1171 firms observed in 2000, 2004 and 2008 described in Section 3. In order to keep the interpretation and discussion of the results tractable, we focus here on continuous measures of innovation[2]: R&D intensity (defined as the ratio of R&D expenditures to firm size) as a measure of innovation input, and the share of innovative sales as a measure of innovation output[3]. The former is measured at the year of the survey, whereas the latter is measured over the last three years.

Our benchmark models are three simple (panel data) linear regressions estimated on the selected sub-samples of innovative firms (i.e. R&D doing firms or product innovators): (1) a productivity equation derived from the "extended production function" à la Griliches and linking labour productivity (LP) to R&D intensity; (2) a productivity equation linking LP to innovation output (i.e., share of innovative sales in total sales) and (3) the "innovation production function", linking innovation output (share of innovative sales) to R&D intensity. To put it in a nutshell, we thus estimate:

(Model 1) $\ln LP_{it} = \beta_0 + \beta_1 \ln(\text{R\&D intensity})_{it} + \varepsilon_{it}$

[2] CIS surveys also provide a binary indicator of the introduction of an innovation, and allow researchers to build a dummy variable indicating whether a firm performs R&D on a continuous basis or not. However, Mairesse Mohnen and Kremp (2005) find that these dummy variables are likely to be less biased by measurement errors than the continuous measures in which we focus here. Moreover, continuous measures were used in the original CDM model, and tend to be preferred by researchers, especially as far as R&D is concerned.

[3] More precisely, so as to have a continuous variable ranging from $-\infty$ to $+\infty$, we use the *Logit transform* of the share of innovative sales, denoted $z = \ln[x/(1-x)]$ when x denotes the share of innovative sales.

(Model 2) $\ln LP_{it} = \beta_0 + \beta_1(\text{logit-share of innovative sales})_{it} + \varepsilon_{it}$

(Model 3) $\text{Logit-share of innovative sales}_{it} = \beta_0 + \beta_1 \ln(\text{R\&D intensity})_{it} + \varepsilon_{it}$.

Note that in Models 1 to 3, the time unit *t* is the four-year period that separates each wave of the CIS.

Model (1), which links productivity to innovation *input* (R&D intensity) is both theoretically grounded and easy to implement when one lacks data on innovation *output*. However, when appropriate data is available (as is the case with CIS data), one can link (*i*) productivity to innovation *output* (here, the share of innovative sales), and (*ii*) innovation *output* to innovation *input*. The former is an alternative productivity equation and the latter is the so-called innovation production function. Taken together, these two equations are at the roots of the CDM framework. Thus, our Model (1) is a kind of "reduced-form" CDM model, whereas Models (2) and (3) are the key equations at the heart of the original CDM model. This is why focusing on these equations is relevant for the task at hand.

After this first analysis, we add control variables in Models 1 to 3, which allows us to explore two different paths. First, we add a single control variable from CIS 0, namely the share of innovative sales in 1990, to control for "initial conditions" (and for the importance of innovation persistence) in ours models. This leads us to perform multiple linear regressions on a balanced panel of 579 firms. The second path we follow consists in estimating the "production functions" described by Models 1 and 2 with appropriate controls for capital, labour and materials. This implies matching our panel with external administrative sources to get measures of Value Added (VA), materials and physical capital. Labour is measured by the number of employees (firm size), a variable that is readily available in the CIS. We also use this variable as a control in the innovation production function (Model 3).

4.1.1. Simple regressions

In order to derive a value of λ, we estimate each of our benchmark models using four different panel estimators: (i) Pooled OLS in levels, (ii) Between in levels, (iii) First Differences and (iv) Between in Differences.[4] To calculate λ, we compare these four estimators by pairs: Pooled OLS in levels with Between in levels, and First Differences with Between in Differences. Each comparison yields two values of σ^2_e and thus two estimations of λ. We now sketch the formulas for these four values of λ when models (1) to (3) are simple regressions (without control variables).

Let β_1 be the true parameter of interest in a simple regression (e.g., in Models 1 to 3). We start by comparing Pooled OLS estimates with Between estimates. Following Mairesse (1990), let TL denote the Pooled OLS estimator in levels, and BL the Between estimator in levels. Let β_{TL} and β_{BL} denote the estimates of β_1 by TL and BL, respectively. Because of the attenuation bias due to the CEV, we can write:

$$\beta_{TL} \to (1 - \lambda_{TL})\beta_1 \qquad \text{and} \qquad \beta_{BL} \to (1 - \lambda_{BL})\beta_1$$

which, in large samples, leads to:

$$(6) \qquad \frac{\beta_{TL}}{1-\lambda_{TL}} = \frac{\beta_{BL}}{1-\lambda_{BL}}.$$

[4] We leave aside the Within (fixed-effect) estimator for two reasons. The first is that, when examining productivity – innovation relationship, fixed-effect may take out too much variability (see Hall and Mairesse, 1995) as well as innovation persistence, which in our panel is likely to be high as stated in Section 3. The second reason is that an individual fixed effect may actually exacerbate the effect of measurement errors in the innovation variables affected by a CEV (Bound, Brown and Mathiowetz, 2001).

Using $\sigma^2_{e\,BL} = \frac{1}{T}\sigma^2_{e\,TL}$ (see Mairesse, 1990), we derive from Equation (6) the formulae for the variance of the measurement error and for λ_{TL} (see Appendix 1 for details):

(7.a) $$\sigma^2_{e\,TL} = \frac{\beta_{TL} - \beta_{BL}}{\frac{\beta_{TL}}{T\sigma^2_{x\,BL}} - \frac{\beta_{BL}}{\sigma^2_{x\,TL}}}$$

(7.b) $$\lambda_{TL} = \frac{\sigma^2_{e\,TL}}{\sigma^2_{x\,TL}} = \frac{\beta_{TL} - \beta_{BL}}{\sigma^2_{x\,TL}\left(\frac{\beta_{TL}}{T\sigma^2_{x\,BL}} - \frac{\beta_{BL}}{\sigma^2_{x\,TL}}\right)}$$

where the β's are estimated parameters (by TL or BL) and all other parameters are constant terms (T is sample time length and the σ^2's are sample statistics).

Of course, since $\sigma^2_{e\,BL} = \frac{1}{T}\sigma^2_{e\,TL}$, once $\sigma^2_{e\,TL}$ has been estimated using (8), it also yields:

(8.a) $$\sigma^2_{e\,BL} = \frac{1}{T}\sigma^2_{e\,TL} = \frac{\beta_{TL} - \beta_{BL}}{\frac{\beta_{TL}}{\sigma^2_{x\,BL}} - \frac{T\beta_{BL}}{\sigma^2_{x\,TL}}}$$

(8.b) $$\lambda_{BL} = \frac{\sigma^2_{e\,BL}}{\sigma^2_{x\,BL}} = \frac{\beta_{TL} - \beta_{BL}}{\sigma^2_{x\,BL}\left(\frac{\beta_{TL}}{\sigma^2_{x\,BL}} - \frac{T\beta_{BL}}{\sigma^2_{x\,TL}}\right)}.$$

We now compare First Differences estimates with Between in Differences estimates. Again, following Mairesse (1990), let TD denote the First Differences estimator in levels, and BD the Between in Differences estimator. Let β_{TD} and β_{BD} denote the estimates of β_1 by TD and BD, respectively. As before, because of the attenuation bias due to the CEV, we can write:

$\beta_{TD} = (1 - \lambda_{TD})\beta_1$ and $\beta_{BD} = (1 - \lambda_{BD})\beta_1$

from which it comes that:

(9) $$\frac{\beta_{TD}}{1-\lambda_{TD}} = \frac{\beta_{BD}}{1-\lambda_{BD}}$$

Rearranging equation (12) and using $\sigma^2_{e\,BD} = \frac{2}{T-1}\sigma^2_{e\,TD}$ (see Mairesse, 1990) leads to:

(10.a) $$\sigma^2_{e\,TD} = \frac{\beta_{TD} - \beta_{BD}}{\frac{2\beta_{TD}}{(T-1)\sigma^2_{x\,BD}} - \frac{\beta_{BD}}{\sigma^2_{x\,TD}}}$$

(10.b) $$\lambda_{TD} = \frac{\sigma^2_{e\,TD}}{\sigma^2_{x\,TD}} = \frac{\beta_{TD} - \beta_{BD}}{\sigma^2_{x\,TD}\left(\frac{2\beta_{TD}}{(T-1)\sigma^2_{x\,BD}} - \frac{\beta_{BD}}{\sigma^2_{x\,TD}}\right)}$$

where the β's are estimated parameters (by TD or BD) and all other parameters are constant terms (T is sample time length and the σ^2's are sample statistics).

Again, since $\sigma^2_{e\,BD} = \frac{2}{T-1}\sigma^2_{e\,TD}$, once $\sigma^2_{e\,TD}$ has been computed using (13), it also yields:

(11.a) $$\sigma^2_{e\,BD} = \frac{2}{T-1}\sigma^2_{e\,TD} = \frac{2(\beta_{TD} - \beta_{BD})}{\frac{2\beta_{TD}}{\sigma^2_{x\,BD}} - \frac{(T-1)\beta_{BD}}{\sigma^2_{x\,TD}}} = \frac{\beta_{TD} - \beta_{BD}}{\frac{\beta_{TD}}{\sigma^2_{x\,BD}} - \frac{T-1}{2}\frac{\beta_{BD}}{\sigma^2_{x\,TD}}}$$

(11.b) $$\lambda_{BD} = \frac{\sigma^2_{e\,BD}}{\sigma^2_{x\,BD}} = \frac{\beta_{TD} - \beta_{BD}}{\sigma^2_{x\,BD}\left(\frac{\beta_{TD}}{\sigma^2_{x\,BD}} - \frac{T-1}{2}\frac{\beta_{BD}}{\sigma^2_{x\,TD}}\right)}$$

Comparing TL and BL thus yield two estimates of λ (λ_{TL} and λ_{BL}) and comparing TD and BD yield another pair of estimates (λ_{TD} and λ_{BD}). We compute standard errors for all estimates of λ using a bootstrapping procedure described in Appendix 2.

Before going any further, we have to mention a specificity of Model 2, i.e. the equation linking labour productivity to the share of innovative sales. There are two ways to estimate this model with a differences estimator (i.e., TD or BD). First, one can take the "first-difference", Δ, of both the dependent and the explanatory variables, as is done when estimating Model 1. This is written:

(TD1) $\Delta \ln LP_{it} = \beta_0 + \beta_1.\Delta\ln(\text{logit-share of innovative sales})_{it} + \varepsilon_{it}$

(BD1) $\Delta \ln LP_{i\bullet} = \beta_0 + \beta_1.\Delta\ln(\text{logit-share of innovative sales})_{i\bullet} + \varepsilon_{i\bullet}$

In the above, Δ corresponds to a 4-year jump since our time unit t is the 4-year gap between two CIS. A second possibility with Model 2, however, is to take the "first-difference" Δ of labour productivity (which again corresponds to a 4-year jump) and the value of the explanatory variable in the previous CIS (i.e. 4 years earlier). This is written:

(TD2) $\Delta \ln LP_{it} = \beta_0 + \beta_1.\ln(\text{logit-share of innovative sales})_{it\text{-}1} + \varepsilon_{it}$

(BD2) $\Delta \ln LP_{i\bullet} = \beta_0 + \beta_1.\ln(\text{logit-share of innovative sales in } t\text{-}1)_{i\bullet} + \varepsilon_{i\bullet}$

This second approach makes sense because, in CIS data, the share of innovative sales is measured over the last 3 years (e.g., 2002-2004 for CIS 4 or 1998-2000 for CIS3), which makes it a "lagged difference". This second approach cannot be applied to Model 1 because R&D expenditures (used to build R&D intensity) are only observed at the year of the survey.

4.1.2. Multiple regressions

Even with panel data, it is difficult to estimate λ when the regression of interest involves multiple regressors (i.e., control variables besides the explanatory variable of interest). We get around this problem using the Frisch-Waugh procedure, which can be described as follows:

1. Using OLS, regress the dependent variable y on a vector z including all regressors except x, our (imperfectly measured) variable of interest.
2. Predict u_1, the residuals of this first regression
3. Regress x on z using OLS
4. Predict u_2, the residuals of this second regression
5. Regress u_1 on u_2 using panel estimators to get the estimated β's associated with x and use these estimates to compute the λ's.

This procedure allows us to compute λ using the formulas given in Equations (9), (11), (14) and (16). We apply the Frisch-Waugh procedure to each benchmark model (Models 1 to 3) with additional

control variables, in order to derive four estimates of λ (λ_{TL}, λ_{BL}, λ_{TD}, λ_{BD}) for which we bootstrap standard errors.

As explained in the header to the present section, we add control variables to our benchmark models to add appropriate controls for the usual inputs (capital, labour and materials) in the "production functions" described by Models 1 and 2, and a control for firm size in the "innovation production function" described by Model 3. While the labour input or firm size, both measured by the number of employees, are readily available in the CIS, getting measures of additional inputs (namely, materials and physical capital) implies matching our initial panel with external administrative sources. Using the French yearly firm census (EAE), we are able to retrieve investment in physical capital at the end of the previous year (which we use as our measure of capital) and a measure of materials. Due to the large number of firms covered by the EAE, the matching process does not significantly bias our dataset.

4.2. Further analyses

To consolidate our main analysis, we extend it in two directions. First, we use our estimates of the attenuation bias λ to build measures of reliability for error-in-variables regressions, i.e. linear regressions that use weights to take into account the measurement error on certain key variables (Draper and Smith, 1998, pp. 89–91; Kmenta, 1997, pp. 352–357). The weights used in error-in-variables regressions are often referred to as reliability measures, and can be defined simply as:

(12) $r = 1 - \lambda$

When applying errors-in-variable regression, it is generally not possible to determine the reliability and researchers mostly use it as a tool to determine the sensitivity of OLS estimates to a CEV problem (by setting different values for r). In the present paper, the estimates of λ obtained in our different models also give us some plausible values for r. This allows us to examine how the estimated elasticity of the innovation variable evolves in a given model when we take into account the attenuation bias.

The second extension to our main analysis consists in estimating a "bare bones" CDM-type model, which simply consists in instrumenting the innovation output variable in CDM-type productivity equation, using the innovation production function. In other words, we simply "plug" the innovation production function into the CDM-type productivity equation (i.e., we plug the multiple regression versions of Model (2) into Model (3), respectively). In this model, the innovation output variable is treated as endogenous and R&D intensity is treated as a "weak instrument". The idea here is to use an IV-like approach to try and correct for the measurement error in the *Innov* variable, as an alternative to our main approach based on the comparison of panel estimators. Thus the regression parameter associated with the (instrumented) share of innovative sales in the "bare bones" CDM model can be compared to the same parameter in the multiple regression version of Model (2), after taking into account the attenuation. While our bare bones CDM is very far from an ideal IV model, it is enough to give us an indication of how the usual way to estimate CDM-type models deals with the CEV problem in the key innovation variables. Given that the productivity equation contains control variables (such as capital and materials) that are not in the innovation production function, and given that R&D intensity is a weak instrument at best, 2SLS estimates may not be very reliable. We thus also estimate our model using 3SLS.

5. Results

5.1. Simple linear regressions

The results obtained by estimating Models 1 to 3 as simple linear regressions are presented in Tables 3 to 5, respectively. Each table features in its upper panel the parameter estimate for the relevant innovation variable (R&D intensity or share of innovative sales) as well as goodness-of-fit statistics. The lower panel presents the estimates of the variance of the error σ^2_e, the sample variance of the imperfectly measured variable x (the relevant innovation variable for the model at hand) and the estimates of the λ parameter, together with its bootstrapped standard errors. Since σ^2_e, the variance of the measurement error, is calculated according to one of the formulas given by Equations (7.a), (8.a), (10.a) and (11.a), it can occur that, during a bootstrap iteration, the value of σ^2_e be negative, yielding a negative value of λ for that iteration (because the sample variance of x is always positive, as a variance should be). To take this into account, we have followed two distinct bootstrapping procedures, which are detailed in Appendix 2, and which ensure that our estimate of λ is always between 0 and 1, as it should be.

TABLE 3 ABOUT HERE

Table 3 presents the results obtained when estimating Model 1, i.e. the productivity equation derived from the extended production function, with R&D intensity as the relevant innovation variable. We observe a significantly positive association between productivity and R&D with the Pooled OLS and Between estimators, but no significant correlation with the difference estimators. Although the model fit is overall rather poor, the sign of the association between the dependent and explanatory variable is reassuringly positive whenever significant. Some readers may think this rather poor fit results from the timing of the R&D intensity variable, which in innovation surveys is measured in the same year as labour productivity. It is well known that a certain time lag has to elapse for R&D to yield positive results on innovation output, let alone on firm productivity (see for instance Hausman *et al.*, 1984). This is one of the reasons why econometric studies often use long R&D series to build an "R&D capital stock". Therefore, it might be more appropriate for our R&D intensity variable to be lagged with respect to labour productivity. While this is not possible when using CIS data only, we can obtain a lagged R&D intensity variable by matching our CIS panel with the yearly French R&D survey. However, experimenting with lagged values of R&D intensity yielded results very similar to those presented in Table 3. This is most certainly because our estimations are performed on a sub-group of firms that conduct R&D on a continuous basis. This leads us to think that, in the present study, the poor model fit is better explained by the estimators used, by the short time dimension of our panel, and by the fact that we are considering – in a first analysis – simple regressions.

Our purpose, however, is to propose estimates of the attenuation bias resulting from CEV, and this is best done at first in the context of simple regressions. Moving to the lower panel of Table 3, we find that the estimates of λ obtained with the first bootstrap procedure are always significantly different from 0, with the differenced (TD and BD) estimators leading to much higher estimates than the levels (TL and BL) estimators. Moreover, although both bootstrap procedures (see Appendix 2) yield very similar estimates, the second bootstrap procedure causes an inflation of the standard errors that makes the estimates slightly less significant. According to the levels estimators, the value of λ could be as low as 0.07 and as high as 0.17, whereas according to the differenced estimators, the value of λ could lie within a range of 0.43 to 0.58. Differenced estimators thus tend to aggravate the effect of measurement errors.

TABLE 4 ABOUT HERE

Table 4 presents the results obtained when estimating Model 2, i.e. the productivity equation linking labour productivity to the (logit-)share of innovative sales, which now becomes the relevant innovation variable. What is especially interesting here is that we have two distinct approaches available when designing differences estimators. In the present case, the first approach, denoted TD1 and BD1 in Sub-section 4.1, could be interpreted as a "long differences" estimator, as both the dependent and explanatory variables are differenced across a 4-year period. By contrast, the second approach, denoted TD2 and BD2 in Sub-Section 4.1, could be interpreted as a "short differences" estimator, as the explanatory variable (which is lagged four years with respect to the dependent variable) is aggregated over three years[5].

In Table 4, we observe a significantly positive association between productivity and the share of innovative sales with the TL and BL estimators on the one hand, and with the TD2 and BD2 estimators on the other. The TD1 and BD1 estimators (which correspond to the TD and BD estimators in Table 3) provide a non-significant or weakly significant effect, as was already the case in Table 3. Again, model fit is overall rather poor: Although the Fisher test is always significant, the R^2 suggests a poor linear fit.

The lower panel of Table 4 presents the estimate of the attenuation bias, i.e. that of λ, which is significantly different from 0 with two estimators only: TL and BD1. The former provides an estimate of λ of around 0.36 or 0.37, and the latter an estimate of around 0.50 (0.47 at the lowest and 0.51 at the highest). Again, the differenced estimator leads to a much higher estimate than the level estimator.

It is interesting to compare Table 3 and Table 4 since they present econometric results derived from a very similar theoretical model (a productivity equation), the only difference being that the explanatory variable is innovation input in one case and innovation output in the other. Overall, measurement errors seem to be slightly larger in the innovation output (share of innovative sales) variable than in the innovation input (R&D intensity) variable, which is consistent with Mairesse *et al.* (2005). As a caveat, though, we should keep in mind that the results in Tables 3 and 4 will certainly be improved when additional control variables (such as materials and the stock of physical capital) are added to the productivity equation.

TABLE 5 ABOUT HERE

Table 5 presents the results of the estimation of the innovation production function, linking the (logit-)share of innovative sales to the log-R&D intensity. In the upper panel of the table, we observe a significantly positive effect of the explanatory variable for all but the Between Differences estimator, which suggests that a higher R&D intensity is associated with a higher degree of innovation (i.e., a higher “innovativeness” or “innovativity”). As before, although the Fisher test is always significant, the R^2 suggests a rather poor linear fit.

Moving to the lower panel of Table 5, we observe that both the level (TL and BL) and the differenced (TD and BD) estimators yield estimates of λ that are significantly different from zero. Again, the attenuation bias seems to be of a higher magnitude when estimated with differenced rather than level estimators, while both bootstrapping procedure yield similar estimates. Interestingly, contrary to what we previously observed, the second bootstrapping procedure does not lead, in this model, to an inflation of the standard errors. Given the difference across the two procedures (see Appendix 2), this suggests that the formulas for the variance of the measurement error σ^2_e, given by

[5] For instance, if we consider CIS3 and CIS4, the dependent variable is differenced across 2004 and 2000, whereas the explanatory variable, observed in 2000, is aggregated over 1998-2000.

Equations (7.a), (8.a), (10.a) and (11.a), lead more systematically to consistent (i.e., positive) values. Overall, level estimators suggest that the value of λ is around 0.20 or slightly higher, whereas differenced estimators give an estimate of around 0.30 at its lowest, and of more than 0.50 at its highest.

5.2. Multiple linear regressions

As explained in Sub-Section 4.1.2, we use multiple linear regressions to control for the usual inputs (capital, labour and materials) in the production functions (Models 1 and 2), and for firm size in the innovation production function (Model 3). These are presented in Tables 6 to 8, which are all organized in the same way: The upper panel presents the results of the multiple regression estimated with each one of our four panel estimators (TL, BL, TD and BD), whereas the lower panel presents the estimates of the attenuation bias obtained by applying the Frisch-Waugh procedure to the aforementioned panel estimators. As in the simple linear regressions of Section 5.1, the standard errors of the attenuation bias are bootstrapped using the two alternative bootstrapping procedures described in Appendix 2 to ensure consistent estimates of λ (i.e., with values comprised between 0 and 1).

TABLE 6 ABOUT HERE

Table 6 presents the results of the productivity equation derived from the extended production function, which links labour productivity to R&D intensity. As can be seen in the upper part of the table, moving from a simple to a multiple linear regression with controls for the usual C, L and M inputs considerably improves the estimation results: The Fisher test is always strongly significant, and the adjusted R^2 ranges from 0.70 to 0.92 (depending on the estimator), suggesting a very good linear fit. The conventional inputs show the expected signs and are significantly different from zero with all estimators (except for the capital input, which becomes non-significant with the BD estimator). The elasticity of capital ranges from 0.01 to 0.04, that of labour from 0.22 to 0.37[6] and that of materials from 0.58 to 0.75. The elasticity of R&D intensity is consistently positive, with a value of about 0.02 to 0.03 depending on the estimator used. This is much lower than the value of about 0.11 to 0.12 obtained with simple linear regressions, which is normal since productivity is now explained by the conventional inputs in addition to R&D intensity.

Moving to the lower panel in Table 6, we see that the estimate of the attenuation bias, λ, is almost never significantly different from zero, except when we implement the levels estimators TL and BL with the first bootstrapping procedure. In this case, the value of λ falls within the range of 0.14 to 0.32, which lies within the range of 0.07 to 0.58 obtained with simple linear regressions. Comparing the results presented in Table 6 to those of Table 3 (in which λ was overall largely significant) suggests that adding appropriate controls for the usual inputs in the extended production function tends to lessen the CEV problem. This could be because the measurement error in the R&D variable is largely correlated with that in the capital stock variable, and controlling for the latter leads to reducing the error in the former. To check this hypothesis, we reapplied our methodology to the capital stock variable, keeping the same multiple linear regressions as in Table 6, but assuming no measurement error in the R&D intensity variable. The estimates of λ we then obtained for the capital stock variable were overall not significantly different from zero, which suggests that there is no

[6] As far as the labour input is concerned, Tables 6 and 7 display the estimate of β_L-1, since the equation we estimate is not $\ln Q_{it} = \beta_0 + \beta_C \ln C_{it} + \beta_L \ln L_{it} + \beta_M \ln M_{it} + \beta_K \ln K_{it} + \varepsilon_{it}$, but the productivity equation $\ln (Q_{it}/L_{it}) = \ln LP_{it} = \beta_0 + \beta_C \ln C_{it} + (\beta_L\text{-}1) \ln L_{it} + \beta_M \ln M_{it} + \beta_K \ln K_{it} + \varepsilon_{it}$, where K denotes the knowledge input (R&D in Table 6 and share of innovative sales in Table 7).

significant bias from measurement errors in the capital stock variable. This finding concurs with the above interpretation, because it suggests that the measurement error in the capital stock variable could indeed be correlated with that in the R&D variable, and that controlling for the latter leads to reducing the error in the former.

TABLE 7 ABOUT HERE

Table 7 presents the results of the CDM-type productivity equation, which links labour productivity to a measure of innovation output (share of innovative sales) rather than a measure of innovation input. As was already the case with the previous productivity equation, including controls for the conventional inputs drastically improves the model fit (compared to the simple regression case). Again, conventional inputs show the expected signs and are significantly different from zero with all estimators (except for the capital input, which becomes non-significant with both variants of the BD estimator). The elasticity of capital ranges from 0.02 to 0.03, that of labour from 0.21 to 0.30 and that of materials from 0.69 to 0.77. These estimates are reassuringly similar to those obtained in Table 6 for the same inputs. The elasticity of the innovation output variable, however, displays a slightly different picture compared to the simple linear regression case: Controlling for the conventional inputs not only lessens the estimates of this elasticity, it also makes it less significant overall. In the simple regression case, the elasticity was significantly positive when using the levels estimators (TL and BL) and the second variant of the differenced estimators (TD2 and BD2), with values ranging from 0.04 to 0.07. In the multiple regressions case, it is significantly positive only when using the TD2 and BD2 estimators, with a value of about 0.01.

The lower panel of Table 7 displays an interesting feature. It shows that λ is very significant and very high (within the range of 0.40 to 0.71), both with the levels estimators (BL and TL) and with the TD1 and BD1 differenced estimators, at least when we implement the first bootstrapping procedure. These four estimators are those for which the elasticity of the innovation output variable is not significantly different from zero. Taken together with those of the upper panel, these results suggest that the attenuation bias is important enough to make the estimates of the elasticity of the innovation variable small enough (with respect to their standard errors) to become insignificant. We are thus in a situation where multiple regression emphasizes the effect of the CEV. Such a situation, which is the exact opposite of the one encountered in the results of Table 6, may indeed occur, depending on the direction of the partial correlation between the measurement error in the innovation variable and the various control variables. To explore this interpretation further, we again reapplied our methodology to the capital stock variable, keeping the same multiple linear regressions as in Table 7 and assuming no measurement error in the innovation output variable. This time, the estimates of λ for the capital stock variable tended to be significantly different from zero, especially with the TD and BD estimators. This reinforce the above interpretation, according to which measurement errors in the innovation output and capital stock variables could mutually reinforce each other.

TABLE 8 ABOUT HERE

Lastly, Table 8 presents the results of the innovation production function, which links the share of innovative sales to R&D intensity and lies at the heart of the CDM model. Here, the added controls simply consist in a single measure of firm size (number of employees). Since this variable is overall not significantly different from zero, including it does not really improve the model fit. Moving to the lower panel, we similarly see that the estimates of the attenuation bias are quite comparable to those obtained in the simple linear regression case, especially when implementing the first bootstrapping procedure (the second one tends to inflate the estimates and their standard errors).

5.3. Further analyses

5.3.1. Errors-in-variables regressions

Table 9 presents the results of our errors-in-variables regressions for Models (1) to (3): Extended productivity equation, CDM-type productivity equation and innovation production function. The estimations were conducted on the panel we used in Sub-Section 5.2. Each model is presented in a distinct column and is estimated twice: (1) in levels, first with the reliability *r* of the explanatory innovation variable set to 1, which correspond to a pooled OLS regression (TL estimator) and then with *r* set to its TL estimated value; (2) in differences, first with *r* set to 1, which correspond to a first-difference regression (TD estimator) and then with *r* set to its TD estimated value. In addition to these estimations, Appendix 3 present cross-sectional estimates: for each wave of the CIS that we consider here (CIS3, CIS4 and CIS 2008), we estimate Models (1) to (3) thrice: (1) assuming *r* set to 1, (2) assuming *r* set to 1 and controlling for initial conditions with the share of innovative sales in CIS 0[7] (as a proxy for an individual fixed effect measuring innovation persistence) and (3) assuming *r* set to its TL estimated value.

TABLE 9 ABOUT HERE

A cursory look at Table 9 shows that, whenever the estimated parameter associated with the innovation (input or output) variable is significant in a model, correcting for the attenuation bias leads to an increase in the said parameter, which remains significant at the same level. Correcting for the attenuation bias also increases the absolute values of non-significant parameters, but without making them significant.

Looking first at the TL estimates presented in the upper panel of Table 9 reveals that, in the extended productivity equation (Model (1)), the parameter associated with the R&D variable is estimated at 0.02 (significant at the 1% level) when we assume full reliability. This baseline value exactly corresponds to the estimate obtained with the Pooled OLS (TL) estimator[8] in Table 6. With reliability r_{TL} going down to 0.68 (which corresponds to $\lambda_{TL} = 0.32$), the parameter estimate rises to 0.03 (and remains significant at the 1% level). In the CDM-type productivity equation (Model (2)), the parameter associated with the innovation variable remains non-significant throughout, although its absolute value does rise as reliability goes down to 0.29. In the innovation production function (Model (3)), the parameter associated with R&D intensity is estimated at 0.13 (significant at the 1% level) under full reliability (again, this exactly corresponds to what we obtained with the TL estimator in Table 8). The parameter increases to 0.18 (and remains significant at the 1% level) when reliability r_{TL} decreases to 0.73.

In the lower panel of Table 9, we present the estimates obtained with the TD estimator. In neither of the productivity equations is the parameter associated with the innovation variable significantly different from zero. Correcting for the attenuation bias increases the absolute value of this parameter from 0.004 to 0.007 in the extended productivity equation, and from 0.002 to 0.01 in the CDM-type productivity equation. In both equations, however, the estimate remains non-significant. In the innovation production function (Model (3)), the estimate of the parameter

[7] In CIS 0, the share of innovative sales is coded as a categorical variable (with values 1 for a share of less than 10%, 2 for a share of 10 to less than 30%, 3 for a share of 30% to less than 70% and 4 for a share of 70% or more). We introduce it in our regressions as a set of dummies, using the first category as the reference.

[8] Although close, the standard errors do not necessarily correspond, because we implemented the TL estimator with heteroskedasticity-robust standard errors, whereas Stata's eivreg procedure (which we used to estimate the errors-in-variables regression) does not allow for robust standard errors.

associated with R&D intensity rises from a baseline value of 0.11 to 0.20 as reliability goes down to $r_{TL} = 0.57$, remaining significant at the 5% level throughout.

The above-mentioned results are confirmed in all three models by the cross-sectional estimates presented in Appendix 3. What we find in this Appendix is that, overall, including a proxy for innovation persistence does not change to the baseline OLS estimates, whereas controlling for the attenuation bias does indeed lead to higher parameter values, at least when the parameter estimates are significant. This is especially the case when the innovation variable included in the regressors is a measure based on R&D. In the end, the rise in the parameter estimates associated with R&D in Model (1) and Model (3), both in panel and in cross-section, is simply another way to present the attenuation bias measured by λ. This way might nevertheless give the reader a better, more intuitive grasp of the consequences of the CEV problem at work in the CDM model. To see how the usual approach to CDM deals with this problem, we move on to present the estimates of our "bare bones" CDM model.

5.3.2. Bare bones CDM model

Table 10 gives the estimates of our "bare bones" CDM model, which simply consists in the innovation production function (Stage 1) plugged into a CDM-type labour productivity equation (Stage 2). In this Table we present both 2SLS and 3SLS estimates. Although both estimators yield very similar second-stage estimates, we deem the 3SLS to be more reliable, since our second stage equation includes variables (namely, the usual inputs of a production function) that are not present in the first stage equation[9].

TABLE 10 ABOUT HERE

Focusing on our key variables, we see that the parameter associated with R&D intensity in the first stage equation (innovation production function) is equal to 0.10 (significant at the 1% level), which roughly corresponds to the estimate of 0.13 obtained in Model (3) assuming full reliability of the R&D intensity variable. More interesting to us is the estimate of the parameter associated with the share of innovative sales in the second stage equation (i.e., the CDM productivity equation). This parameter was never significant when we estimated Model (2) by Pooled OLS[10]: Even when we tried to compensate for the attenuation bias, the estimate remained negative (with a very low absolute value) and non significant. Here, with the innovation output variable being instrumented by R&D, we find that 1% increase in the share of innovative sales is associated with an increase in labour productivity of about 0.49 % to 0.52% (according to 2SLS and 3SLS estimates, respectively).

While it would be tempting to conclude that IV regression, by eliminating the measurement error in the innovation output variable thanks to the first stage regression, yields a more correct estimate of the innovation parameter, several reasons may actually explain the drastic change in this estimate. One prominent reason could be the correction of an underlying endogeneity bias in the productivity equation (more innovative firms may be more productive, but more productive firms may also be more innovative). A related reason is that R&D is obviously a very weak instrument for innovation output, and it is well-known that weak instruments tend to inflate the parameter estimate associated with the instrumented variable. Then again, after all is said and done, correcting measurement errors may be part of the story, and it would be interesting to apply our methodology for estimating λ to a simple CDM-like setting such as the one we have just described. Since panel

[9] We also implemented LIML, but this estimator yielded the same estimates as 2SLS so we downplay it here.
[10] It was not significant either with the conventional differenced estimators TD and BD. This suggests that this is not merely a matter of properly handling firm-level unobserved heterogeneity.

estimators for IV regression are scarce (Wooldridge and Semykina, 2013, suggest to rely on 2SLS and FE-2SLS[11]), we save this for further research, possibly with better data (e.g., a panel with a longer time dimension).

6. Conclusion

This paper aimed at giving an assessment of the importance of measurements errors in the key innovation variables of the CDM model, namely R&D intensity and the share of innovative sales. Do to so, we have estimated linear models (two productivity equations and an innovation production function) using original panel data obtained by merging three waves of the French CIS. By comparing different panel estimators, we were able to derive estimates of the λ parameter that provide an indication of the attenuation bias entailed by the CEV problem in our context.

We find that, in the productivity equations, the attenuation bias entailed by measurement errors seems to be more important (especially in terms of overall significance) in the innovation output variable (share of innovative sales) than in the innovation input variable (R&D intensity). This finding is consistent with the intuition expressed in Mairesse, Mohnen and Kremp (2005). Further analyses suggest that instrumenting the innovation output variable with the R&D variable in a very simplified CDM-type framework yield a significantly positive estimate of the effect of innovation output on labour productivity. While instrumentation may indeed reduce the effect of measurement errors in the innovation variable, other factors (such as the control for endogeneity provided by an IV approach) may also drive this result. Further research is thus needed to better disentangle the effect of measurement errors in a CDM-type framework that would allow us to correct simultaneously for endogeneity and/or selectivity (see for instance Wooldridge and Semykina, 2013). Other improvements to be dealt with in further research include the possibility to let measurement errors be autocorrelated, and the possibility to allow for correlated measurement errors in the innovation variables and capital stock variables.

References

Benavente, J.M. (2006) "The role of research and innovation in promoting productivity in Chile", *Economics of Innovation and New Technology*, 15(4/5), 301-315

Bound, J., Brown, C. and N. Mathiowetz (2001) "Measurement Error in Survey Data", in Heckman J.J. and E.E. Leamer (Eds), *Handbook of Econometrics*, Vol. 5, Chap. 59, pp. 3707-3843, North-Holland.

Chudnosky, D., Lopez, A. and G. Pupato (2006) "Innovation and productivity in developing countries: A study of Argentine manufacturing firms' behavior (1992-2001)", *Research Policy*, 35, 266-288.

Crépon B., Duguet E. and J. Mairesse (1998) "Research, Innovation and Productivity: An Econometric Analysis at the Firm Level", *Economics of Innovation and New Technology*, 7, 115-158.

Draper, N., and H. Smith (1998) *Applied Regression Analysis*, 3rd Ed., Wiley, New York.

Griffith, R., Huergo, E., Mairesse, J. and B. Peters (2006) "Innovation and Productivity across Four European Countries", *Oxford Review of Economic Policy*, 22(4), 483-498.

Hall, B. and J. Mairesse (1995) "Exploring the relationship between R&D and productivity in French manufacturing firms", *Journal of Econometrics*, 65, 263-293.

Hausman, J., Hall B. and Griliches Z. (1984) "Econometric Models for Count Data with an Application to the Patents-R & D Relationship", *Econometrica*, 52(4), 909-938.

Kmenta, J. (1997) *Elements of Econometrics*, 2nd Ed., Ann Arbor: University of Michigan Press.

[11] Note that, since FE tend to aggravate the attenuation bias, having only these two estimators at our disposal may make the implementation of our methodology somewhat more complicated.

Mairesse J.(1990) "Time Series and Cross Sectional Estimates on Panel Data: Why are They Different and Why Should They be Equal?" in Hartog, J., Ridder G. and J. Theeuwes (Eds) *Panel data and labor market studies*, Elzevier, 81-95.

Mairesse J., Mohnen P. and E. Kremp (2005), "The Importance of R&D and Innovation for Productivity: A Reexamination in Light of the French Innovation Survey", *Annals of Economics and Statistics*, 79/80, 489-530.

Mairesse J. and S. Robin (2012) "The Importance of Research for Innovation and Productivity: Comparing Different Estimators of the Innovation Production Function", Chapter 6 in *Innovation and Growth: From R&D Strategies of Innovating Firms to Economy-wide Technological Change*, Andersson, Johansson, Karlsson and Lööf (Eds), Oxford University Press, 368, 128-159.

Raffo, J., Lhuillery, S. and L. Miotti (2008) "Northern and southern innovativity: a comparison across European and Latin American countries", *European Journal of Development Research*, 20(2), 219-239.

Woodridge J. and A. Semykina (2013) "Estimation of dynamic panel data models with sample selection", *Journal of Applied Econometrics*, 28(1), pp. 47-61

Table 1: summary statistics on key variables, by wave of the French CIS

Variable	CIS 0	CIS 3	CIS 4	CIS 2008
# of employees in year *t*	328.7 (2267.6)	505.9 (2990.8)	362.2 (2230.0)	392.9 (1621.5)
# of employees in *t*-2	NA	498.6 (2939.6)	376.4 (2298.2)	414.4 (2330.9)
Labour productivity (firm sales/firm size) in year *t*	136.6 (470.4)	192.9 (784.1)	244.0 (982.7)	262.3 (682.9)
Labour productivity in year *t*-2	NA	184.9 (671.8)	230.0 (922.6)	245.6 (696.7)
Continuous R&D (1/0)	NA	0.37 (0.48)	0.34 (0.47)	0.35 (0.48)
R&D Intensity	NA	2.81 (8.00)	5.86 (10.77)	3.85 (15.24)
Product innovator (1/0)	0.67 (0.47)	0.49 (0.50)	0.43 (0.50)	0.46 (0.50)
% of innovative sales (coded 0/1/2/3 in CIS 0)	0.46 (0.77)	0.07 (0.13)	0.10 (0.20)	0.11 (0.21)

Mean values of the variables, with standard errors in parentheses

Table 2: innovation persistence from CIS 0 to subsequent waves of the survey

% of innovative sales in 1990	Continuous R&D in CIS 3	Continuous R&D in CIS 4	Continuous R&D in CIS 2008
0 to 10%	0.27	0.26	0.27
10% to 30%	0.52	0.5	0.49
30% to 70%	0.64	0.58	0.56
70% to 100%	0.69	0.54	0.59
>10%	**0.57**	**0.53**	**0.52**

% of innovative sales in 1990	Innovator in CIS 3	Innovator in CIS 4	Innovator in CIS 2008
0 to 10%	0.39	0.35	0.38
10% to 30%	0.65	0.58	0.62
30% to 70%	0.75	0.66	0.64
70% to 100%	0.72	0.68	0.66
>10%	**0.69**	**0.61**	**0.63**

Table 3: Estimates from Model 1 (productivity equation derived from extended production function)

	Simple linear regressions			
Variable	*Panel estimator*			
	TL	BL	TD	BD
Ln R&D intensity	0.11*** (0.02)	0.12*** (0.02)	0.003 (0.02)	-0.0003 (0.01)
R^2	0.10	0.08	0.004	0.004
Fisher *F*	41.73***	23.47***	2.35*	1.02
Observations	1674	1674	1167	1167
σ^2_e *(mean value)*	0.46	0.15	1.31	1.31
σ^2_x	2.96	2.29	2.62	1.25
λ	**0.17** (0.08)**	**0.07** (0.04)**	**0.46* (0.24)**	**0.58** (0.27)**
	0.16* (0.09)	**0.07* (0.04)**	**0.43* (0.25)**	**0.54* (0.29)**

Significance levels: *** 1%; ** 5%; * 10%

All standard errors are heteroskedasticity-robust whenever possible. Standard errors for λ are bootstrapped.

All models include time dummies. "Observations" refers to panel observations, not firms.

Table 4: Estimates from Model 2 (labour productivity regressed on innovation output)

	Simple linear regressions					
Variable	*Panel estimator*					
	TL	BL	TD1	BD1	TD2	BD2
% innovative sales (logit transform)	0.04*** (0.01)	0.04*** (0.02)	-0.03** (0.01)	-0.02 (0.02)	0.07*** (0.02)	0.06*** (0.02)
R^2	0.01	0.01	0.03	0.03	0.04	0.04
Fisher *F*	14.15***	8.03***	21.23***	25.37***	21.78***	29.06***
Observations	2112	2112	1398	1398	1398	1398
σ^2_e *(mean value)*	0.36	0.12	1.17	1.17	0.12	0.12
σ^2_x	3.06	1.72	5.85	2.88	4.05	1.52
λ	**0.37** (0.19)**	**0.24 (0.16)**	**0.33 (0.20)**	**0.51* (0.28)**	**0.08 (0.06)**	**0.21 (0.15)**
	0.36* (0.21)	**0.26 (0.20)**	**0.31 (0.21)**	**0.47* (0.28)**	**0.08 (0.06)**	**0.20 (0.16)**

Significance levels: *** 1%; ** 5%; * 10%

All standard errors are heteroskedasticity-robust whenever possible. Standard errors for λ are bootstrapped.

All models include time dummies. "Observations" refers to panel observations, not firms

Table 5: Estimates from Model 3 (innovation production function)

	Simple linear regressions			
Variable	*Panel estimator*			
	TL	BL	TD	BD
Ln R&D intensity	0.13*** (0.03)	0.15*** (0.03)	0.11** (0.05)	0.08 (0.06)
R^2	0.05	0.04	0.02	0.004
Fisher *F*	29.43***	7.96***	18.81***	3.94**
Observations	1506	1506	1047	1047
σ^2_e *(mean value)*	0.59	0.20	0.51	0.51
σ^2_x	2.92	2.29	2.61	1.33
λ	**0.23** (0.12)**	**0.10* (0.05)**	**0.31* (0.17)**	**0.54** (0.26)**

	0.27 (0.08)**	**0.20** (0.06)**	**0.35** (0.14)**	**0.69*** (0.13)**

Significance levels: *** 1%; ** 5%; * 10%

All standard errors are heteroskedasticity-robust whenever possible. Standard errors for λ are bootstrapped.

All models include time dummies. "Observations" refers to panel observations, not firms.

Table 6: Estimates from Model 1 (productivity equation derived from extended production function)

Variable	**Multiple linear regressions** *Panel estimator*			
	TL	BL	TD	BD
Ln R&D intensity	0.02*** (0.003)	0.03*** (0.005)	0.004 (0.003)	0.005 (0.005)
Ln C_{it-1} (investment)	0.04*** (0.005)	0.04*** (0.01)	0.014** (0.007)	0.003 (0.007)
Ln L (# employees)	-0.78*** (0.01)	-0.78*** (0.01)	-0.66*** (0.05)	-0.63*** (0.03)
Ln M	0.75*** (0.01)	0.75*** (0.01)	0.62*** (0.05)	0.58*** (0.02)
Adjusted R^2	0.92	0.92	0.70	0.70
Fisher *F (p-value)*	0.000	0.000	0.000	0.000
Observations	1296	1296	856	856
σ^2_e *(mean value)*	0.73	0.24	1.03	1.03
σ^2_x	2.30	1.72	2.03	1.04
λ	**0.32** (0.14)** **0.32 (0.25)**	**0.14 **(0.06) **0.14 (0.11)**	**0.41 (0.26) **0.39 (0.27)**	**0.52 (0.28) **0.48 (0.28)**

Significance levels: *** 1%; ** 5%; * 10%

All standard errors are heteroskedasticity-robust whenever possible. Standard errors for λ are bootstrapped.

All models include time dummies. "Observations" refers to panel observations, not firms.

Table 7: Estimates from Model 2 (labour productivity regressed on innovation output)

Variable	**Multiple linear regressions** *Panel estimator*					
	TL	BL	TD1	BD1	TD2	BD2
% innovative sales (Logit transform)	-0.004 (0.003)	-0.007* (0.003)	-0.002 (0.002)	-0.005 (0.002)	0.006** (0.003)	0.01*** (0.002)
Ln C_{it-1} (investment)	0.03*** (0.005)	0.03*** (0.005)	0.02** (0.01)	0.01 (0.01)	0.02** (0.01)	0.01 (0.01)
Ln L (# employees)	-0.79*** (0.01)	-0.79*** (0.01)	-0.70*** (0.05)	-0.72*** (0.02)	-0.70*** (0.05)	-0.71*** (0.02)
Ln M	0.77*** (0.01)	0.78*** (0.01)	0.69*** (0.04)	0.72*** (0.02)	0.69*** (0.04)	0.72*** (0.02)
Adjusted R^2	0.93	0.93	0.75	0.76	0.76	0.76
Fisher *F (p-value)*	0.000	0.000	0.000	0.000	0.000	0.000
Observations	1506	1506	992	992	992	992
σ^2_e *(mean value)*	2.13	0.71	9.71	9.71	1.40	1.40
σ^2_x	2.97	1.75	5.79	3.12	3.79	1.52
λ	**0.71*** (0.15) 0.67 (0.57)**	**0.40*** (0.11) 0.38 (0.30)**	**0.76*** (0.19) 0.73*** (0.22)**	**0.67** (0.30) 0.59* (0.31)**	**0.35 (0.35) 0.46 (0.28)**	**0.22 (0.20) 0.49* (0.28)**

Significance levels: *** 1%; ** 5%; * 10%

All standard errors are heteroskedasticity-robust whenever possible. Standard errors for λ are bootstrapped.

All models include time dummies. Observations" refers to panel observations, not firms.

Table 8: Estimates from Model 3 (innovation production function)

Variable	**Multiple linear regressions** *Panel estimator*			
	TL	BL	TD	BD
Ln R&D intensity	0.13*** (0.03)	0.15*** (0.03)	0.11** (0.05)	0.08 (0.06)
Ln Size	-0.01 (0.04)	-0.04 (0.05)	0.30 (0.25)	-0.45 (0.30)
Adjusted R^2	0.05	0.04	0.02	0.01
Fisher *F (p-value)*	0.000	0.000	0.000	0.018
Observations	1506	1506	1046	1046
σ^2_e *(mean value)*	1.41	0.47	1.52	1.52

σ^2_x	2.88	2.27	2.63	1.30
λ	**0.27* (0.14)** **0.55** (0.25)**	**0.12* (0.06)** **0.26* (0.15)**	**0.43** (0.19)** **0.56** (0.26)**	**0.66*** (0.25)** **0.63** (0.29)**

Significance levels: *** 1%; ** 5%; * 10%
All standard errors are heteroskedasticity-robust whenever possible. Standard errors for λ are bootstrapped.
All models include time dummies. "Observations" refers to panel observations, not firms.

Table 9: Error-in-Variable regressions

	Model (1) – extended productivity equation		**Model (2) – CDM productivity equation**		**Model (3) – innovation production function**	
	Level estimator (TL)					
	r = 1	*r = 0.68*	*r = 1*	*r = 0.29*	*r = 1*	*r = 0.73*
Ln R&D intensity	**0.02*** (0.003)**	**0.03*** (0.005)**	—	—	**0.13*** (0.03)**	**0.18*** (0.03)**
% Innovative sales	—	—	**-0.004 (0.003)**	**-0.01 (0.01)**	—	—
Ln C	0.04*** (0.005)	0.04*** (0.005)	0.03*** (0.004)	0.03*** (0.004)	—	—
Ln L (= ln firm size)	-0.78*** (0.01)	-0.77*** (0.01)	-0.79*** (0.01)	-0.79*** (0.01)	-0.01 (0.04)	-0.01 (0.04)
Ln M	0.75*** (0.01)	0.74*** (0.01)	0.77*** (0.01)	0.77*** (0.01)	—	—
R^2	0.92	0.93	0.93	0.93	0.05	0.06
Fisher *F* (p-value)	0.000	0.000	0.000	0.000	0.000	0.000
Observations	1296	1296	1506	1506	1506	1506
	Differenced estimator (TD)					
	r = 1	*r = 0.59*	*r = 1*	*r = 0.24*	*r = 1*	*r = 0.57*
Ln R&D intensity	**0.004 (0.003)**	**0.007 (0.006)**	—	—	**0.11** (0.04)**	**0.20** (0.08)**
% Innovative sales	—	—	**-0.002 (0.002)**	**-0.01 (0.01)**	—	—
Ln C	0.01*** (0.005)	0.01*** (0.005)	0.02*** (0.006)	0.02*** (0.004)	—	—
Ln L (= ln firm size)	-0.66*** (0.02)	-0.66*** (0.02)	-0.70*** (0.02)	-0.70*** (0.02)	0.30 (0.26)	0.32 (0.26)
Ln M	0.62*** (0.01)	0.62*** (0.01)	0.69*** (0.01)	0.69*** (0.01)	—	—
R^2	0.70	0.70	0.76	0.76	0.02	0.02
Fisher *F* (p-value)	0.000	0.000	0.000	0.000	0.000	0.000
Observations	856	856	992	992	1046	1046

Significance levels: *** 1%; ** 5%; * 10% (standard errors in parentheses).
All regressions include time dummies.
"Observations" refers to panel observations, not firms.

Table 10: "bare bones" CDM-type model (IV regression)

	Estimated by 2SLS	**Estimated by 3SLS**
	First stage equation *(Dependent variable = % of innovative sales)*	
Ln R&D intensity	0.10*** (0.03)	0.10*** (0.03)
Ln L (= log firm size)	-0.02 (0.05)	-0.02 (0.05)
R^2	0.04	0.04
Fisher *F* (p-value)	0.000	0.000
	Second stage equation *(Dependent variable = ln Labour Productivity)*	
% Innovative sales	0.49*** (0.18)	0.52*** (0.18)

Ln C	0.15*** (0.03)	0.11*** (0.02)
Ln L	-0.34*** (0.06)	-0.29*** (0.04)
Ln M	0.22*** (0.04)	0.22*** (0.02)
Fisher *F* (p-value)	0.000	0.000
Observations	1166	1166

Significance levels: *** 1%; ** 5%; * 10% (standard errors in parentheses).
All regressions include time dummies.
"Observations" refers to panel observations, not firms.

APPENDIX 1: Formulae for the variance of the measurement and for the attenuation bias

We start from Equation (6) in the core of the paper:

$$(6) \qquad \frac{\beta_{TL}}{1-\lambda_{TL}} = \frac{\beta_{BL}}{1-\lambda_{BL}} .$$

Using $\sigma^2_{e\,BL} = \frac{1}{T}\sigma^2_{e\,TL}$ (see Mairesse, 1990), we derive from Equation (6):

$$(7) \qquad \beta_{TL} - \lambda_{BL}\beta_{TL} = \beta_{BL} - \lambda_{TL}\beta_{BL}$$

$$\Leftrightarrow \beta_{TL} - \frac{\sigma^2_{e\,BL}}{\sigma^2_{x\,BL}}\beta_{TL} = \beta_{BL} - \frac{\sigma^2_{e\,TL}}{\sigma^2_{x\,TL}}\beta_{BL}$$

$$\Leftrightarrow \beta_{TL} - \frac{\sigma^2_{e\,BL}}{\sigma^2_{x\,BL}}\beta_{TL} = \beta_{BL} - \frac{\sigma^2_{e\,TL}}{\sigma^2_{x\,TL}}\beta_{BL}$$

$$\Leftrightarrow \beta_{TL} - \frac{\frac{1}{T}\sigma^2_{e\,TL}}{\sigma^2_{x\,BL}}\beta_{TL} = \beta_{BL} - \frac{\sigma^2_{e\,TL}}{\sigma^2_{x\,TL}}\beta_{BL}$$

Which, after rearranging, leads to:

$$(7.a) \qquad \sigma^2_{e\,TL} = \frac{\beta_{TL} - \beta_{BL}}{\frac{\beta_{TL}}{T\sigma^2_{x\,BL}} - \frac{\beta_{BL}}{\sigma^2_{x\,TL}}},$$

Replacing Equation (7.a) in the definition of the attenuation bias ($\lambda = \sigma^2 e/\sigma^2 x$) yields the formula for λ_{TL}:

$$(7.b) \qquad \lambda_{TL} = \frac{\sigma^2_{e\,TL}}{\sigma^2_{x\,TL}} = \frac{\beta_{TL} - \beta_{BL}}{\sigma^2_{x\,TL}\left(\frac{\beta_{TL}}{T\sigma^2_{x\,BL}} - \frac{\beta_{BL}}{\sigma^2_{x\,TL}}\right)}$$

where the β's are estimated parameters (by TL or BL) and all other parameters are constant terms (T is sample time length and the σ^2's are sample statistics).

Of course, since $\sigma^2_{e\,BL} = \frac{1}{T}\sigma^2_{e\,TL}$, once $\sigma^2_{e\,TL}$ has been estimated using (7.a), it also yields:

$$(8.a) \qquad \sigma^2_{e\,BL} = \frac{1}{T}\sigma^2_{e\,TL} = \frac{\beta_{TL} - \beta_{BL}}{\frac{\beta_{TL}}{\sigma^2_{x\,BL}} - \frac{T\beta_{BL}}{\sigma^2_{x\,TL}}}$$

from which, using again the definition of the attenuation bias ($\lambda = \sigma^2_e/\sigma^2_x$), we can derive a value for λ_{BL} :

(8.b) $$\lambda_{BL} = \frac{\sigma^2_{e\,BL}}{\sigma^2_{x\,BL}} = \frac{\beta_{TL} - \beta_{BL}}{\sigma^2_{x\,BL}\left(\frac{\beta_{TL}}{\sigma^2_{x\,BL}} - \frac{T\beta_{BL}}{\sigma^2_{x\,TL}}\right)}.$$

The formulae for σ^2_{eTD}, λ_{TD}, σ^2_{eBD} and λ_{BD} are obtained using similar calculations.

APPENDIX 2: Bootstrap procedures for the standard errors of λ

The estimates of the attenuation bias λ are derived from theoretical formulas. These formulas are valid under the usual CEV assumptions, which may or may not hold when using real data. The standard errors of λ are bootstrapped. We have defined λ as the ratio of σ^2_e (the variance of the measurement error) to σ^2_x (the variance of the variable measured with errors). In the data, the latter is measured by the sample variance of x, whereas σ^2_e is calculated according to one of the formulas given by Equations (8), (10), (13) and (15). It can then occur that, during a bootstrap iteration, the value of σ^2_e be negative, yielding a negative value of λ for that iteration.

To take this into account, we have followed two distinct bootstrapping procedures, which ensure that our estimate of λ is always between 0 and 1, as it should be. The first procedure simply consists in running a fixed number of bootstrap iterations (for the purpose of our application, we chose to run 3000 iterations), retaining only those that yield a consistent value of λ (i.e., such that $0 \leq \lambda \leq 1$). Thanks to the large number of iterations, we generally end up with a sample of values of λ that is large enough for inference.

In the second bootstrap procedure, we define σ^2_e using:

(A.1) $$\sigma^2_e = \sqrt{\sigma^4_e},$$

which ensures that σ^2_e is always positive, as a variance should be. It may, however, still lead to values of λ that are larger than 1, in which case we have to retain only those values that are smaller than or equal to 1. Unsurprisingly, the first procedure tends to yield smaller samples of values of λ and smaller standard errors, whereas the second one yields larger samples but tends to inflate the bootstrapped standard errors.

APPENDIX 3: errors-in-variables regressions estimates by CIS wave

Since the CDM equations are most often estimated on cross-sectional data (correcting for selectivity and endogeneity biases), it is interesting to examine how cross-sectional estimates of our three simple models (the extended productivity equation, the CDM-type productivity equation and the innovation production function) behave when we take into account the attenuation bias estimated using the CIS panel. Thus, this Appendix presents the results of errors-in-variables regressions for each of the three aforementioned models in each wave of the CIS present in our panel: 2000 (Table A1), 2004 (Table A2) and 2008 (Table A3).

Every year, each model is estimated thrice: First, assuming reliability r set to 1 (its maximum value), second, assuming r set to 1 and using the share of innovative sales in CIS 0 to capture innovation persistence and third, assuming r set to its TL estimated value ($r_{TL} = 1 - \lambda_{TL}$). The rationale for including the value of the share of innovative sales in CIS 0 in the second series of estimates is that much of the endogeneity bias present in a cross-sectional estimation of the CDM model is likely to be caused by firms' "innovativeness" (i.e. by the fact that some firms remain more innovative over

time than others). In this respect, and given the high degree of innovation persistence observed in Table 2, the share of innovative sales in CIS 0 is a useful proxy variable for "innovativeness", which may help reduce the aforementioned endogeneity bias. Overall, Tables A1 to A3 suggest that the estimates obtained when adding the share of innovative sales in CIS 0 remain very close to the baseline estimates, whereas controlling for the attenuation bias indeed leads to higher estimates.

Table A1: Errors-in-variables regressions estimates for CIS 3 (conducted in 2000)

	Model (1) – extended productivity equation			**Model (2) – CDM productivity equation**			**Model (3) – innovation production function**		
	r = 1	*r = 1 & CIS 0*	*r = 0.68*	*r = 1*	*r = 1 & CIS 0*	*r = 0.29*	*r = 1*	*r = 1 & CIS 0*	*r = 0.73*
Ln R&D intensity	**0.02*** (0.004)**	**0.02*** (0.004)**	**0.03*** (0.006)**	—	—		**0.09** (0.04)**	**0.07* (0.04)**	**0.13** (0.06)**
% Innovative sales	—	—	—	**-0.01*** (0.004)**	**-0.01*** (0.01)**	**-0.05*** (0.02)**	—	—	
Ln C	0.04*** (0.007)	0.04*** (0.007)	0.04*** (0.007)	0.03*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	—	—	—
Ln L (= ln firm size)	-0.79*** (0.01)	-0.80*** (0.01)	-0.79*** (0.01)	-0.80*** (0.01)	-0.80*** (0.01)	-0.80*** (0.01)	0.07 (0.06)	0.07 (0.06)	0.06 (0.06)
Ln M	0.77*** (0.01)	0.77*** (0.01)	0.76*** (0.01)	0.78*** (0.01)	0.78*** (0.01)	0.79*** (0.01)	—	—	—
% Innovative sales in 1990 *(Ref.: 0 to 10%)*									
10% to 30%	—	-0.02 (0.02)	—	—	-0.03 (0.02)	—	—	0.34** (0.16)	—
30% to 70%	—	-0.02 (0.02)	—	—	-0.01 (0.02)	—	—	0.42** (0.19)	—
70% to 100%	—	0.01 (0.03)	—	—	0.03 (0.03)	—	—	0.81** (0.34)	—
R^2	0.94	0.94	0.94	0.94	0.94	0.94	0.02	0.04	0.02
Fisher *F* (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Observations	419	413	419	492	486	492	455	449	455

Table A2: Errors-in-variables regressions estimates for CIS 4 (conducted in 2004)

	Model (1) – extended productivity equation			**Model (2) – CDM productivity equation**			**Model (3) – innovation production function**		
	r = 1	*r = 1 & CIS 0*	*r = 0.68*	*r = 1*	*r = 1 & CIS 0*	*r = 0.29*	*r = 1*	*r = 1 & CIS 0*	*r = 0.73*
Ln R&D intensity	**0.01*** (0.004)**	**0.01*** (0.004)**	**0.02*** (0.006)**	—	—		**0.13*** (0.05)**	**0.11** (0.05)**	**0.18*** (0.06)**
% Innovative sales	—	—	—	**-0.001 (0.003)**	**-0.001 (0.003)**	**-0.004 (0.01)**	—	—	—
Ln C	0.04*** (0.01)	0.04*** (0.01)	0.04*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	—	—	—

Ln L (= ln firm size)	-0.80*** (0.01)	-0.80*** (0.01)	-0.80*** (0.01)	-0.79*** (0.01)	-0.80*** (0.01)	-0.79*** (0.01)	-0.11 (0.08)	-0.11 (0.08)	-0.11 (0.08)
Ln M	0.76*** (0.01)	0.76*** (0.01)	0.76*** (0.01)	0.78*** (0.01)	0.78*** (0.01)	0.78*** (0.01)	—	—	—
% Innovative sales in 1990 *(Ref.: 0 to 10%)*									
10% to 30%	—	-0.04** (0.02)	—	—	-0.04*** (0.01)	—	—	0.36* (0.21)	—
30% to 70%	—	0.02 (0.02)	—	—	-0.01 (0.02)	—	—	0.61** (0.25)	—
70% to 100%	—	0.01 (0.04)	—	—	0.03 (0.04)	—	—	1.58*** (0.51)	—
R^2	0.94	0.94	0.94	0.95	0.95	0.95	0.02	0.05	0.02
Fisher *F* (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Observations	459	455	459	514	507	514	550	546	550

Table A3: Errors-in-variables regressions estimates for CIS 2008 (conducted in 2008)

	Model (1) – extended productivity equation			**Model (2) – CDM productivity equation**			**Model (3) – innovation production function**		
	r = 1	*r = 1 & CIS 0*	*r = 0.68*	*r = 1*	*r = 1 & CIS 0*	*r = 0.29*	*r = 1*	*r = 1 & CIS 0*	*r = 0.73*
Ln R&D intensity	**0.03*** (0.01)**	**0.03*** (0.01)**	**0.05*** (0.01)**	—	—		**0.16*** (0.04)**	**0.13** (0.04)**	**0.22*** (0.06)**
% Innovative sales	—	—	—	**-0.0004 (0.005)**	**0.0004 (0.005)**	**-0.001 (0.02)**	—	—	—
Ln C	0.03*** (0.01)	0.03*** (0.01)	0.03*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	0.02*** (0.01)	—	—	—
Ln L (= ln firm size)	-0.74*** (0.02)	-0.74*** (0.02)	-0.73*** (0.02)	-0.77*** (0.02)	-0.77*** (0.02)	-0.77*** (0.02)	0.03 (0.07)	0.04 (0.07)	0.02 (0.07)
Ln M	0.72*** (0.01)	0.72*** (0.01)	0.71*** (0.02)	0.76*** (0.01)	0.76*** (0.01)	0.76*** (0.01)	—	—	—
% Innovative sales in 1990 *(Ref.: 0 to 10%)*									
10% to 30%	—	-0.04* (0.02)	—	—	-0.04** (0.02)	—	—	0.48** (0.19)	—
30% to 70%	—	-0.01 (0.03)	—	—	0.01 (0.03)	—	—	0.26 (0.23)	—
70% to 100%	—	-0.10 (0.06)	—	—	-0.05 (0.05)	—	—	1.28*** (0.45)	—
R^2	0.89	0.89	0.89	0.91	0.91	0.91	0.03	0.05	0.03
Fisher *F* (p-value)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Observations	418	413	418	500	495	500	501	495	501