Labour Economics 76 (2022) 102182

Contents lists available at ScienceDirect

Labour Economics

journal homepage: www.elsevier.com/locate/labeco

The reliability of students' earnings expectations

Luis Diaz-Serrano [a,b,*], William Nilsson [c]

[a] Universitat Rovira i Virgili, ECO-SOS, Department of Economics, Av. de la Universitat, 1, 43204 Reus, Spain
[b] Universidad Antonio Nebrija, ECEMIN, C/ Sta. Cruz de Marcenado, 27, 28015 Madrid, Spain
[c] Universitat de les Illes Balears, Department of Applied Economics, Ctra Valldemossa Km 7,5, 07122 Palma de Mallorca, Spain

ARTICLE INFO

JEL:
C46
C83
I26

Keywords:
Earnings expectations
Test-retest
Reliability
Measurement error

ABSTRACT

The elicitation of subjective expectations in surveys has gained significant interest in economics. Students' earnings expectations have been widely used to model school and occupational choices, but quantification of the reliability of this variable has not yet been done. To what extent are earnings expectations affected by random measurement error? To assess this issue, in this paper we use different waves of a survey carried out in a Spanish university eliciting the earnings expectations of economics students. A test-retest method is applied with different time spans between the first wave and subsequent repetitions of the survey: two weeks and two and a half months later. A significant number of students declared large differences in earnings expectations, even in the case where the time span between surveys was short. Apart from the reliability, we also provide an estimate of the measurement error variance. With this information we show how a sensitivity analysis can be performed in external earnings expectations data research.

1. Introduction

Reliability testing of self-reported subjective measures and scales is quite common in the field of psychology but very rare in economics. Reliability refers to the consistency of a measure, and it can be expressed as the proportion of true variance of the variable of interest relative to the total observed variance.[1] An exception in the economics literature is the study by Krueger and Schkade (2008), who tested the reliability of self-reported life satisfaction using the test–retest methodology.[2] It is important to test the reliability of subjective self-reported measures in economics with a suitable methodology because the use of this type of variables to explain individuals' economics decisions has grown dramatically during the last decades.

Consider that a researcher wants to evaluate the decision that students face after finishing secondary schooling. They can enter the labour market or continue with university studies; if they decide to pursue higher education, they must also choose a subject to study. The decision to attend higher education is one of the most important of an individual's life since it implies incurring substantial costs. Therefore, in this context, it is not only the individual's tastes or cognitive abilities that play a role in the decision. According to human capital theory, one of the most important variables for such a decision is the earnings expectations in different scenarios, that is, with or without university studies. During the last two decades, interest in knowing students' earnings expectations has increased, and several studies have reported positive results regarding how valuable information can be obtained (Dominitz and Manski, 1996; Manski, 2004). In this regard, we agree with Jensen (2010), who claimed that, "Though many studies estimate these returns with earnings data, it is the perceived returns that affect schooling decisions, " (page 1, abstract). In this context, the initial interest in earnings expectations stems from an interest in using this variable to explain schooling and occupational choices. Recent evidence has shown that expected earnings elicited in surveys determine individuals' choice not only of the level of education to achieve (Attanasio and Kaufmann 2014; Belfield et al. 2016; Boneva et al., 2021) but also of the type of university studies to undertake (Arcidiacono et al., 2012; Schweri and Hartog, 2017). Further, empirical evidence has indicated that wage expectations can easily become self-fulfilling in terms of reservation wages (Caliendo et al., 2017), which in turn will determine future earnings in the labour market. Taking all these studies into account, no one can cast doubt on the importance of self-reported earnings expectations in economic analysis.

* Corresponding author at: Universitat Rovira i Virgili, ECO-SOS, Department of Economics, Av. de la Universitat, 1, 43204 Reus, Spain.
E-mail address: luis.diaz@urv.cat (L. Diaz-Serrano).

[1] Nunnally (1975) conducted a historical review of psychometric theory, of which reliability theory is an important part.

[2] The test–retest methodology consists of asking the same individuals the same question in two different periods. This methodology is the one that we use in this study, and it will be explained in more detail in Section 3. Kristensen and Westergaard-Nielsen (2006) analysed the reliability of job satisfaction using a survey in which this question was asked twice at different points. The immediacy of both questions makes these data invalid for reliability analysis since individuals may easily recall their first answer, which may cause the reliability to be overestimated.

https://doi.org/10.1016/j.labeco.2022.102182
Received 23 October 2020; Received in revised form 22 April 2022; Accepted 16 May 2022
Available online 20 May 2022

In the literature, some papers analysing earnings expectations have claimed that this variable is likely to be affected by random measurement errors (Arcidiacono et al., 2012; Attanasio and Kaufmann, 2017; Huntington-Klein, 2015; Reuben et al., 2015; Zafar, 2013). However, despite the awareness of the problem, no papers have analysed it. In this literature, due to the lack of empirical evidence on the magnitude of measurement errors, it is most common simply to disregard these concerns or add a very general discussion about reliability. In this paper, we test for the first time the reliability of students' earnings expectations. We use the test–retest methodology, and we quantify the magnitude of the random measurement error.[3] We believe this analysis to be important since, as some previous studies cited above have suggested, random measurement errors could be an important concern.

Measurement errors in survey questions can occur for many different reasons, and, in this context, it is important to make clear the difference between "validity" and "reliability". The questions could be vague and open to interpretation and misunderstanding, and the respondents could answer untruthfully or their answer might be driven by some kind of cognitive bias. These types of errors are somehow systematic and related to validity. In previous studies, the proposed solutions to this problem have consisted of reformulating and improving questions to avoid logical inconsistencies (Dominitz, 1998; Manski, 2004; Zafar, 2011). Measurement errors can also occur to different degrees if an interview is conducted or if the respondents self-report the answers; for example, respondents may fail to read the question carefully or mistakenly give the answer incorrectly. These types of errors can be random instead of systematic and affect the reliability of earnings expectations. In this paper, random measurement errors are the type of errors that we study. It is important to keep in mind that the test–retest approach that we use here only captures *random* measurement errors. Accordingly, measurement errors that are systematic, such as the ones that we explained above, will not be revealed by this methodology. For example, if the same misinterpretation is made on both test occasions, it cannot be detected in a test–retest study.

The concerns regarding the use of variables containing a random measurement error originate from the fact that it will have a harmful impact on regression models. For example, if earnings expectations are measured with errors and used as an explanatory variable, the coefficient associated with this variable and its variance will be biased and the conclusions from the model will not be correct (Carroll et al., 2006). A common assumption is that measurement errors are independent of the true variable, and these are referred to as *classical measurement errors*. Textbooks covering measurement errors in linear regression models have often shown that the estimated coefficient for a covariate measured with errors will tend to zero; that is, there is an *attenuation* bias. Stefanski and Carroll (1985) analysed the problem of covariates measured with errors in logistic regression. The attenuation bias is present in most situations for logistic regression, but they also clarified that, if there is a high imbalance in the sample regarding the amount of zeros and ones, the bias can be in the opposite direction.

We study random measurement errors in students' earnings expectations, and our three main contributions are the following. Firstly, we propose and perform a test–retest analysis of earnings expectations, elicited through a survey carried out with a sample of university students enrolled in a degree in economics and management. Secondly, we quantify the reliability and the measurement error variance, and the random measurement errors in students' earnings expectations are found to be important. Thirdly, we clarify how random measurement errors can be dealt with in analyses using students' earnings expectations as a covariate. We show how the measurement error variance that we estimate in our study can be used to gauge the sensitivity of the results from regression models to random measurement errors. Our results indicate that the reliability of students' earnings expectations is low.

The remainder of the paper is structured as follows. Section 2 discusses the related literature. In Section 3, we introduce the theoretical foundations of measurement errors and review how reliability in a measure can be evaluated and how this information can be used in other studies. Section 4 explains the data. In Section 5, we assess the degree of random measurement errors found in earnings expectations, and illustrate how to use this information with external data. Finally, in Section 6, we summarize and discuss the main implications of our results.

2. Review of the literature

The concept of measurement errors has already been considered in the previous literature on earnings expectations (Dominitz, 1998; Manski, 2004), but the method of dealing with the problem has been to reformulate and improve questions to increase the answering frequencies and avoid logical inconsistencies. In other words, until now, the focus has been on evaluating the validity of the questions – that is, the extent to which the given answer captures the concept that it is aimed to measure – but not the underlying random measurement error in the elicited responses. Basically, are respondents willing and able to respond in a meaningful way? Zafar (2011) analysed students' subjective earnings expectations and cognitive biases and did not find evidence of cognitive biases systematically affecting the reporting of beliefs. Again, Zafar's (2011) concern was the validity of the expectations and not the reliability.

In this paper, our focus is not on whether a systematic bias is affecting the answer but rather on whether the answer is affected by a random measurement error. We are interested in evaluating the reliability of students' earnings expectations. The reliability of an observed measurement can be quantified using repeated measures of the same phenomenon. The idea is to observe the extent to which the same values are elicited if individuals are asked once or several times during a short period, when true expectations at that moment should remain the same. If the repeated answers are very similar, the measurement has high reliability, but, if large changes are found, the reliability is low due to random measurement error. The closest analysis to ours is perhaps that of Wiswall and Zafar (2015). They carried out an experiment to test the extent to which individuals revise their earnings expectations when they receive new information regarding the realization of real earnings in the labour market. Despite the possibility that this analysis would be perceived as a kind of reliability analysis, it was not. Indeed, as we mentioned above, the questions were repeated in the same session, and, while this was sufficient for the authors' purpose, the immediacy of the repeated answers rendered the experiment inappropriate for reliability analysis since respondents may easily have recalled their first answer.

Another branch of the literature on students' earnings expectations has provided information on college returns and costs in experimental settings and evaluated the effect on decisions (see, for example, Bleemer and Zafar 2018, Hastings et al. 2015; Jensen 2010). In their experimental set-up, Bleemer and Zafar (2018) included a control group that did not receive any information. A follow-up survey was conducted two months later, and the data could have been used to measure test–retest reliability, but such an analysis was not performed.

The literature on expectations has paid substantial attention to the validity of measurements, asking such as "Are the questions understood correctly?", "Are the answers logically consistent?" and "Do answers bunch around specific values?" For example, the use of probabilistic questions is becoming more common in questionnaires, and the initial concerns about asking for probabilities were that individuals would tend to choose from a few probabilities, for example 0, 50 and 100%. The literature has found that survey respondents tend to use rounding

[3] Wiswall and Zafar (2015) investigated whether individuals revise their earnings expectations when they receive new information regarding real earnings in the labour market. For a control group, no new information was provided, but the repeated questions were asked only about 25 min later in the same survey; hence, this can hardly be regarded as a reliability analysis.

but not as heavily as first expected. Manski and Molinari (2010) analysed the way in which different respondents use different degrees of rounding and found that patterns can be detected using different survey questions on the probability that a future event will occur. In a recent paper, Giustinelli et al. (2020) reported that there is relative stability in respondents' rounding of probabilistic expectations over time. Gouret (2017) proposed a measure of coherence that allows the identification of uninformative subjective probabilities. In an empirical example, he observed that answers of 50% are often found to be uninformative, but it is important to be aware that this is not the only answer that can be affected. Delavande et al. (2011) investigated how different elicitation designs using visual aids can help respondents to express probabilistic concepts. Gouret and Hollard (2011) did not measure reliability but used a measure of respondents' coherence, with which noisy data can be distinguished from more valuable data. The idea is to retain data that are expected to have higher reliability based on more coherent answering behaviour. Van Santen et al. (2012) detected respondents who answered inconsistently, but they showed that simply excluding these cases implies an endogenous sample selection problem.

While not measuring reliability, some studies have investigated how transitory or persistent expectations behave over time. Zafar (2011) analysed students' subjective earnings expectations and repeated the questions about a year later. The correlation coefficients for different earnings expectations based on 16 different scenarios for eight different areas of study were found to be between 0.02 and 0.72. A correlation coefficient of 0.6 represents a coefficient of determination of 0.36; hence, quite a large share of the variation in earnings expectations in the second survey, one year after the first, cannot be explained by knowing the answer given regarding the same expectations a year ago. A low correlation does not necessarily indicate measurement error, although, in the absence of measurement error, students would have to make large changes in their earnings expectations from one year to another. It is possible that earnings expectations are measured with errors in both periods, which would imply a biased regression coefficient.

3. Theoretical foundations of measurment errors

3.1. Quantifying reliability and measurement error variance

Consider that one or more questions are used to create variable W to describe construct X. The true variable, X, is, accordingly, not observed, and W can be an error-prone measurement. When measurement errors are independent of the true variable, they are referred to as *classical measurement errors*:

$$W = X + \in, \in \sim \mathrm{N}\left(0, \sigma^2_{\in}\right) \tag{1}$$

The observed measure, W, is hence an unbiased measure of X, but it is accompanied by a random error term, $\in$. The errors are assumed to be independent of X, and $\sigma^2_{\in}$ is the measurement error variance. With these assumptions, the reliability is the ratio of the variance of the true variable to the variance of the observed measurement. The reliability can be estimated if we have two (or more) observed repeated measurements of the same variable (replicate measurements). Sometimes the replicate measurement is referred to as a parallel measurement. A parallel measurement can be obtained by adding another set of questions to measure the same construct, for example by asking the same individuals the same question(s) in two different periods.[4] The reliability of a variable can be calculated as the correlation, $\rho_{W_1W_2}$, between two parallel measurements, W_1 and W_2 (Carmines and Zeller 1979).

$$\rho_{W_1W_2} = \frac{\sigma_{W_1W_2}}{\sigma_{W_1}\sigma_{W_2}} = \frac{\sigma_{(X+\in_1)(X+\in_2)}}{\sigma_{W_1}\sigma_{W_2}} = \frac{\sigma^2(X)}{\sigma^2(W)} \tag{2}$$

[4] Parallel measurements have the same true underlying score and equal variance. The measurement errors found in different parallel measurements are assumed not to be correlated.

The covariance of these two measures is equal to the variance of the "true" variable, X, because the measurement errors, $\in_r$, are not correlated with each other or with X. With this estimate of the reliability ratio, we can easily find an estimate of the variance of the measurement error, which in fact is more easily transferable to external data.

The idea behind the test–retest method is to obtain two parallel measurements – that is, to perform the test once and then repeat it at least one more time. However, the appropriate timing of the measurements is difficult to determine. If the retest is repeated too soon, the reliability can be overestimated because individuals can remember their previous answer. On the other hand, if too much time is left between the first survey and the collection of the parallel measurement (retest), it is possible that the underlying true score has changed, because their true expectation has actually changed, and hence the reliability will be underestimated.

Another criticism is that the reliability could be low because of a reactivity problem (Carmines and Zeller, 1979). Measuring a concept once can cause a change in the true score of that concept. Carmines and Zeller (1979) specified some properties that parallel measurements should have: first, the expected value and variance of the parallel measurements should be equal, $E(W_1) = E(W_2)$ and $\sigma^2_{W_1} = \sigma^2_{W_2}$; second, the correlation of several parallel measurements should be equal for different pairs, $\rho_{W_1W_2} = \rho_{W_1W_3} = \rho_{W_2W_3}$; and, third, the correlation of parallel measurements and other variables (for example, Y) should also be equal, $\rho_{W_1Y} = \rho_{W_2Y} = \rho_{W_3Y}$. Once data have been collected, it is of course possible to test whether these assumptions are fulfilled.

3.2. Transferability of reliability and random measurement error variance to other data

Reliability induction is a concept that refers to the use of a reliability measure that has been obtained from a previous study instead of calculating one from the data that will be analysed (Vacha-Haase et al., 2002). Reliability should be considered as a property of a measured variable in a specific sample. When a new measure is introduced, the reliability is often evaluated to provide support of the usefulness of the variable, but the reliability does not need to be correct for subsequent studies. Accordingly, reliability induction can often be inappropriate because the sample composition and variability can be largely different from the original study. It can therefore be difficult to justify transferring a reliability ratio from a different previous study. The best option, of course, is for each study to include its own calculation of reliability that can be used in the analysis. However, we are aware that frequently it is not possible to perform this analysis. For example, in a large-scale survey, performing a retest is often not an option. Therefore, it is often easier to justify the transfer of measurement error variance. A consistent estimate of reliability is

$$\hat{\lambda} = \frac{\hat{\sigma}^2_w - \hat{\sigma}^2_{\in}}{\hat{\sigma}^2_w}, \tag{3}$$

where $\hat{\sigma}^2_w$ is the sample variance of W and $\sigma^2_{\in}$ is an estimate of the measurement error variance. If we assume that the measurement error variance is similar in both studies, we can resort to reliability induction and use $\sigma^2_{\in}$ from an external source. Kimball et al. (2008) also suggested recomputing the true-to-proxy variance ratio because the variability of the proxy could be different in different samples. The degree of measurement error can differ for many reasons, however – for example different phrasing of the question, a different testing environment, different characteristics of the individuals who answer and so on. Performing more test–retest studies can shed light on which situations give rise to more or less measurement error.

The assumption of equal measurement error variance is important. Studies on earnings expectations work with incomplete scenarios; that is, the information provided for the hypothetical situation for which the earnings expectations are considered is often not complete.

Manski (1999) distinguished *feasible* from *counterfactual* incomplete scenarios. In feasible incomplete scenarios, the respondents consider that the question includes information on a possible "to-be-realized" scenario and make their prediction, choice or expectation in accordance with the question. In subjectively counterfactual incomplete scenarios, the scenario does not support possible situations for a non-neglectable set of respondents. In *feasible* incomplete scenarios, the respondents need to speculate about the future, while, in *counterfactual* incomplete scenarios, it is also necessary to speculate about the past, which in some ways is inconsistent with the actual past. The latter is, of course, much more challenging, and it is not clear how ideal respondents, and even less how actual respondents, logically interpret such questions.

Reliability generalization (see, for example, Vacha-Haase et al. 2002), which is a meta-analysis of reliability, is of course not possible for students' earnings expectations due to the lack of quantification of reliability. In the future, this may be a path to obtain further knowledge on how differences in the sample composition and test environment are related to reliability.

3.3. Dealing with measurement errors in regression models

Carroll et al. (2006) provides a good starting point for someone who is unfamiliar with the literature. The key difference among the different approaches is the available information that can solve the problem. First, the information can be found either *internally*, that is, within the primary data, or *externally*, that is, in an independent study. When external information is used, we need to assume transferability so that the model parameters are also relevant to the primary data. Second, there are different categories within the extra information available: *validation* data, *replication* data and *instrumental* data. *Validation* data refer to the situation in which the correctly measured variable is available for a subsample of the observations in addition to the variable measured with error. This situation is rare, and, for our study, it is not even possible. *Replication* data are obtained when the variable measured with error is collected more than once, which allows us to estimate the variance of the measurement error. This is the approach that we adopt in this study. *Instrumental* data consider the case in which a second measurement, that is, another variable, which is (highly) correlated with the variable measured with error, among other characteristics, is available. The secondary measurement should not be part of the outcome model.

Both validation data and replication measurements are rare in economics, and, from this perspective, the most common approach in economics is to use the instrumental variable methods. Still, finding a variable that is correlated with a variable measured with error but that also meets other necessary requirements is not an easy task. Few studies use replication data; therefore, measures of reliability or the variance of the measurement error are generally not known. While the data requirements are less strict for instrumental data than for both validation data and replication data, it is important to ensure that the assumptions for instrumental variable methods are fulfilled. Violation of these assumptions can produce significant bias, even in a case in which the actual variance of the measurement error is zero (Carroll et al., 2006). It is also relevant to note that nonlinear error-in-variables regression models are not identified with the standard assumption for instrumental variables, and stronger assumptions are necessary (Amemiya, 1985).

3.3.1. Dealing with measurement errors in linear regressions

The implications of and solutions to classical measurement errors in linear regression have often been described in textbooks. A simple model, in which Y is the dependent variable and X is the only explanatory variable, is a starting point:

$$Y = \alpha + \beta_x X + \varepsilon \tag{4}$$

α and β_x are the coefficients to be estimated, and ε is a random error term. W, a realization of X that contains a classical measurement error, is available instead of X. Ordinary least squares regression of Y on W implies an attenuated regression slope, $\beta_{x*} = \beta_x \lambda$, where $\lambda = \frac{\sigma_x^2}{\sigma_x^2+\sigma_\epsilon^2} < 1$, and β_x is the regression slope if X is used.[5] The regression slope is, accordingly, biased towards zero with an attenuation factor λ, which has already been defined as *reliability*. The reliability ratio is the ratio of the true variance (σ_x^2) to the observed variance ($\sigma_x^2 + \sigma_\epsilon^2$). It is easy to extend the model to include other explanatory variables, Z, which are not measured with errors. When the measurement error is specified as above, a least squares regression of Y on (W, Z) implies $\beta_{x**} = \beta_x \lambda_1$, where λ_1 is *conditional reliability* and $\lambda_1 = \frac{\sigma_{x|z}^2}{\sigma_{w|z}^2} = \frac{\sigma_{x|z}^2}{\sigma_{x|z}^2+\sigma_\epsilon^2}$. $\sigma_{x|z}^2$ and $\sigma_{w|z}^2$ are the residual variances from the regressions of X on Z and of W on Z.

3.3.2. Dealing with measurement errors in generalized linear latent models

In a generalized linear latent and mixed model (*gllamm*), a measurement error model is considered in the relationship between the observed variables and the true variable (Rabe-Hesketh et al., 2004).[6] Rabe-Hesketh et al. (2003a) proposed a further refinement of the methodology and relaxed the normality assumptions of the true covariates and the measurement error. They developed a nonparametric maximum likelihood estimator for *gllamm*.[7] In Section 5.2, we provide a description of how to use an external measurement of the measurement error variance in a discrete choice model. This is performed with a generalized linear latent and mixed model using the command *cme* (covariate measurement error) and below, in this section, we provide the key features of this model.

Rabe-Hesketh et al.'s (2003a) study is our key reference for the model, although we keep the assumption of normality throughout our exposure. In this model, the unobserved true value of the explanatory variable is modelled as a latent variable, and three sub-models are used: an outcome model, a measurement model and a true covariate model. The outcome model, which takes the form of a generalized linear model, is of primary interest. The measurement model specifies the relationship between the true covariate and its measurements, while the true covariate model (or the exposure model in epidemiology) considers the relationship between the true covariate and the other explanatory variables. The first model is the *true covariate model*:

$$X_i = \gamma_0 + \gamma_1 Z_i + u_i,\ u_i \sim \mathrm{N}\left(0, \tau^2\right) \tag{5}$$

where X_i is the true (unobserved) covariate. The residuals, u_i, are assumed to be independent of the explanatory variable, Z_i. γ_1 are regression parameters. More than one explanatory variable is allowed in the model, although, for simplicity, we use only one regressor in our exposition of the model. Rabe-Hesketh et al. (2003a) refer to u_i as a latent variable. The reason for including explanatory variables in the specification is to avoid the unrealistic assumption that the true covariate, X_i, is independent of the other variables in the outcome model.

Next, we consider a *classical measurement error model*:

$$W_{ir} = X_i + \in_{ir},\ \in_{ir} \sim \mathrm{N}\left(0, \sigma_\epsilon^2\right) \tag{6}$$

where W_{ir} is the rth exposure measurement that is available for unit i and is an unbiased measurement of the true covariate. The measurement error, $\in_{ir}$, is independent of the true covariate X_i. Note that

[5] We refer to Carroll et al. (2006) for further details.

[6] The implementation of this methodology in Stata can be found in Rabe-Hesketh et al. (2002).

[7] The Stata program *gllamm* can also handle this situation. When a classical measurement error model is assumed and a single covariate is measured with errors, a wrapper command, *cme* (covariate measurement error) is available (Rabe-Hesketh, 2003). Rabe-Hesketh et al. (2003b) provided the details of its implementation in Stata. While *cme* uses the command *gllamm*, the syntax is more user friendly. Both the option to use replication measures and the option to use a plug-in measure of the measurement error variance are implemented.

we label the variance of the measurement error as $\sigma^2_{\in}$, while Rabe-Hesketh et al. (2003a) referred to it as σ^2_f. The property of a nondifferential measurement error applies to this specification, meaning that W_{ir} does not include any additional information about the outcome, Y_i, other than what is found in X_i and Z_i. The measured exposure is, accordingly, conditionally independent of the outcome given the true covariate, and different exposure measurements for the same i are conditionally independent given the outcome and are sometimes labelled *surrogates* (Carroll et al. 2006).

If the true covariate model is substituted into the measurement error model, we obtain the reduced-formed measurement error model, $W_{ir} = \gamma_0 + \gamma_1 Z_i + u_i + \in_{ij}$. Conditionally on Z_i, the variance of the measurement is $\sigma^2_{x|z} = \sigma^2_{\in} + \tau^2$, and we have more details of the *conditional* reliability, λ_1, which was introduced in Section 3.3. If the explanatory variables explain a large share of the variance in the true covariate, the conditional reliability can be substantially lower than the unconditional reliability (Rabe-Hesketh et al., 2003b).

The *outcome model* is specified as follows:

$$g(\mu_i) = \alpha_0 + \alpha_1 Z_i + \beta X_i \tag{7}$$

The outcome variable Y_i is modelled using a link function, $g(\cdot)$; for example, if the dependent variable is binary, we can choose a logit link function. Z_i is a set of explanatory variables measured without error, X_i is the true covariate and W_i is the observed realization of X_i, which is measured with errors. α and β are coefficients to be estimated. Generally, replacing X_i with W_i, will provide biased estimates of both α and β. The method will take advantage of replication measures, instrumental variables or a known measurement error variance to overcome this problem. The reduced-form outcome model is obtained by inserting the true covariate model into the expression above:

$$g(\mu_i) = \alpha_0 + \alpha_1 Z_i + \beta(\gamma_0 + \gamma_1 Z_i + u_i) = \delta_0 + \delta_1 Z_i + \beta u_i \tag{8}$$

where δ_0 and δ_1 are parameters of the reduced-form equation, that is, $\delta_0 = \alpha_0 + \beta\gamma_0$ and $\delta_1 = \alpha_1 + \beta\gamma_1$. A logit specification of the link function would imply $logit(P[d_i = 1|X_i]) = \delta_0 + \delta_1 X_i + \beta u_i$, where d_i is a binary dependent variable and P indicates a probability. When only one measurement, W_i, is available, $\sigma^2_{\in}$, τ^2 and β are not jointly identified by the first- and second-order moments of the observed variables. However, if we have prior knowledge of the measurement error variance, $\sigma^2_{\in}$, we can plug in this information to estimate the rest of the parameters. A limitation of this method is that the estimation uncertainty for that parameter is removed and the standard errors of the rest of the parameters will be slightly underestimated. Therefore, with uncertainty in the precision, it is reasonable to refer to the procedure as a sensitivity analysis.

The overview of the model is made in a rather strict way, but many assumptions can be relaxed in weaker versions. As mentioned above, Rabe-Hesketh et al. (2003a)(2003b) suggested the use of nonparametric models when assumptions of normality of the true covariates and the measurement errors are unrealistic. A congeneric measurement model can replace the classical measurement model explained above (Rabe-Hesketh et al., 2003b). This allows the use of an instrumental variable instead of a replicate measurement.

The estimation is made with maximum likelihood, and the likelihood function is

$$L(\theta_D, \theta_M, \tau) = \prod_i \int P(d_i|u_i;\ \theta_D) \prod_{r=1}^{n_i} g(W_{ir}|u_i;\ \theta_M) g(u_i;\ \tau) \mathrm{d}u_i \tag{9}$$

θ_D and θ_M are parameters from the outcome model and the measurement model, respectively. n_i is the number of replicate measures that are available for individual i.

For logistic regression with classical measurement error, as suggested above, the following specifications are used in the construction of the likelihood:

$$P(d_i|u_i;\ \theta_D) = \frac{\exp\{d_i[\alpha_0 + \beta\gamma_0 + (\alpha_1 + \beta\gamma_1)Z_i + \beta u_i]\}}{1 + \exp\{d_i[\alpha_0 + \beta\gamma_0 + (\alpha_1 + \beta\gamma_1)Z_i + \beta u_i]\}} \tag{10}$$

$$g(W_{ir}|u_i;\ \theta_M) = \frac{1}{\sqrt{2\pi\sigma^2_{\in}}} \exp\left(-\frac{[W_{ir} - (\gamma_0 + \gamma_1 Z_i + u_i)]^2}{2\sigma^2_{\in}} \right) \tag{11}$$

$$g(u_i;\ \tau) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{u_i^2}{2\tau^2} \right) \tag{12}$$

The likelihood can be integrated numerically using Gauss–Hermite quadrature. Notice that the expressions above include the structural parameters, α, γ, β, rather than the reduced-form parameters.[8]

In Section 5, we provide an empirical example of this model. The estimate of the measurement error variance for students' log earnings expectations in our example is presented in Section 5.1. In Section 5.2, we use this variance as external information in a generalized linear latent and mixed model. More specifically, we estimate a model that uses students' log earnings expectations to explain their expectations of achieving a university degree.

4. The data

4.1. Collection of the data used in the reliability analysis

The data for the reliability analysis were collected in 2015 from students taking the Analysis of Economic Data course at the University of the Balearic Islands (UIB). The course was taught in the first year of the economics and business administration degrees. We carried out three waves of a survey, the first of which was held in a computer lab during class time in the first week of the course. In addition to questions relevant to our study, the survey included a wide variety of questions to create a dataset that students would later use during the introductory course in statistics. The key questions for this study will be explained below. The second wave took place outside class time 14 days after the first wave. Finally, the third wave was conducted outside class time, and the average time span between the first survey and this third wave was about 74 days. The number of students participating in the third wave (301) was lower than the number of participants in the first two waves (399). The number of students who participated in all the waves was 280.

Generally, in a test–retest analysis, there is no optimal spacing between the collection of the parallel measurements. Test–retest studies only tend to consider two waves, and the spacing between waves depends on the topic. To the best of our knowledge, previous studies have not offered a discussion about the optimum timing between waves or stated why the authors decided to apply a specific timing between waves. In our case, as in Krueger and Schkade (2008), we opted to use a time span of two weeks between the first and the second wave. Another important issue that other studies have not taken into account is that reliability could be overestimated if the respondents remember their initial answer. Therefore, to improve our reliability analysis, we also decided to include a third wave with a longer time span between the second and the third wave.

4.2. Key variable: Students' earnings expectations

The key interest for this study is students' earnings expectations. This variable is obtained from the elicited responses to the following question:

What do you think your average gross monthly salary will be after you have graduated in the studies that you are currently attending?

This question was repeated in all three waves. In Fig. 1, we depict the estimated kernel densities of the log earnings expectations for the three waves. The densities for the different waves are very similar, suggesting

[8] The model is implemented in Stata, using the command *cme* or directly using *gllamm*. These Stata commands provide estimates of these structural parameters.

Table 1
Descriptive statistics for earnings expectations.

Panel A									
	Wave 1			Wave 2			Wave 3		
	Mean	*s.d.*	*n*	*Mean*	*s.d.*	*n*	*mean*	*s.d.*	*n*
Earnings expectations	1634.3	771.0	464	1645.0	733.7	419	1628.7	619.9	310
Log earnings expect.	7.32	0.38	464	7.33	0.37	419	7.33	0.36	310
Panel B									
	Wave 2 – Wave 1			Wave 3 – Wave 1			Wave 3 – Wave 2		
	mean	*s.d.*	*n*	*Mean*	*s.d.*	*n*	*mean*	*s.d.*	*n*
Log earnings expect.	0.0177	0.3081	314	0.0056	0.2945	231	0.0065	0.3434	219

Notes: Earnings expectations are measured in Euros per month. *n* refers to the number of valid answers. The logarithmic transformation is the natural logarithm.

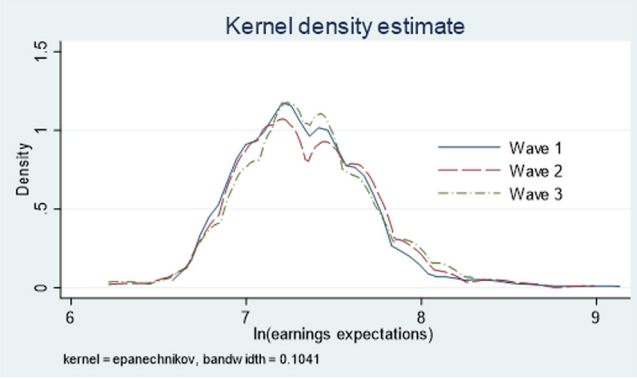

Fig. 1. Distribution of the three repeated measures of earnings expectations.

Table 2
Correlation matrix on log earnings expectation.

	Wave 1	Wave 2	Wave 3
Wave 1		0.6118 (399)	0.6197 (301)
Wave 2	0.5898 (280)		0.5632 (284)
Wave 3	0.6378 (280)	0.5684 (280)	

Notes: The three cells in the lower left corner refer to the group of students that answered all three waves. The three cells in the upper right corner refer to students in the two waves analysed, but not necessarily in the third. The sample size is included in parenthesis.

that the elicited earnings expectations in the three waves might be good parallel measurements. In Panel A of Table 1, we report the corresponding means and standard deviations. The average earnings expectation in the three waves ranges between €1628 and €1645 per month. The standard deviations are also fairly similar in the three waves. In Panel B of Table 1, we report the distribution of the within-individual differences between waves, and we observe that the average is close to zero, while the standard deviation is rather high. This is the first indication that some students have made large changes to their answers. This is evaluated in more detail in the following section.

5. Results

5.1. Test-retest reliability of log earnings expectations

Table 2 contains a correlation matrix on test–retest reliability; the values in parentheses refer to the sample size. The correlation coefficient is calculated for both a restricted sample composed of the students who participated in all three waves (hereafter restricted sample) and an unrestricted sample consisting of the students participating in each pair of waves (hereafter unrestricted sample).

For the unrestricted sample in the first two waves, the correlation of own log earnings expectations, that is, our estimate of the ratio of the variance of the true variable (X) to the variance of the observed variable (W), is about 0.61, which implies a measurement error variance of 0.0516. The correlation for this measure for the first and third waves is practically identical, 0.62, while that for the second and third waves is 0.56. The correlations for the restricted sample are fairly similar: 0.59, 0.64 and 0.57, respectively. All these correlations taken together suggest fairly low reliability; hence, in our data, random measurement error seems to be important. Unfortunately, we cannot compare our results with those of other studies on the reliability of earnings expectations because of the lack of such studies. However, surveys including questions about actual observed earnings are also found to have measurement errors, and our results can be compared with these results, though they do not measure the same concept. Bound and Krueger (1991) reported a ratio of variance of the signal to the total variance for earnings of 0.82 for men and 0.92 for women. This information was obtained with survey and complementary data from administrative social security payroll tax records. Gottschalk and Huynh (2010) also used administrative data to find a reliability ratio for survey data. Their measure of annual earnings was based on monthly observations, but the data contained missing values; therefore, they imputed these values. The reliability was 0.67 for the full sample (n=3742) and 0.73 for a restricted sample (n = 2587) with no imputed earnings. This reliability is not far from what we estimate with our data.

In our case, about 26% of the respondents in the first and second waves maintained the same earnings expectation in both waves. About 27% kept the same earnings expectation in the first and third waves and about 26% in the second and third waves. Approximately 14% of those who participated in the three waves provided the same earnings expectation in all three waves. The mean of the absolute difference in earnings expectations between the first and the second wave was about 374 euros. The median of the absolute difference was 200 euros, while the third quartile was 500 euros. The corresponding values between the first and the third wave were 399, 300 and 500 euros, respectively. These results indicate that many students changed their reported earnings expectations in a reasonably short period of time, implying that the variable could contain a random measurement error. This circumstance has been already suggested in previous studies; however, they did not offer much discussion. Our measure of reliability is evaluated further in the following sections.

5.1.1. Are the repeated measurements of earnings expectations parallel measurements?

We performed several tests to determine whether our repeated measurements of earnings expectations are parallel measurements. The test involves testing the equality of means and variances of the earnings ex-

Table 3
p-Values on hypothesis of equal mean and variances.

	Mean			Variance		
	Wave 1	Wave 2	Wave 3	Wave 1	Wave 2	Wave 3
Wave 1		0. 4186 (399)	0.9758 (301)		0.8229 (399)	0.8040 (301)
Wave 2	0.5654 (280)		0.8249 (283)	0.2509 (280)		0.3429 (283)
Wave 3	0.7133 (280)	0.8168 (280)		0.7699 (280)	0.3920 (280)	

Notes: The three cells in the lower left corner refer to the group of students that answered all three surveys. The three cells in the upper right corner refer to students found in the two surveys analysed, but not necessarily in the third. The same idea concerns the variance. The sample size is included in parenthesis.

Table 4
Correlation of log earnings expectations and other variables.

	Self-assessed skills				Risk
	Maths	Verbal	Social	Commercial	aversion
Wave 1	0.0347	0.1025	0.0824	0.1030	0.0866
Wave 2	0.0895	0.0783	0.1248	0.1577	0.1645
Wave 3	0.0628	0.0808	0.1239	0.0844	0.1094

Note: Risk aversion is measured on a scale where higher values indicate *less* risk aversion.

pectations elicited in the different waves, three in our case. We report the results of these tests in Table 3. The table contains the *p*-values of these tests, which include results for both the restricted (below the main diagonal) and the unrestricted sample (above the main diagonal). The hypothesis of equal means is never rejected at the conventional significance levels. The same applies to the hypothesis of equal variances. We also performed Kolmogorov–Smirnov tests for paired data to test the equality of the distribution functions for all the combinations of pairs of waves. The hypothesis of equality in distribution was also never rejected.

An additional test to confirm that the measurements of earnings expectations observed during the different waves can be considered as parallel measurements consisted of calculating the correlation coefficients of these repeated measures of earnings expectations with other variables picked up from the same study. In this regard, in the first wave of the survey, we asked the surveyed students to self-assess their own self-perceived skills in the following four domains: mathematics, verbal, social and commercial. Self-assessment was carried out on the basis of an ordinal scale ranging from 1 to 10. We also constructed a subjective risk aversion measure by asking students their willingness to take risk using a 10-point ordinal scale, on which (1) is not at all willing to take risks and (10) is fully prepared to take risks. See the Appendix for the exact formulation of these survey questions. In Table 4, we show the correlation analysis of students' earnings expectations with these five self-assessed measures of self-perceived skills and risk aversion.

The correlation between the log earnings expectations – measured in any of the three waves – and the different self-assessed skills and willingness to take risk is found to be very weak. Since log earnings expectations appear to be measured with random errors, the estimated correlation coefficients are likely to be underestimations of the correlation of the true variables. It is also possible that other variables are also measured with errors, which would exacerbate the measurement error problem. The hypothesis that the correlation is equal in the two waves is never rejected. The results of the tests reported in Tables 3 and 4 support the assumption that the repeated measurements of earnings expectations in our study can be used as parallel measurements.

5.1.2. Is the reactivity problem, new information or memory altering the reliability?

One reason for the low correlation between the first and the second wave could be that students had discussed their answers after the first wave and that their answers in the retest carried out two weeks later might have been adjusted as a result of these exchanges of information with their classmates. Another possibility is that, after the first wave, which was conducted in the computer lab, some students might have searched for additional information about earnings, but we think that this option is very unlikely. During the survey, students were compelled not to discuss their answers; however, as we mentioned above, there is no guarantee that they did not discuss it afterwards. In the second wave, we asked the students whether they had discussed their responses with someone, and 277 respondents answered that they had not. For this subgroup, the correlation is 0.6163, which is very similar to that found for the full sample. This suggests that, in this case, the reactivity problem is not causing an underestimation of the reliability.

It is worth noting that, despite the time spans between periods being very different, about 14 and 74 days, the correlations across periods are very similar. If students remember their first answer, this can cause the reliability from the first to the second wave to be overestimated. On the contrary, if students receive new information, thus potentially prompting them to revise their expectations, the reliability from the first to the third session will be underestimated. These biases work in opposing ways, which would make the first measure "too large" and the second "too small", but, despite this, the point estimates are found to be almost identical. To some extent, these results mitigate the concerns about possible biases due to memory or new information. It is important to keep in mind that many different scenarios of earnings perception were included in the first survey, making it more difficult to remember the answers. If a small survey with few questions is used, the risk that students will remember their answers will be higher and a time span of only two weeks between waves could be too short.

5.1.3. Is the variance of the log earnings expectation particularly small in our sample?

The low reliability that we find in Section 5 could be a consequence of analysing a sample with low variance in log earnings expectations. Transferring the measurement error variance to samples with higher variance would imply higher reliability (Eq. (3)). Accordingly, it is interesting to compare the variance of the log earnings expectations in our sample with that estimated in other studies. The literature on students' earnings expectations has often used samples that consist of a group of students, generally from the same university and in most cases studying for the same degree. The variance of earnings expectations can hence be rather low. Hartog and Diaz-Serrano (2013) reviewed the literature and found that the coefficient of variation of students' earnings expectations ranged from 0.2 to 0.4 in the different studies. Brunello et al. (2004) analysed earnings expectations in 10 European countries and reported a standard deviation of after-college log expected earnings of 0.56. We consider this figure to be rather low since these authors pooled earnings expectations from different countries. Webbing and Hartog (2004) reported a coefficient of variation of 0.37, which fits within the values reported by Hartog and Diaz-Serrano (2013), and a standard deviation of log earnings expectations of 0.29. In our study, in the first wave, the coefficient of variation of earn-

Table 5
Reliability and descriptive statistics for log earnings expectation.

Description	Notation	Value in current study
Reliability	$r_{w_1 w_2}$	0.6118
Variance of measurement error	$\hat{\sigma}^2_{\in}$	0.0516
Mean of log earnings expectation	$\bar{W}$	7.3260
Variance of log earnings expectation	$\hat{\sigma}^2_{W}$	0.1330
Variance of true log earnings expectations	$\hat{\sigma}^2_{X}$	0.0814

Note: The reliability is calculated using answers from the first and second wave because the sample size is largest for this combination. The mean and variance refer to the joint mean and variance for the two waves.

ings expectations (i.e., before calculating the logarithm) is about 0.47. The standard deviation of log earnings expectations is about 0.37. These values are quite large compared with the ones estimated in other studies, and it is clear that our sample does not stand out in the literature as being a case of particularly low variation in log earnings expectations. All this evidence taken together suggests that samples with low variation in log earnings expectations are common and, accordingly, the low reliability found in our study is what we should expect if a reliability analysis had been carried out in the previous studies dealing with earnings expectations.

An interesting feature of our survey is that we also collected students' earnings expectations within different hypothetical scenarios. More specifically, we randomly assigned a different university field of study (medicine, sociology, engineering, etc.) to each student and asked them to report their earnings expectations in the hypothetical case that they graduated from this randomly assigned university field of study. This randomization was only made in the first wave; therefore, the earnings expectations for the hypothetical field of study assigned to each participant refers to the same field in the three waves.[9] In these randomly assigned fields of study, the coefficient of variation in the first wave is 0.46, the estimated correlation of the log earnings expectations in the test–retest is 0.6575 (n=396) and the measurement error variance is 0.0558. This correlation is quite similar to the one estimated for "regular" log earnings expectations (0.6118). If we restrict the sample to those students who did not discuss earnings expectations with anyone, for the hypothetical randomly assigned fields of study, we estimate a correlation of 0.6374 (n=274), which is very similar to the one estimated using the full sample and the one referring to their current field of study (0.6163). This result indicates that the measurement errors in the two types of earnings expectations are practically the same.

5.2. Using the measurement error variance as a sensitivity analysis in regression models

The overall impression from the previous section is that repeatedly asking students about their earnings expectations with a time span of either 14 or 74 days in both cases provides suitable parallel measurements for quantifying reliability and obtaining the measurement error variance for students' log earnings expectations. This section shows how to use the measurement error variance that we obtained in our study in another similar study using different data. The key assumption is that the variances of the measurement error are similar; hence, for this exercise, we consider a survey in which expectations have been elicited in a similar way to the data that we used in our reliability study. From our current study, we obtained the variance of the true log expectations, $\hat{\sigma}^2_{W} r_{w_1 w_2} = 0.1330 \times 0.6118 \approx 0.0814$; accordingly, the variance of the measurement error is $\hat{\sigma}^2_{\in} = \hat{\sigma}^2_{W} - \hat{\sigma}^2_{X} = 0.1330 - 0.0814 \approx 0.0516$. Table 5 provides the relevant information.

As we explained previously, the consequences of using an explanatory variable with a classical measurement error include, for example, that the estimated coefficient is biased. However, as we explained above, under certain assumptions, the variance of the measurement error obtained in one study can be used to correct the bias of the estimated coefficient in another study in which the explanatory variable of interest (e.g., earnings expectations) is measured with errors and reliability analysis is not possible.[10] Note that, before our study, there was no previous information on the magnitude of the variance of log earnings expectations' measurement error. When repeated measurements are available, the methodology proposed by Gillen et al. (2019) can also be applied to take the measurement error into account in a regression model. In this example, we estimate a logit model, but this correction can also be used in a linear regression model. [11] To assess the extent of the random measurement error problem, in this example we use three samples, which are described below.

5.2.1. Samples

Sample 1: University of Balearic Islands (UIB)

The first sample we use in our example is the same we have used to carry out our reliability analysis. That is, the sample of students attending the Economics degree and the Management degree at the University of Balearic Islands (UIB). This data has been described in detail in Section 4. In this example, we estimate the impact of log earnings expectations to explain the choice between being attending the management degree (1) vs. the economics degree (0)

Sample 2: Rovira i Virgili University (URV)

The second sample consists of data coming from a survey carried out in 2013 to students attending the Management degree and the Finance & Accounting degree at Rovira i Virgili University (URV). This is a Catalan university, which is located 100 kms to south of Barcelona. In the URV study, we used the same questionnaire as the one used in the UIB study. The interesting feature of the URV data is that the sample is very similar to the UIB sample, and the standard deviation of log earnings expectations in both samples is the same. In this context, to assume that the measurement error variance in both samples is similar is very plausible. In this example, we estimate the impact of log earnings expectations to explain the choice between being attending the management degree (1) vs. the finance and accounting degree (0)

Sample 3: High school students in Catalonia

This sample comes from a survey conducted in April/May 2016 in public secondary education schools in Catalonia, a region in the northeast of Spain, of which the capital is Barcelona. The survey was specially designed to collect data for a research project on the determinants of students' expectations and school choices. The database contains information regarding socio-demographic characteristics, personality traits, cognitive and non-cognitive skills, family background, school outcomes, students' academic intentions for the next year and academic and labour market expectations in the medium and the long run, including earnings expectations.

The invitation to participate in the survey was sent to all the public secondary schools in Catalonia (564 in total), of which 92 agreed to participate.[12] Almost half of the surveyed students (45.5%) were in their last year of compulsory secondary education (aged 16), and the rest were enrolled in post-compulsory secondary education (aged 17–

[9] In the Appendix, we provide the complete set of randomly assigned hypothetical university fields of study.

[10] The Stata program *gllamm* (see Rabe-Hesketh et al. 2003a) implements a generalized linear model with covariate measurement errors, and, in the absence of replicate measurements, the measurement error variance can be provided by an external study.

[11] Gillen et al. (2019) named their methodology *obviously related instrumental variables* (ORIV), and the idea was to stack two models, one in which the second measure is used to instrument the first and another in which the opposite is used.

[12] For those schools that decided to participate, the questionnaire was administered in the classroom. Since class attendance is compulsory for those attending secondary education, all the students present on the day of the survey completed the questionnaire. Thus, we consider that attrition is not an issue of concern.

Table 6
Modelling expectation to obtain a university degree.

	High School sample Expectation of university degree		URV sample Management (1) vs. Finance & Accounting (0)		UIB sample Management (1) vs. Economics (0)	
	Logit	CME	Logit	CME	Logit	CME
Log(expected earnings)	0.368***	0.464***	-1.395***	-2.701***	-0.867**	-1.686**
	(0.0956)	(0.121)	(0.381)	(0.826)	(0.410)	(0.834)
Female	-0.252*	-0.263**	-0.259	-0.366	-0.563**	-0.419*
	(0.134)	(0.134)	(0.245)	(0.268)	(0.275)	(0.243)
Born abroad	0.813***	0.827***	-0.463	-0.450	0.447	0.302
	(0.100)	(0.101)	(0.355)	(0.365)	(0.587)	(0.475)
Grades CE	0.686***	0.688***	0.399***	0.367***	0.0153	0.102
	(0.0535)	(0.0536)	(0.134)	(0.135)	(0.0822)	(0.0713)
Self-reported math skills	0.0836***	0.0823***	0.0139**	0.0176**	0.109	0.0145
	(0.0215)	(0.0215)	(0.00620)	(0.00686)	(0.0809)	(0.0640)
Father college degree	0.385***	0.387***	0.447	0.530	-0.765*	-0.417
	(0.145)	(0.145)	(0.351)	(0.362)	(0.404)	(0.309)
Mother college degree	0.529***	0.526***	-0.0870	-0.0251	0.210	0.144
	(0.130)	(0.130)	(0.338)	(0.350)	(0.327)	(0.254)
Constant	-7.265***	-7.957***	7.596***	17.29***	6.562**	12.44**
	(0.778)	(0.945)	(2.741)	(3.039)	(3.062)	(5.974)
Assumed ME variance		0.05		0.05		0.05
Estimated reliablity		0.79		0.53		0.53
St. dev. (log exp.earnings)		0.50		0.34		
Observations	2,921		2,921	374		268

Standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

18). Among the latter group, 35.7% were students attending high school and 18.8% were following a lower vocational training programme. In this example, we focus on adolescent students in their last year of compulsory secondary education.

In these data, adolescent students were asked about their earnings expectations under different scenarios. The specific scenario that we are interested in regards the earnings expectations in the hypothetical situation that respondents enter the labour market with a university degree. More specifically, adolescents were asked to answer the following question: *"What monthly salary do you think you would earn if you found a job with a college degree?"*. In this example, we estimate a logit binary model, which requires the variance of the measurement error as an input unless a repeated measurement is available.[13] We use log earnings expectations, as defined above, to explain students' expectations of achieving a university degree. The dependent variable is created from the question "*What is the highest level of education you expect to achieve?*". We create a dummy variable that takes the value 1 if the surveyed students answer that they expect to achieve a university degree and 0 otherwise. According to human capital theory, we should expect a positive relationship between earnings expectations and the expectations of achieving a university degree. To test the extent to which the results are affected by measurement errors, we first estimate the model without considering measurement errors. Afterwards, we re-estimate the same model assuming the existence of measurement errors, and we specify the variance of the measurement error in the logit model (*cme*) obtained from our reliability analysis using the data on earnings expectations of the UIB students. We estimate the measurement error variance as 0.05 (Table 5). In this example, we make several simplifying assumptions. We assume classical measurement errors, in which the true covariate is normally distributed (Rabe-Hesketh et al. 2003b), and that the other covariates in the model are measured without errors.

5.2.2. Results

The results of our example are reported in Table 6. We will start by commenting on the results obtained in the sample of high school students. The results indicate that the estimated coefficient for earnings expectation is highly statistically significant and that the estimates that do not consider the existence of measurement error in log earnings expectations provide a downward-biased coefficient compared with the models that do: 0.368 vs. 0.464. The implication of this requires further elaboration in the logit model. A 25% increase in the earnings expectations increases the log odds of expecting to achieve a university degree by $0.368 \times \log(1.25) = 0.082$, while, if we consider the existence of measurement errors (cme), this increase is $0.464 \times \log(1.25) = 0.103$. The odds ratios from an increase of 25% in the earnings expectations are calculated as $\exp(\beta \times \log(1.25))$. Accordingly, the odds ratios are 1.085 and 1.109, respectively. These results indicate that a 25% increase in the earnings expectations raises the odds of expecting to achieve a university degree by 8.5% with the conventional logit model, whereas it rises by 10.9% with the covariate measurement error model (*cme*). A simple example can explain the meaning of this result in terms of a change in the probability. The proportion of students who expect to achieve a university degree is 0.76. If we assume that this is the probability for the average individual, then the odds are 3.17. With these figures, a 25% increase in the earnings expectation raises the odds to $3.17 \times 1.085 = 3.437$ in the logit model and up to $3.17 \times 1.109 = 3.512$ in the covariate measurement error model (*cme*). Transferring these numbers back to a probability scale, we have 0.774 and 0.778, respectively. Therefore, the change in the probability of expecting to achieve a university degree due to a 25% increase in earnings expectations is, accordingly, 0.014 for the logit model and 0.018 for the covariate measurement error model (*cme*). These are very small changes in the probability, in response to an important change in earnings expectations. We should, however, keep in mind that the initial probability is already fairly high in this sample. The estimated reliability for the data used in the example (Table 6) is 0.79, which is higher than the reliability that we estimated in our test–retest study (0.61). This is because the standard deviation of log earnings expectations in these data (0.50) is higher than that in the data used to carry out the test–retest analysis (0.34). This underlines the importance of transferring the measurement error variance instead of the reliability.

Results regarding the UIB and the URV samples are quite different to the ones obtained with the sample of high school students. In the UIB sample, the percentage of students attending the management degree is 64%. According to the estimated coefficients reported in Table 6, and following the steps described above, we obtain that a 25% increase in the log earnings expectations causes an increase of the probability of being attending the management degree of 4.3 percentage points with

[13] This model can be estimated using the *cme* Stata command (Rabe-Hesketh, 2003b).

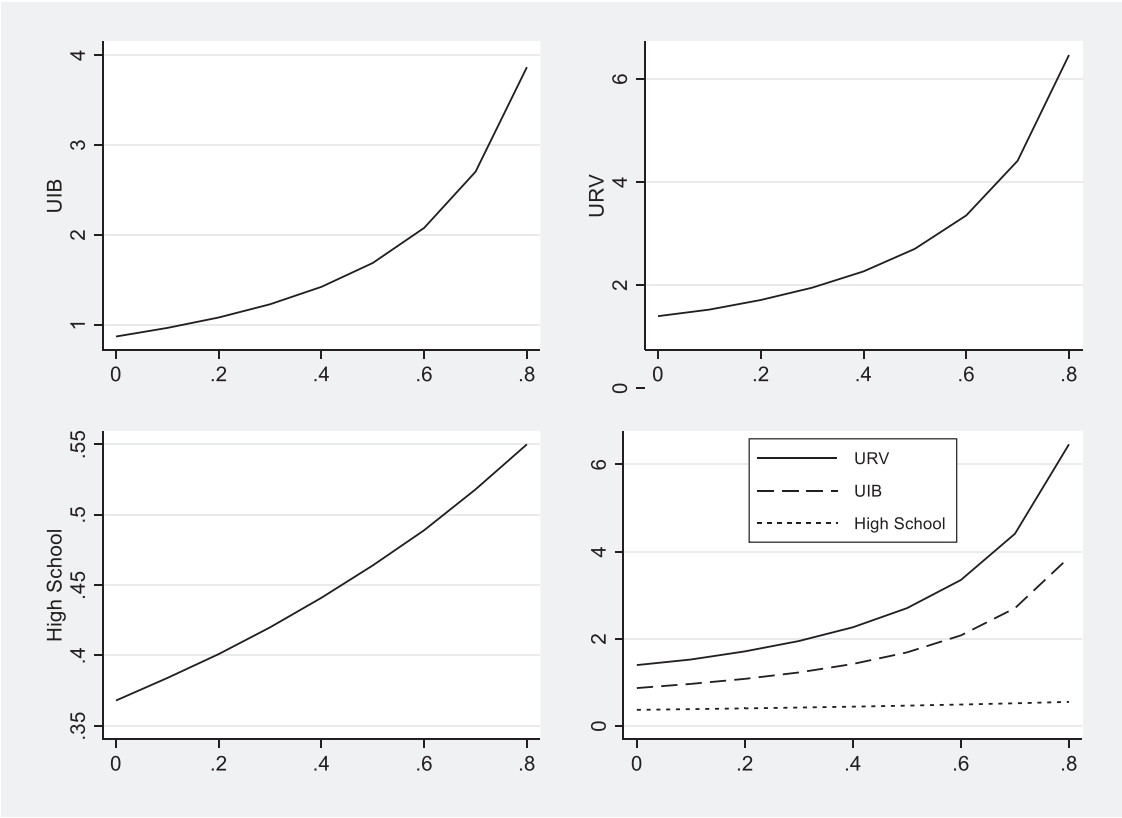

Fig. 2. Estimated coefficients from different measurement error variances in *cme* models
Note: Coefficients in absolute value.

the logit model, and of 8.1 percentage points with the *cme* model. In the URV sample, with a 69% of students attending the management degree, the corresponding probability increases are of 6.2 and 11.2 percentage points with the logit and the *cme* model, respectively. In these two samples, the bias and the probability changes are much higher that with the sample of high school students. Note that in the UIB and the URV samples, the estimated reliability for the data used in the regressions is 0.53, which is significantly smaller than the estimated reliability in the sample of high school students, 0.79.

We cannot be sure that the measurement error variance in the data used in the example is the same as that in the data that we used in our test–retest analysis; therefore, we carry out an additional sensitivity analysis to cover a wider range of possible measurement error variances. We re-estimate the covariate measurement error model (*cme*) using values of the measurement error variance ranging from 0.01 to 0.08, with steps of 0.01. The estimated coefficients, together with the coefficient from the conventional logit model, are depicted in Fig. 2.

This example shows how to take measurement error into account in a sensitivity analysis and highlights that the magnitude of the estimated effect can be very different from that in a model that assumes log earnings expectations measured without errors.

5.3. *Extensions: nonclassical measurement errors?*

Most of the literature on measurement errors, like our study, has relied on the assumption of a classical measurement error, that is, that the measurement error is not correlated with the latent true variable. While the existence of classical measurement errors in earnings expectations is very likely, the possibility of nonclassical measurement errors could also be plausible. For example, it could be that earnings expectations suffer from mean reversion, which is a type of nonclassical measurement error. Students might answer untruthfully, such that students with "truly" high earnings expectations will tend to understate their answer, thus appearing to be modest, while students with "truly" low earnings expectations will tend to overstate their answers to appear more confident. The literature has provided some methodologies that would allow us to test for the potential existence of nonclassical measurement errors in earnings expectations if this was the case.[14]

Studies analysing the existence of measurement errors in individuals' reported income in surveys have generally relied on a validation sample, in which the true variable is measured together with the error-prone measurement (e.g., Bingley and Martinello 2017, Bound et al. 1994, Chen et al. 2005). In Bound et al. (1994) and Chen et al. (2005) it is assumed that the validation variable does not include measurement errors. However, Bingley and Martinello's (2017) study is different since they allow for classical measurement error in the validation variable. They support the assumption of classical measurement error of the validation sample because the information is based on third-party reports, that is, reports originating from employers, banks and so on. Their results reported a mild measurement error in the *validation* variable for income, and the measurement error in the survey data was found to be classical for the annual gross income. Bingley and Martinello (2017) also showed that, when the validation data are (incorrectly) assumed to be free from measurement error, a nonclassical measurement error in the survey data appears.

There is an important difference between earnings expectations and realized earnings. These "true earnings" are not available for earnings expectations, and the methods based on validation data are not an option in our case. In general, validation data are rarely available; therefore, more recent studies have proposed alternative identification strategies that could also be used with earnings expectations. Instrumental

[14] Chen et al. (2011) reviewed the recent literature on nonclassical measurement errors.

variables are one of the options (Hahn and Ridder, 2017; Hu and Schennach, 2008). Hu and Schennach (2008) suggested an instrumental variable strategy, with an additional assumption that, conditional on the value of the true regressor, a measure of the location of the measurement error is equal to zero. The idea is that, even in situations in which the conditional mean is not equal to zero, it could be the case that, for example, the conditional median is equal to zero. An advantage of their method is that it can handle nonclassical measurement errors in "most widely used models, including probit, logit, tobit, and duration models …". The Hahn and Ridder (2017) instrumental variable solution is based on a two-step control variable estimator that can handle both nonclassical measurement errors and endogeneity problems for the same explanatory variable. This methodology requires "an exogenous shifter" as a valid instrument, which should also be independent of the measurement errors, equation errors and first-stage errors. We note that, in Hu and Schennach's (2008) methodology, a repeated measure could be used as an instrumental variable, whereas this is not a feasible instrument in Hahn and Ridder's (2017) solution.

Another practicable solution was proposed by Carroll et al. (2010). They suggested an identification strategy for nonclassical measurement errors using two samples, neither of which needs a correct measurement of the variable. They assumed that the samples differ with respect to an accurately measured discrete covariate and that the marginal distribution for the latent true variable is different. The latent model of interest was, however, assumed to be the same for the two samples.

Hu et al. (2022) considered a nonparametric regression model with nonclassical measurement errors in an explanatory variable. The exposition of the model only includes one explanatory variable, but additional error-free covariates can be included in practice. Identification is obtained through "monotonicity of the regression function, independence of the regression error, and completeness of the measurement error distribution". The monotonicity implies that that the fitted regression line should be either non-decreasing (or non-increasing) everywhere. The independence of the regression error of the latent regressor, X, and its measurement, W, is a standard assumption of nondifferential error, which implies that, knowing the true variable, the error-prone measurement does not provide any additional information about the dependent variable. The assumption of completeness implies that the conditional density, $f_{w|x}$, is complete in both w and x, which is a weaker assumption than the independence between x and $w - x$. The model proposed by Hu et al. (2022) is, accordingly, identified without the need for any of the most common identification strategies: i) a validation sample with the true measurement; ii) a secondary measurement of the error-prone variable; iii) an instrumental variable; and iv) an auxiliary sample.

The models that are briefly described above show that there are feasible options to allow for nonclassical measurement errors in studies of earnings expectations. Note, however, that knowledge of the measurement error variance from an external source is not enough for identification if the measurement error is nonclassical. The implementation requires careful planification in the data collection process to fulfil the assumptions. Carroll et al. (2010), Hu and Schennach (2008) and Hu et al. (2022) estimated their models using "sieve" maximum likelihood. The "sieve" method refers to maximizing, while keeping a subspace of the parameters constrained, and then relaxing the constraint as the sample size increases (Geman and Hwang, 1981). Statistical software usually includes maximum likelihood and the option to impose constraints, but there is still an important practical hurdle when these methods are not directly available.

6. Conclusions and discussion

The interest in eliciting expectations has increased in the economics literature, however, very little attention has been paid to the reliability of these measures. Our study focused on the reliability of students' log earnings expectations, and a test–retest evaluation was performed. Our test–retest analysis revealed that the reliability of earnings expectations is low, which suggests that the problem of measurement errors is important and it should be considered.[15]

Transient influences are, of course, difficult to capture in survey questions and, for the researcher, this will cause the variable to be measured with a random error. Students can change their earnings expectations if new information arrives between sessions. Students were asked about their earnings expectations three times., and in the second wave they were asked whether they had discussed the question regarding earnings expectations with their peers or family. We found that the reliability for the group of students who provided a positive answer to this question was the same as that estimated in the main sample. The third reason for changed expectations could be that, after facing the survey question for the first time, the respondents may make a more careful consideration of the answer in subsequent waves of the survey through a more in-depth analysis when reading the same question again. If students "learn" how to answer the earnings expectation questions, the extent of measurement errors should decline and the measured reliability should increase across successive survey waves. This is, however, not the case in our data, and random measurement errors are equally relevant after facing the survey question once more.

While it would be ideal for every study to perform its own evaluation of reliability, we are aware that this is generally not feasible. In these cases, we suggest using a measure of reliability (or quantification of the variance of the measurement error) from a study in which this problem can be assumed to be similar. To show how this could be achieved, we used our reliability analysis of earnings expectations carried out with data from university students enrolled in the University of the Balearic Islands to correct the potential measurement error in earnings expectations from data on secondary education students. The necessary assumption is that the variance of the measurement error in the two studies is similar. We used a binary logit model (expectations of achieving a college degree in the future) as an example. Our results reveal that, if the measurement error in earnings expectations is not accounted for, the estimated coefficient associated with this variable is biased downwards.

The findings reported throughout this analysis suggest that the results associated with the use of earnings expectations should be taken with some caution. Sensitivity analysis based on estimates or assumptions about the variance of the measurement error should be a natural component of research using this variable. However, it is important to consider the population of interest. Earnings expectations for secondary education students are expected to be more widely dispersed, which implies that the measurement error problem will be less severe than in a sample extracted from a population of college students enrolled in the same field of study (Eq. (3)). The variance in log earnings expectation is, for example, 0.48 for students in secondary education compared with 0.38 observed for the first wave in the UIB. In this context, the variance of the measurement error could also be different for the two groups. Hopefully, measures of the variance of the measurement error and reliability in different studies will be collected in different situations, which will provide a more complete picture of the measurement error in students' log earnings expectations. These measurements are important starting values for a sensitivity analysis in other empirical studies using log earnings expectations.

Funding

The authors acknowledge financial support from the *Obra Social "La Caixa"* (grant # 2014ACUP0130) and from the Spanish Ministry of Science and Innovation through Grant # RTI2018-094733-B-I00.

[15] It is interesting to note that the reliability of earnings expectations we find here is fairly similar to that Krueger and Schkade's (2008) find for life satisfaction.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgements

We would like to thank the comments of the anonymous reviewers. We are particularly grateful to Prof. Wilbert van der Klaauw, who has crucially contributed with his comments to shape the final version of this article

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.labeco.2022.102182.

References

Attanasio, O.P., Kaufmann, K.M., 2014. Education choices and returns to schooling: Mothers' and youths' subjective expectations and their role by gender. J. Dev. Econ. 109, 203–216.
Amemiya, Y., 1985. Instrumental variable estimator for the nonlinear error-in-variables model. J. Econom. 28, 273–289.
Arcidiacono, P., Hotz, V.J., Kang, S., 2012. Modeling college major choices using elicited measures of expectations and counterfactuals. J. Econom. 166, 3–16.
Attanasio, O.P., Kaufmann, K.M., 2017. Education choices and returns on the labor and marriage markets: evidence from data on subjective expectations. J. Econ. Behav. Organ. 140, 35–55.
Belfield, C., Boneva, T., Rauh, C, Sha, J., 2016. Money or fun? Why students want to pursue further education, IZA Discussion Paper 10136 Bonn, Germany.
Bingley, P., Martinello, A., 2017. Measurement Error in Income and Schooling and the Bias of Linear Estimators. J. Lab. Econ. 35 (4), 1117–1148.
Bleemer, Z., Zafar, B., 2018. Intended college attendance: Evidence from an experiment on college returns and costs. J. Public Econ. 157, 184–211.
Boneva, T., Golin, M., Rauh, C., 2021. Can perceived returns explain enrollment gaps in postgraduate education? Lab. Econ. In press.
Bound, J., Brown, C., Duncan, G.J., Rodgers, W.L., 1994. Evidence on the validity of cross–sectional and longitudinal labor market data. J. Lab. Econ. 12 (3), 345–368.
Bound, J., Krueger, A.B., 1991. The extent of measurement error in longitudinal earnings data: do two wrongs make a right? J. Lab. Econ. 9 (1), 1–24.
Brunello, G., Lucifora, C., Winter-Ebmer, R., 2004. The wage expectations of European business and economics students. J. Hum. Resour. 39, 1116–1142.
Caliendo, M., Lee, W.S., Mahlstedt, R., 2017. The gender wage gap and the role of reservation wages: new evidence for unemployed workers. J. Econ. Behav. Organ. 136, 161–173.
Carmines, E.G., Zeller, R.A., 1979. Reliability and Validity Assessment. Sage University Papers series on Quantitative Applications in the Social Sciences, 07-017. SAGE Publications, Beverly Hills and London.
Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., 2006. Measurement Error in Nonlinear Models: A Modern Perspective, 2nd ed. CRC Press.
Carroll, R.J., Chen, X., Hu, Y., 2010. Identification and estimation of nonlinear models using two samples with nonclassical measurement errors. J. Nonparametr. Stat. 22 (4), 379–399.
Chen, X., Hong, H., Tamer, E., 2005. Measurement error models with auxiliary data. Rev. Econ. Stud. 72, 343–366.
Chen, X., Hong, H., Nekipelov, D., 2011. Nonlinear models of measurement errors. J. Econ. Lit. 49 (4), 901–937.
Delavande, A., Giné, X., McKenzie, D., 2011. Eliciting probabilistic expectations with visual aids in developing countries: How sensitive are answers to variations in elicitation design? J. Appl. Econ. 26, 479–497.
Dominitz, J., 1998. Earnings expectations, revisions, and realizations. Rev. Econ. Stat. 80, 374–388.
Dominitz, J., Manski, C.F., 1996. Eliciting student expectations of the returns to schooling. J. Hum. Resour. 31, 1–26.
Geman, S., Hwang, C.R., 1981. Nonparametric maximum likelihood estimation by the methods of sieves. Ann. Stat. 10 (2), 401–414.
Gillen, B., Snowberg, E., Yariv, L., 2019. Experimenting with measurement error: techniques with applications to the caltech cohort study. J. Polit. Econ. 127 (4), 1826–1863.
Giustinelli, P., Manski, C.F., Molinari, F., 2020. Tail and center rounding of probabilistic expectations in the health and retirement study. J. Econom. In press.
Gottschalk, P., Huynh, M., 2010. Are earnings inequality and mobility overstated? The impact of nonclassical measurement error. Rev. Econ. Stat. 92, 302–315.
Gouret, F., 2017. What can we learn from the fifties? J. Forecast. 36, 756–775.
Gouret, F., Hollard, G., 2011. When kahneman meets Manski: using dual systems of reasoning to interpret subjective expectations of equity returns. J. Appl. Econom. 26, 371–392.
Hahn, J., Ridder, G., 2017. Instrumental variable estimation of nonlinear models with nonclassical measurement error using control variables. J. Econom. 200, 238–250.
Hastings, J., Nielson, C.A., Zimmerman, S.D., 2015. The Effect of Earnings Disclosure on College Enrollment Decisions. NBER Working Paper, p. 21300.
Hartog, J., Diaz-Serrano, L., 2013. Schooling as a risky investment: a survey of theory and evidence. Found. Trends Microecon. 9 (3-4), 159–331.
Hu, Y., Schennach, S.M., 2008. Instrumental variable treatment of nonclassical measurement error models. Econometrica 76 (1), 195–216.
Hu, Y., Schennach, S., Shiu, J.L., 2022. Identification of nonparametric monotonic regression models with continuous nonclassical measurement errors. J. Econom. 226, 269–294.
Huntington-Klein, N., 2015. Subjective and projected returns to education. J. Econ. Behav. Organ. 117, 10–25.
Jensen, R., 2010. The (Perceived) returns to education and the demand for schooling. Q. J. Econ. 125 (2), 515–548.
Kimball, M.S., Sahm, C.R., Shapiro, M.D., 2008. Imputing risk tolerance from survey responses. J. Am. Stat. Assoc. 103 (483), 1028–1038.
Kristensen, N., Westergaard-Nielsen, N., 2006. Reliability of job satisfaction measures. J. Happiness Stud. 8 (2), 273–292.
Krueger, A.B., Schkade, D.A., 2008. The reliability of subjective well-being measures. J. Public Econ. 92, 1833–1845.
Manski, C.F., 1999. Analysis of choice expectations in incomplete scenarios. J. Risk Uncertain. 19 (1/3), 49–66.
Manski, C.F., 2004. Measuring expectations. Econometrica 72, 1329–1376.
Manski, C.F., Molinari, F., 2010. Rounding probabilistic expectations in surveys. J. Bus. Econ. Stat. 28, 219–231.
Nunnally, J.C., 1975. Psychometric THEORY. 25 years ago and now. Educ. Res. 4 (10), 19–21 7-14+.
Rabe-Hesketh, S., 2003. CME: Stata program to Estimate Generalized Linear Models with Covariate Measurement Error. Statistical Software Components S434701. Boston College Department of Economics revised 04 Sep 2004.
Rabe-Hesketh, P., S, A., Skrondal, A., 2003a. Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. Stat. Model. 3, 215–232.
Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2002. Reliable estimation of generalized linear mixed models using adaptive quadrature. Stat. J. 2 (1), 1–21.
Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2003b. Maximum likelihood estimation of generalized linear models with covariate measurement error. Stat. J. 3 (4), 386–411.
Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2004. Generalized multilevel structural equation modeling. Psychometrika 69 (2), 167–190.
Reuben, E., Wiswall, M., Zafar, B., 2015. Preferences and biases in educational choices and the labour market expectations: shrinking the black box of gender. Econ. J. 127, 2153–2186.
Schweri, J., Hartog, J., 2017. Do wage expectations predict college enrollment? Evidence from healthcare. J. Econ. Behav. Organ. 141, 135–150.
Stefanski, L.A., Carroll, R.J., 1985. Covariate measurement error in logistic regression. Ann. Stat. 13 (4), 1335–1351.
Vacha-Haase, T., Henson, R.K., Caruso, J.C., 2002. Reliability generalization: moving toward improved understanding and use of score reliability. Educ. Psychol. Meas. 62 (4), 562–569.
Van Santen, P., Alessie, R., Kalwij, A., 2012. Probabilistic survey questions and incorrect answers: retirement income replacement rates. J. Econ. Behav. Organ. 82, 267–280.
Webbink, D., Hartog, J., 2004. Can students predict their starting salaries? Yes!. Econ. Educ. Rev. 23, 103–113.
Wiswall, M., Zafar, B., 2015. How do college students respond to public information about earnings? J. Hum. Cap. 9, 117–169.
Zafar, B., 2011. Can subjective expectation data be used in choice models? Evidence on cognitive biases. J. Appl. Econ. 26, 520–544.
Zafar, B., 2013. College major choice and the gender gap. J. Hum. Resour. 48 (3), 545–595.