UNIVERSITAT ROVIRA I VIRGILI
DEPARTAMENT D'ECONOMIA

CREIP
CENTRE DE RECERCA EN ECONOMIA INDUSTRIAL I ECONOMIA PÚBLICA

WORKING PAPERS

Col·lecció "DOCUMENTS DE TREBALL DEL DEPARTAMENT D'ECONOMIA - CREIP"

The Reliability of Students' Earnings Expectations

Luis Diaz Serrano
William Nilsson

Document de treball n.04- 2017

DEPARTAMENT D'ECONOMIA – CREIP
Facultat d'Economia i Empresa

UNIVERSITAT ROVIRA I VIRGILI

DEPARTAMENT D'ECONOMIA

Edita:

Departament d'Economia
www.fcee.urv.es/departaments/economia/public_html/index.html
Universitat Rovira i Virgili
Facultat d'Economia i Empresa
Av. de la Universitat, 1
43204 Reus
Tel.: +34 977 759 811
Fax: +34 977 758 907
Email: sde@urv.cat

CREIP
www.urv.cat/creip
Universitat Rovira i Virgili
Departament d'Economia
Av. de la Universitat, 1
43204 Reus
Tel.: +34 977 758 936
Email: creip@urv.cat

Adreçar comentaris al Departament d'Economia / CREIP

ISSN edició en paper: 1576 - 3382
ISSN edició electrònica: 1988 - 0820

DEPARTAMENT D'ECONOMIA – CREIP
Facultat d'Economia i Empresa

The Reliability of Students' Earnings Expectations

Luis Diaz-Serrano* and William Nilsson**

Abstract: Eliciting expectation and introducing probabilistic questions into surveys have gained important interest. In this study, we focus on the reliability of students' earnings expectations. To what extent is observed log earnings expectations affected by random measurement error (noise)? A test-retest method is applied and reliability is found to be fairly low; about 0.59 in 2015 and about 0.67 in 2016. Particularly homogeneous samples exaggerate problems of measurement error. The analysis show how these measures of reliability can be adjusted to become more suitable to other studies, where different degrees of homogeneity are present.

JEL: C46; C83; I26

Keywords: Earnings expectations, reliability, measurement error.

* CREIP - Universitat Rovira i Virgili. Department of Economics, Av. de la Universitat, 1, 43204 Reus, Spain. E-mail: luis.diaz@urv.cat

** Contacting author: University de les Illes Balears, Department of Applied Economics, Ctra Valldemossa Km 7,5, E-07122 Palma de Mallorca, Spain. Phone: +34 971 17 13 77, Fax: +34 971 17 23 89, E-mail: william.nilsson@uib.es

The authors acknowledge the financial support from the Spanish Ministry of Science and Innovation (ECO2014-59055-R).

1. Introduction

Consider that a researcher would like to evaluate the decision that students face after finishing secondary schooling. They can enter the labour market or continue with university studies, and accordingly they have to choose what to study. One variable that could matter for such decision is their earnings expectations for the different alternatives. The researcher would like to ask the students for their earnings expectation exactly when the decision was taken. However, all of the students do not make their decision at the same time, and the researcher would have to decide when to ask the students. Let us say that the students are asked one month before the application must be filled in. Some of the students might have already made their decision, and others may still not have decided and the collected earnings expectations will be a proxy for earnings expectations that were used, or will be used, to finally make the decisions. If students have clearly defined earnings expectations these will not change much over a short period of time and even though a proxy is used it will be highly correlated with the true expectations that are used for the decision. If, on the other hand, students have very difficult to form their expectation, their declared expectation could vary a lot over time, even though the students do not receive new information that causes their actual expectations to change.

The interest to elicit expectations has increased and several studies report positive results on how such questions can provide valuable information (Dominitz & Manski, 1996, Manski, 2004). The questions are usually formulated to incorporate not only expectations but also subjective evaluations on how certain these expectations are. The literature on expectations mentions the issue of measurement error (Dominitz, 1998 and Manski, 2004), but the avenue to deal with the problem has been to reformulate and improve questions to increase answering frequencies and avoid logical inconsistencies.

The focus has been to evaluate the *validity* of the questions, i.e. to what extent the given answer captures the concept that it is aimed to measure? Basically, are respondents willing and able to respond in a meaningful way? Zafar (2011) analyze students' subjective earnings expectations and cognitive biases and concludes "I do not find evidence of cognitive biases systematically affecting the reporting of beliefs". Again, the concern was the validity of the expectations.

The literature acknowledges and includes subjective measures of the uncertainty in the expectation, but how much uncertainty that is found in the answering itself has not been considered. This is the focus in the present study. We are interested to evaluate the *reliability* of students' earnings expectations. The reliability of a measure can be quantified using repeated measures of the same phenomena. The idea is to see to what extent the same result is obtain if the measuring procedure is repeated once, or several times, during a short period when true expectations should remain the same. If repeated answers are very similar the measure has high reliability, but if large changes are found, the reliability is low due to random measurement error. The concern is not that a bias is affecting the answer, but rather that the answer is affected by random noise. Wiswall & Zafar (2015) use repeated measures to calculate a reliability ratio, i.e. the ratio of true variance to observed variance, but the repeated measures are collected on the same occasion and it is likely that the reliability is severely overestimated simply because individuals can remember their firstly stated expectation.

The purpose with this study is to evaluate the degree of measurement error in earnings expectations for a specific group of university students. The measurement error is quantified using a test-retest method and the information is useful not only for the present study but also for other studies on students' earnings expectations. The main data covers students from the University of the Balearic Islands, UIB, that were enrolled

in the course Analysis of Economic Data in 2015. The sample is, accordingly more homogenous than a random sample of university students would be. Earnings expectations were asked at three times where the second session and third session were about 14 days respective 74 days after the first session. The time period is short, and little new information should arrive in that period. The reliability is found to be fairly low and this implies important concerns on using earnings expectations in research.

The study contributes to the literature in several important aspects. First, reliability is quantified and random measurement error is found to be important in students' earnings expectation. Second, the study clarifies how homogenous samples interact with random measurement error and exaggerate its consequences. In addition, we also suggest how to adapt the measure or reliability to different degree of homogeneity in the current study and other studies of log earnings expectations. The study provides a very clear message: studies of log earnings expectations should either incorporate an analysis of its reliability or, if this is, for some reason, not possible, the analysis should be adapted by using a homogeneity adjusted measure of reliability from a study where the measurement error is likely to be similar.

Section 2 discusses related literature and section 3 introduces a theoretical model to clarify the problem of measurement error and the consequences of using a homogenous sample when the coefficient of correlation is calculated. The model is general and different degrees of measurement error and homogeneity is allowed for both variables. The section also reviews how to evaluate reliability in a measure and how this information can be used in other studies. Section 4 explains the data and section 5 evaluate the degree of measurement error found in earnings expectations as well as the stated probability to either end up having earnings below 50% or above 150% of the earnings expectation.

2. Related literature

The literature on expectation and the use of probabilistic questions has paid a lot of attention on the validity of the measurements; are the questions understood correctly, are answers logically consistent, do answers bunch around specific values, and so on. For example, initial concerns on asking for probabilities were that individuals would tend to choose from a few probabilities, for example, 0%, 50% and 100%. The literature has, overall discarded these concerns and the use of probabilistic questions is becoming more common in questionnaires. Manski & Molinari (2010) analyse how different respondents can use different degrees of rounding practice and that pattern can be detected using different survey questions on the probability that a future event will occur. Delavande et al. (2011) analyse how different elicitation designs using visual aids can help respondents to express probabilistic concepts.

Wiswall & Zafar (2015) is, to our knowledge, the only article that measures a reliability of expectations. The questions are, as mentioned above, repeated on the same occasion and while it is sufficient for their purpose, it is not a valid test of the reliability of expectations. Apart from this case, we have not found any study on the reliability of the measurements. Gouret & Hollard (2011) do not measure reliability, but they develop a measure of respondents' coherence which can be used to distinguish noisy data from more valuable data. The idea is to retain the data that is expected to have higher reliability, based on more coherent answering behaviour. Van Santen et al. (2012) distinguish between respondents that answer inconsistently, but they show that simply excluding these cases implies an endogenous sample selection problem.

While not measuring reliability, there are a few studies on how transitory or persistent expectations are over time. For example, Dominitz & Manski (2003) analyse repeated questions on expectations in a working paper, but in the published version this

analysis is excluded. They ask about the percentage chance (probability) that the individual would increase his income in the next 12 month compared to the previous 12 months. The same question is repeated 6 months later for a sub-sample. They perform linear auto-regressions, which are interpreted as showing "substantial predictive power". The slopes in these regressions are between 0.43 and 0.63 for 6 different periods and 0.53 for the pooled sample. They do not show descriptive statistics for the sub sample with repeated questions, but if the standard deviations are fairly similar (as it is for the complete sample) for the different periods the slope is close to the coefficient of correlation. The purpose of their analysis is not to study measurement error, but to evaluate temporal fluctuations in the variable. With a period of six months the correlation is expected to be lower than 1 due to both measurement error and actual changes in the response due to actual changes in the expectations.

Zafar (2011) analyze students' subjective earnings expectations and repeated the questions about a year later. The data set is available on Journal of Applied Econometrics Data Archive and correlation coefficients for different earnings expectation based on different scenarios can be found in Table A1 in Appendix. A coefficient of correlation of about 0.6 represents a coefficient of determination of 0.36, and hence a quite large share of the variation in the earnings expectation the second year cannot be explained by knowing the answer to the same expectation a year earlier. The sample sizes are, however, fairly small and a large variation in correlation coefficients is found for different majors.

The earnings expectations can change quite a lot from one year to another, and while the low correlation does not necessarily indicate measurement error, it gives the impression of students' earnings expectations to be perishable items. Using earnings expectations measured a year before the decision can be a poor proxy for true earnings

expectations. It is possible that earnings expectations are measured with error in both periods, which would imply a biased regression coefficient as well as coefficient of correlation. If earnings expectations have classical measurement error the temporal stability measured with the regression coefficient or the coefficient of correlation will be underestimated. Below we consider the effect of measurement error and the use of homogeneous samples on measuring temporal stability of earning expectations using the coefficient of correlation. The analysis is, however, general and it is, of course, also useful for analysing two different variables.

3. Measurement error and homogenous samples

The purpose of this section is to clarify how the coefficient of correlation is affected by measurement error interacted with using a particularly homogenous sample. The model is general and different degree of measurement error and also different degree of homogeneity compared to the population is allowed for the two variables. Solon (1989) studies the interaction of these two sources of bias, but in a context where the sample either is homogeneously chosen based on X or Y. When studying temporal stability of earnings expectations the sample will often be homogenous with respect to both variables. Gottschalk and Huynh (2010) studies how measurement error affects earnings inequality and earnings mobility when measurement error can be found in both periods, but they do not consider the effects of working with a homogenous sample. While they consider both non-classical measurement error and classical measurement error the focus below is on classical measurement error.

Conceptually each student has a "true" earnings expectation which is unobservable and cannot be measured directly. If we would be able to ask each student an infinite amount of times, and after answering, he/she would instantly forget his/her

answer, we could calculate the average expectation to find the true earnings expectation. This is, however, not possible and true earnings expectations remain unobserved. Accordingly, if a student is asked about his earnings expectation, a random measurement error sometimes makes the answered expectation to be above, and in other occasions below, the true score. We are interested in the relationship between E^* and E^*_{-1},

$$E^* = \beta E^*_{-1} + \varepsilon \quad (1)$$

Where E^* and E^*_{-1} refer to "true" log earnings expectations measured in two different occasions in time and ε is an error term. Both variables are measured as deviations from their means. The slope, β, is the intertemporal elasticity of earnings expectations. The "true" log earnings expectations are not observed and instead E and E_{-1} are collected in each period.

$$E = E^* + \upsilon$$

$$E_{-1} = E^*_{-1} + \mu$$

Where υ and μ are measurement errors assumed to be neither correlated with each other nor ε, E^* or E^*_{-1}. For example, the covariance, $\sigma_{E^*\upsilon}$, and the covariance, $\sigma_{E^*_{-1}\mu}$, are both zero. We are particularly interested in the correlation of log earnings expectations in the two periods,

$$\rho = \beta \frac{\sigma(E^*_{-1})}{\sigma(E^*)}$$

For this case of classical measurement error Gottschalk and Huynh (2010) show that the estimated correlation, $\hat{\rho}$, underestimates the actual correlation, ρ, according to;

$$\hat{\rho} = \rho \frac{\sigma(E^*_{-1})\sigma(E^*)}{\sigma(E_{-1})\sigma(E)} < \rho \qquad (2)$$

This refers to cases where the sample is not homogenous (compared to the population) and since we are interested in the interaction effect it is necessary to adapt the expression. Note that $\sigma^2(E) = \sigma^2(E^*) + \sigma^2(\upsilon)$, because the covariance, $\sigma_{E^*\upsilon}$, is cero by assumption. In the same way $\sigma^2(E_{-1}) = \sigma^2(E^*_{-1}) + \sigma^2(\mu)$. Accordingly, we prefer to express the formula in terms of variances to be able to clarify how a homogenous sample, for example, $s^2(E^*_{-1}) < \sigma^2(E^*_{-1})$, would affect the estimated correlation. $s^2(.)$ is the sample estimate of the population variance $\sigma^2(.)$, which will be lower if the sample is incorrectly collected form a particular homogeneous group.

$$\hat{\rho}^2 = \rho^2 \frac{\sigma^2(E^*_{-1})\sigma^2(E^*)}{\sigma^2(E_{-1})\sigma^2(E)} = \rho^2 \frac{\sigma^2(E^*_{-1})}{\left(\sigma^2(E^*_{-1}) + \sigma^2(\mu)\right)} \times \frac{\sigma^2(E^*)}{\left(\sigma^2(E^*) + \sigma^2(\upsilon)\right)}$$

And,

$$\hat{\rho} = \rho\left[\frac{\sigma^2(E^*_{-1})}{\left(\sigma^2(E^*_{-1})+\sigma^2(\mu)\right)} \times \frac{\sigma^2(E^*)}{\left(\sigma^2(E^*)+\sigma^2(\upsilon)\right)}\right]^{0.5} \quad (3)$$

The *reliability* of E to measure E^* is the ratio of true variance to observed variance. We find that the attenuation factor is a geometric mean of the reliability for the two measurements, E_{-1} respective E. If a random sample is used instead of the entire population, sample variances replace the population variances in expression (3). A homogenous sample, where $s^2(E^*_{-1}) < \sigma^2(E^*_{-1})$ and $s^2(E^*) < \sigma^2(E^*)$ will exacerbate the measurement error and the intertemporal correlation will be even more underestimated. A homogenous sample makes the effect of measurement error more severe because the signal is lower than what is found in the population. Section 5.1.3 discusses why the problem of a homogenous sample can be of particular relevance in the case of students' earnings expectations.

3.1 Quantifying reliability and correcting for random measurement error

The reliability of a measure can be obtained by calculating the correlation between two parallel measures (Carmines & Zeller, 1979). Parallel measures have the same true underlying score and equal variances. The measurement errors found in different parallel measures are assumed to not be correlated.

$$\rho_{EE'} = \frac{\sigma_{EE'}}{\sigma_E \sigma_{E'}} = \frac{\sigma_{(E^*+\upsilon)(E^*+\upsilon')}}{\sigma_E \sigma_{E'}} = \frac{\sigma^2(E^*)}{\sigma^2(E)} \quad (5)$$

If E is a measure of log earnings expectations, and E' is a parallel measure the covariance is equal to the variance of the "true" log earnings expectations because the

measurement error υ are not correlated with each other and nor with E^*. Hence, the reliability for E in equation (3) can be found by calculating the correlation of two parallel measures E and E'. If E_{-1} is a different variable the correlation of two parallel measures E_{-1} and E'_{-1} would identify its reliability. If E and E_{-1} refers to the same variable and if we assume that the reliability is the same for the two periods it would be enough to obtain a parallel measure to one of the time periods.

The idea behind the test-retest method is to obtain two parallel measures, that is, to perform the test once and then repeat the same test at least one more time. The appropriate time between the measurements is, however, difficult to know. If the retest is repeated too soon, reliability can be overestimated because individuals can remember their previous answer. On the other hand, if too much time is left until the retest it is possible that the underlying true score has changed, in which case the reliability is underestimated. The trade-off when asking for earnings expectations is that students can remember their first answer or that their true expectation actually has changed. Another critique is that reliability can seem to be low because of a reactivity problem (Carmines & Zeller, 1979). Measuring a concept once can cause a change in the true score of that concept. Carmines & Zeller, 1979 specify some properties that parallel measure should have; (1) The expected value and variance of parallel measures are equal; $E(E) = E(E')$ and $\sigma^2(E) = \sigma^2(E')$, (2) The correlation of several parallel measures are equal for different pairs; $\rho_{EE'} = \rho_{EE''} = \rho_{E'E''}$ and (3) The correlation of parallel measures and other variables (for example Y) are equal; $\rho_{EY} = \rho_{E'Y} = \rho_{E''Y}$. Once data is collected it is of course possible to test if these assumptions are fulfilled. The advantage of calculating the reliability is that this information can be used to "correct" for attenuation as showed in Carmines & Zeller (1979).

$$\rho = \hat{\rho} / \sqrt{\rho_{E_{-1}E'_{-1}} \times \rho_{EE'}} \quad (6)$$

This is also evident using equation (3) and equation (5). Notice that the test-retest method does not distinguish between measurement error and the interaction of measurement error and a homogenous sample. Accordingly, the calculated reliability can be used to correct for attenuation bias for the same sample or for a sample that has the same overall effect of measurement error and homogeneity. The sample version of expression (6) is $r = \hat{r} / \sqrt{r_{E_{-1}E'_{-1}} \times r_{EE'}}$, i.e. the "true" correlation is the calculated correlation divided by the square root of the multiplied reliability measures. If a more heterogeneous sample is used the calculated reliability would overcompensate due to using a too small reliability. If an even more homogenous sample is used the correction would not be sufficient. Accordingly, it is interesting to clarify how the information in a test-retest analysis can be used to correct the bias in other studies with different degree of homogeneity.

3.2 Using quantified reliability to correct for measurement error in other studies

The best option is of course that each study includes its own calculation on the reliability which can be used in the analysis. In some cases this is not possible and treating the data as if no measurement error is present seems careless. The idea to use a correction factor is not new. For example, in the context of intergenerational earnings correlation, using earnings in a single year implies transitory variance apart from permanent earnings and a bias towards cero. Information on noise-to-signal from other sources can be used to correct this problem. (See, for example, Solon, 1989). If we assume that the interaction of measurement error and working with a homogenous sample is exactly the same, the calculated reliability can be used as it is. If we assume

that the variance of the measurement error is the same, it is possible to adapt the reliability for different degrees of homogeneity in the sample, and accordingly relaxing the previous mentioned assumption.

First, consider that we are able to collect a new random sample from the population, and in this case the sample variance is not underestimating the population variance due to a too homogenous sample, instead it is correctly chosen from the population. How should we adapt the reliability from the first study? We assume that the sample variance of the measurement error is an unbiased estimate of the population variance of the measurement error, hence, $s^2(\upsilon) = \sigma^2(\upsilon)$. From the study on reliability we calculate the variance of the true score, $s^2(E^*) = s^2(E)r_{EE'}$, i.e. the true variance is equal to the observed variance multiplied by the reliability, which we can see in equation (5). Knowing that $s^2(E) = s^2(E^*) + s^2(\upsilon)$, we obtain $s^2(\upsilon)$. From the sample with the correct degree of homogeneity we use $s^2(E) = s^2(E^*) + s^2(\upsilon)$ to calculate $s^2(E^*)$ which together with $s^2(\upsilon)$ can be used directly in (3) to obtain the correct reliability for E in the new sample. Hence, the problem of homogeneity is no longer included in the adjusted measure of reliability. The same argument can be used to correct the reliability for E_{-1}.

Next, consider that a new random sample is selected but we are still not able to choose a sample with the correct degree of homogeneity. Can we adapt the reliability from the first study to this case? Yes, the procedure is actually the same, but the notation will be slightly different. From the second sample, labelled B, we use $s_B^2(E) = s_B^2(E^*) + s^2(\upsilon)$ to calculate $s_B^2(E^*)$ that is used to substitute $\sigma^2(E^*)$ in equation (3). $s^2(\upsilon)$ is used as before and, accordingly, the reliability is adapted and possible to use for sample B. If $s_B^2(E^*) < s^2(E^*)$, the second sample is more homogeneous and the

measure of reliability will be reduced, while if $s_B^2(E^*) > s^2(E^*)$ the reliability will be higher because sample B is more heterogeneous.

Notice that we assume that $s^2(\upsilon) = \sigma^2(\upsilon)$ and also for sample B. The degree of measurement error can, however, be different due to many reasons, for example, different phrasing of the question, different testing environment, different characteristics of the individuals that answer etc. If more test-retest studies are performed this can shed light on which situations that are accompanied with more or less measurement error.

While this is a case study, section 5.2 offers average log earnings expectation, reliability and variance of measurement error. This enables correcting measurement error in other studies.

4. Data collection

The data for the analysis was collected from students in the course Analysis of Economic Data, University of the Balearic Islands, in 2015. The course was in 2015 included in three different university studies, Degree in Economics, Degree in Business Administration and Double degree: Degree in Business Administration and Law. In 2016 another round was collected to complement the analysis. The more recent data is analyzed in section 5.4.

The first wave was done in a computer-lab on class hour during the first week of the course. The questionnaire was made in Google drive. Some students (41 out of 474) were unable to attend neither the class, nor a recovery class during the same week and they were allowed to answer outside class hour. This survey included many questions and some students needed a complete hour to fill in all the answers. In addition to questions relevant for this study, the survey included a wide variety of questions to

create a dataset that students later would use during the introductory course in statistics. The key questions for this study will be explained below.

The second wave was, for practical reasons, done outside class hour. The students were given a one week deadline to answer the survey from any computer with Internet connection. The time to answer was about 5 to 10 minutes. The students answered about 14 days after the first survey. A variable is available on how many days that had passed since they filled in the first survey. (The mean is 13.8 days and the standard deviation is 2.8). Both these surveys were mandatory to do for students that wanted to participate in an individual work which corresponds to 15% of the assessment of the course.

The third wave was also done outside class hour. The mean time is about 73.8 days after the first wave and the standard deviation is 3.3. Some of the professors used this survey as a way for students to register for a group work which corresponds to 15% of the assessment of the course. The amount of students answering this survey is accordingly lower for two reasons; a) students may have dropped out from the course and did not have the incentive to answer the survey because they simple had no plans to participate in the group work, or b) they were students in groups that did not use the survey as a mandatory requirement to participate in the group work.

As explained above, the second survey was done about 14 days after the first survey and the third survey was done about 74 days after the first survey. The timing of when to perform the re-test is difficult to know, and the concerns are that memory could overestimate the reliability from the first to the second survey, while actual changes could underestimate the reliability from the first survey to the third survey. In addition to this, the reactivity problem could underestimate the reliability, and an additional question is included to be able to evaluate this issue.

4.1 Variables

The key interest for the study is students' earnings expectations, but we also want to know how uncertain they consider these expectations to be. The survey is not interactive and adjustments to the Dominitz & Manski (1996) method are done to avoid logical inconsistencies.

4.1.1. Variables in the first wave

The first question on earnings perceptions is:

What do you think is the average gross monthly salary at age 45 for those who graduated on the studies that you currently are pursuing? Think of the country where you would seek a job, but specify the amount in Euros.

In the questionnaire we also allow students to reveal how certain they are about their perception, but since this is not the focus of the analysis in this study, we include these details in an Appendix. Once students have answered these questions we go on to ask about their expectation of their own earnings. We ask;

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

We also let the students reveal their uncertainty about their expectations. The set of questions were then repeated for different scenarios: randomly assigned university degree, with only secondary schooling, preferred degree (if it was different compared to current studies) and someone with age 45 that graduated with the randomly assigned

university degree. The list of degrees that were randomly assigned can be found in Appendix. We also asked students about their perceptions on their own mathematical, verbal, social and commercial skills and how they perceive themselves regarding risk taking. These questions are clarified in Appendix.

4.1.2. Variables in the second wave

Three different scenarios were included with the same sets of questions as explain above. First, instead of asking for someone with age 45 we asked for earnings perception for a student pursuing the same studies as the individual. The next scenarios are exactly the same as above, which are own earnings expectation on current studies and own expectations on the "randomly assigned study". The randomization only refers to the first survey, and the study should be the same in this occasion. The randomization was done by the last digit in a generated individual id-code. The students were asked to open a particular link to the questionnaire in Google drive based on this digit. Some students failed to choose the correct link in either the first or the second survey, but this is detectable in the data.

In the survey an additional question was added to know if the students had discussed their earning expectations with peers or family members or received new information. The translated question is;

"After answering Survey A, have you discussed or commented the earnings expectations with someone, for examples other students, friends or parents, or have you received other information that could have led to a change in your earnings expectations?"

The question is very general to include any interaction or information that "*could* have led to a change" in the earnings expectations. Hence, answering "No" should in principle, distinguish the students and the reliability could be calculated for this subgroup were the reactivity problem should be small, or non-existent. Basically, the question tries to identify the channels for a possible change in expectation, and beyond these channels only own reflection, without external information, is left that could have caused a changed expectation.

4.1.3. Variables in the third wave

The only relevant questions for this study is the repeated set of questions on own earnings expectations from graduating on the current studies.

4.2. Summary statistics

Descriptive statistics and an overview of which variables are found in the different waves can be found in table 1.

[Table 1]

The average earnings expectations on the current studies are about 1600€ per months, which correspond to about 7.31-7.33 in log earnings expectations for the different waves.

5. Results

5.1 Test-retest reliability on log earnings expectations

Table 2 includes a correlation matrix on test-retest reliability, and the values in parenthesis refer to sample size. The correlation coefficient is calculated both for a restricted sample where only students found in all three waves are included and an unrestricted sample where the correlation is calculated for those that answered the analyzed waves, but not necessarily the third.

[Table 2]

The correlation on own earnings expectations is about 0.59 when measured in the first wave and the second wave for the unrestricted sample. The same result is found when the correlation is calculated using the answer from the first and the third waves. Similar results are found for the restricted sample. This indicates a fairly low reliability and hence, measurement error is important. About 26% of the complete sample maintained their earnings expectation from the first to the second survey. About 27% kept the same answer from the first to the third and about 26% answered the same in the second and the third survey. Approximately 12% of those that answered all three surveys answered the same amount in all of the three waves. The mean of the absolute difference of the expectation from the first session to the second session is about 377 Euros. The corresponding value from the first to the third session is 401 Euros. Remember that expectations refer to monthly earnings and the average is about 1600 Euros per month. The results indicate that many students do not have clearly defined expectations and a single question on earnings expectations contains important measurement error. In the next section the measure of reliability is critically assess, to assure that this conclusion is correct.

5.1.1 Are the repeated questions parallel measures?

Table 3 includes *p*-values on the hypothesis that the mean earnings expectation is equal and also that the variance is equal for the different occasions. The table includes results both for a restricted sample, where the students are only included if they were present in all waves, and an unrestricted sample.

[Table 3]

The hypothesis of equal mean is never rejected on conventional significance levels. The hypothesis of equal variance cannot be rejected on the 5% significance level in none of the cases. If the significance level of 10% is used the hypothesis is rejected for wave 1 and 3. The tendency is lower standard deviation in the first occasion compared to the third session.

Doing pair-wise tests of the hypothesis that the correlation is equal for different combination of waves in table 1 the obtained *p*-values are between 0.528 and 0.839 for the lower left corner and between 0.745 and 0.998 for the upper right corner. Hence, the hypothesis that the correlation is equal cannot be rejected on any conventional significance level.

Table 4 includes correlation with each of the different earnings expectation and other variables. This table refers to the restricted sample, but some of the pair-wise correlation coefficient is calculated for even fewer observations because of non-response on some of the variables. The samples sizes are between 261 and 281 observations.

[Table 4]

The correlation of log earnings expectations, measured in any of the waves, and different self-evaluated skills and willingness to assume risk is found to be very weak. Since log earnings expectations are measured with random error the correlation coefficients in the table are underestimations of the correlation of the true variables. The hypothesis that the correlation is equal in two sessions is never rejected.

[Table 5]

Log earnings perceptions of other scenarios or earnings for individuals unrelated to the student are other kinds of variables. The correlation coefficients for these variables are included in Table 5. The point estimate of the correlation of log earnings expectation and the perception of log earnings for someone at age 45 year that graduated in the same studies as the student is higher for the first session compared to the second session, but the p-value (0.271) is not small enough to reject the hypothesis that they are equal. The same occurs compared to the third session where the p-value is found to be 0.132. The correlation of own earnings expectation and earnings perception of a student is, however, significantly different for the different sessions (p-value 0.000). The correlation is higher when own earnings expectation and earnings perceptions for students are asked on the same occasion. The same occurs for the randomly assigned study when ask on the second wave, but not for the first wave. The anchoring of expectations on perception on earnings in the market seems to be very strong when the questions are asked on the same occasion, and in particularly, when the perception refers to a group to which the student himself/herself belongs. Using the data collected in 2015 it is not possible to evaluate to what extent log earnings perception on the market is affected by measurement error. This analysis is done in section 5.4 with

data collected in 2016. The particularly strong correlation in the same wave gives the impression that the random error of earnings expectation and earnings perceptions may be correlated. This would increase the correlation beyond correlation of the true expectations and true market perceptions.

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

One reason that the correlation is fairly low between the first wave and the second wave could be that asking the question could make students discuss it, and the re-test that is made about two weeks later could include a different answer due to an adjustment to this new information. A similar argument is that the first session was done in the computer lab while the second wave was done outside class hour. Students were not allowed to discuss their questions and their answers when answering in the computer-lab. It is not possible to control if this was done outside class hour. Neither is it possible to check if students searched for additional information, but these options seem unlikely. If the correlation is calculated for the subgroup (n=279) that answered that they did not discuss it with peers, family nor *received other information,* etc. it is 0.5840, i.e. very similar to what is found for the complete sample. This suggests that the reactivity problem is not causing an underestimation of the reliability in this case.

It is interesting to find that the correlation is very similar despite the different time periods, i.e. about 14 days and about 74 days. If students would remember their first answer this could *over*estimate the reliability from the first session to the second session, while if students would receive new information this would *under*estimate the reliability from the first to the third session. These biases work in a way to make the measure on reliability different, and still the point estimates are found to be almost

identical. Accordingly, concerns about biases due to memory or new information can be discarded. It is important to keep in mind that many different scenarios on perception were included in the first survey and this makes it more difficult to remember the answers. If a small survey, with few questions, is used, the risk that students remember their answer would, of course, be higher, and separating the sessions with only two weeks could be too little.

5.1.3 Is the sample particularly homogenous?

It is possible that the sample is very homogenous, i.e. the current study includes students in the course Analysis of Economic Data, which include students in Economics, Business and Administration and also a small group that studies for a double degree (Business and Administration and Law). The coefficient of variation for earnings expectations (i.e. before calculating the logarithm) collected in the first wave is about 0.48 and the Gini index is about 0.19. The standard deviation of log earnings expectation is about 0.38. The sample is more homogenous compared to a population of high school graduates or university students in general. The coefficient of variation reported here is, however, fairly large, compared to other studies. Hartog and Diaz-Serrano (2013) review the literature and the coefficient of variation ranges from 0.2 to 0.4 for the different studies. Brunello *et al*. (2004) analyse earnings expectations from 10 European countries and report a standard deviation of log expected earnings after college of 0.56. It is, however, expected that the standard deviation is larger due to that they pool expectations from different countries. Webbing & Hartog (2004) report a coefficient of variation of 0.37 and a standard deviation of log earnings expectation of 0.29. While we consider our sample to be more homogenous compared to a population of high school graduates or university students in general, it does not stand out in the

literature to be a case of particularly low variation in log earnings expectations. Hence, homogenous samples, or samples with fairly low variation in log earnings expectations, seem to be a common problem and accordingly something that is expected to make the problems of measurement error more severe.

Earnings perceptions from randomly assigned studies were also asked in the first and second waves in this study. Calculated for the first wave, the coefficient of variation is about 0.47 and the Gini index is about 0.20. The correlation from the test re-test is only 0.5420 (n=375) for the randomly assigned studies. (The correlation refers to after performing the logarithmic transformation). The randomly assigned studies includes more variation in the kind of studies, despite this, the coefficient of variation and the Gini index is similar to the expectations on earnings after the current studies. The test re-test reliability may be lower because students have less information on these studies. Including only the students that did not discuss the earnings expectations makes the correlation even lower; 0.4914 (n=260).

5.2 Applying the measure of reliability on other studies.

The overall impression from the previous section is that asking students on their earnings expectations about 14 days, or 74 days later provides parallel measures suitable to quantify the reliability. The only concern is the particular strong correlation with market perceptions that is found when both questions are asked on the same occasion. Measuring anchoring with the correlation coefficient could be underestimated due to random measurement error in both variables, but a possible positive correlation of the measurement errors mean that the anchoring could be overestimated. The case of non-classical measurement error is not analyzed further in this study. This section show how to use the quantified reliability in another study, and in particularly we show how to

adjust the measure of reliability to become more suitable for studies where the homogeneity may differ.

From the current study we have obtained the variance of the true log expectation, $s^2(E^*) = s^2(E)r_{EE'} = 0.1471 \times 0.5893 \approx 0.0867$, and accordingly, the variance of measurement error is $s^2(\upsilon) = s^2(E) - s^2(E^*) = 0.1446 - 0.0867 \approx 0.0594$. Table 6 provides the information to use.

[Table 6]

The information in Table 6 can be used to adapt the reliability to a different degree of homogeneity. This is done to data on earnings expectations collected in 2014 and 2015 for students at Universitat Rovira i Virgili, URV. For the complete sample the average (standard deviation) were 7.16 (0.50) in 2014 and 7.07 (0.46) for log earnings expectation in their current studies. The sample size was 464 in 2014 and 162 in 2015. A subsample of 82 students had specified an identity number in both years which enabled matching answers from different years to a particular student. The correlation coefficient for this group is $\hat{\rho}_B = 0.3849$. Descriptive statistics for this subsample is, $\overline{E}_{-1,B} = 7.2119$, $\overline{E}_B = 7.0522$, $s_B^2(E_{-1}) = 0.3122^2$ and $s_B^2(E) = 0.3092^2$. Interestingly, the subsample is much more homogenous compared to the complete sample and the standard deviation in log earnings expectations is substantially lower in both years. Notice that the standard deviation is higher for the complete sample compared to what is found for the data from UIB, but since the correlation is calculated for a subsample, the relevant standard deviation is the one found in that sample. Since the standard deviation is smaller in URV the reliability of about 0.59 is too high. Below we calculate an alternative that takes into account the difference in homogeneity. From Table 6

above we find, $s^2_{A'}(\upsilon) = 0.0594$ and accordingly, $s^2_B(E^*_{-1}) = 0.3122^2 - 0.0594 \approx 0.0381$ and $s^2_B(E^*) = 0.3092^2 - 0.0594 \approx 0.0362$ which gives,

$$\hat{r} = r\left[\frac{0.0381}{(0.0381+0.0594)} \times \frac{0.0362}{(0.0362+0.0594)}\right]^{0.5} \approx r0.3847$$

and, accordingly; $r_B = 0.3849 / 0.3847 \approx 1.00$.

If we use the calculation on reliability directly the correlation of 0.3849 is adjusted to about 0.65, but since the sample is more homogenous compared to what is found in the data that provided the measure of reliability, it is more appropriate to use 0.3847. Then it turns out that the low correlation completely is due to measurement error in combination with a homogenous sample. The persistence in true earning expectation is very strong.

This procedure adjusts the reliability due different degree of homogeneity, while the variance of the measurement error is assumed to be the same as what is found in the data that is used to perform the test re-test analysis. This is of course an important assumption. Most studies work with incomplete scenarios and it is natural that both the heterogeneity and the measurement error can be affected by the details in these scenarios. It is, however, difficult to speculate if adding more details will reduce the measurement error more or less than the heterogeneity. For example, in our case, the question does no mention anything about working full time or not. This implies more heterogeneity if some students answer the question considering a full time job, while others keep in mind, for example, their own decisions on the labour supply. If students are consistent in their way to answer, measurement error is not increased, but if they

consider different undeclared features in different moments in time, this increases random measurement error and, accordingly, decreases the measure of reliability. Hence, adding more details will likely reduce heterogeneity but possibly also the measurement error.

It is important to keep in mind that that once the test re-test information is used for a different study it is possible that the variance of the measurement error actually is different. The calculations are done including high precision (in terms of used decimals) but the interpretations should be done in a much rougher way. If the degree of measurement error actually is higher in the current study, the used reliability may be too low and the adjusted correlation may even be above 1. Such result would of course indicate that the assumption of equal variance of measurement error is not fulfilled. Using reliability or the homogeneity adjusted reliability with the purpose to correct for attenuation bias in other studies, are rule of thumbs that are helpful to come closer to a "true" correlation.

5.3 Retest reliability on subjective earnings risk

While subjective earnings risk is not the main focus of this study, below we include measures of reliability. We measure subjective earnings risk as the sum of the stated probabilities to be <50% and >150% of the wage expectations. We expect that the reliability will be even lower because the subjective probabilities are stated once the expectation is settled. A measurement error in the wage expectation will add on to a measurement error in the subjective probability to be below 50% or above 150% of the stated expected earnings. Table 7 includes the measures on reliability.

[Table 7]

The sample size is smaller compared to when earnings expectations are analyzed. The reason is that only cases where the students' calculation of 0.5*wage expectation and 1.5*wage expectation were correct are included in the calculation. Mistakes in the calculation happened to a rate of 12%, 9% and 3% in the different waves. The measure on reliability of students' subjective earnings risk is about 0.43-0.49, and measurement error is accordingly even more severe compared to earnings expectations. Hence, the attenuation bias will be even worse when the variable is used in research.

5.4 Repeating and extending the analysis

The data from 2015 is inconvenient in some respect and these concerns were considered when new data was collected in 2016. First, changing the market perception on earnings from "a person at age 45" to "a student", as was done from the first wave to the second wave, made it difficult to study reliability in market perception. Accordingly, the questions considering earnings for "a person at age 45" were included in three waves in 2016. This addition also enables analyzing reliability in difference of log earnings expectation and log earnings perception from the market. The variables used in table 3 were also repeated in two waves to evaluate their reliability. The reason was simply that the correlations showed in the table could have been particular low due to measurement error both in log earnings expectations and the other variables, such as; self perceived skills or willingness to take risk. The setup for collecting the data was the same as in 2015. Two new double degrees were introduced at the university; Degree in Economics and Tourism and Degree in Administration and Tourism. The course Analysis of Economic Data is included in both these degrees and the composition of students in the data is, accordingly, extended to also include students from these degrees.

Table 8 includes measures of reliability and additional information for each of the variables. This information came from the first and the second wave. The average time between wave 1 and wave 2 was 15.7 days with a standard deviation of 5.4.

[Table 8]

The reliability of log earnings expectations is about 0.67. 72 out of the 386 students were enrolled in the new degrees mentioned above. If wave 1 and 3 are used, the reliability is about 0.59, but again, the sample size is smaller (n=301) for this case. If we use the reliability calculated in 2016 (from wave 1 and 2) to adjust the correlation in log earnings in the application in section 5.2, the correlation becomes about 0.57. If we, in addition, take into account the particular homogenous sample, the adjusted correlation is about 0.88. Again we find a very high persistence in log earnings expectations. It is clear that analyzing and interpreting the correlation of 0.38 would have provided a very misleading conclusion about inter temporal stability of log earnings expectations.

The reliability in log earnings perception for an individual "at age 45" is also about 0.67. In some cases the interest is on a difference of earnings expectation and earning perception on the market. Basically, if students think that the earnings are high (low) on the market, they also expect high (low) earnings. Cross sectional variation in earning expectations (for them self) could, accordingly, be difficult to explain because variation in earnings perception could be unrelated to observed information. A way to overcome this problem is to "anchor" the expectation to market perception by differencing. Since both variables are affected by measurement error and since the correlation is positive it is expected that the reliability will be even worse. The reliability of the difference of log earnings expectations and log earnings perception is

calculated to be about 0.49. Accordingly, if the idea is to use a linear regression model to explain the difference in log earnings expectation and log earnings perception, even using the same variable from two weeks earlier would only explain about 24% of the variation. It is natural that other variables will have difficulties to explain this difference. On the other hand, differencing does leave *some* signal and it is not only noise that is left.

Table 8 also includes reliability for self perceived skills and willingness to take risks. These calculations are only included to provide perspectives on the relatively low correlations found in table 3. It is clear that it is not only earnings expectation that is affected by random measurement error. For example, the reliability for willingness to take risks is about 0.64. Accordingly, the concern of using earnings expectations in empirical analysis without taking into account measurement error is equally relevant for these variables.

5.5 Discussion

The previous section shows how correcting for measurement error can provide different conclusions when transitory log earnings expectations are analyzed. In almost any use of students' earnings expectations the measurement error will affect the results. If earnings expectation is used as the dependent variable measurement error will imply a low coefficient of determination. Majeske et al. (2010) explain how a measure of reliability can be used to correct the adjusted coefficient of determination. Several studies show very low coefficient of determination and the exceptions are examples when scenarios are pooled and dummy variables are added to capture different averages among the scenarios. (See for example, Nicholson & Souleles, 2001, and Schweri et al. 2011). A similar situation is found in Brunello *et al*. (2004) where earnings expectations

from 10 European countries are pooled. Both pooling scenarios and obtaining more heterogeneity by using cross country variation in expectation imply larger variation and a higher coefficient of determination (due to dummy variables), but the explained variation is somewhat trivial. In addition, pooling implies adding restrictions that not necessarily are fulfilled. For example, the effect measured by a coefficient may be different in different countries, or for different scenarios, but the pooled model captures the effect with one single coefficient. Measurement error on the dependent variable also makes the precision lower, and the standard errors are too high. In particular for small samples, this may lead to not rejecting hypothesis that actually should be rejected. If log earnings expectations are used as an explanatory variable the coefficients of the model will be estimated inconsistently and the conclusion from the model will not be correct.

In some applications an interest is on a difference of an expectation and a perception on another scenario. For example, Brunello et al (2004) calculates an expected wage premium as the percentage difference between expected college and high school wages. The coefficient of determinations drops substantially compared to when earnings expectation is the dependent variable. Schweri et al (2011) calculates the difference between expectations and perceived actual earnings in the market. The purpose is to study if private information is related to this difference. The coefficients of determination for the different models are very low. In general, calculating a difference of two variables will reduce the reliability even further if the variables are positively correlated (Revelle, 2015). Table 5 shows a particularly strong anchoring when the own expectations and market perceptions are included in the same survey. In the previous section we find the reliability for a difference of own log earnings expectations and perceived log earnings for someone at age 45 graduated in the same studies. The reliability is only about 0.49. The measure on reliability of students' subjective earnings

risk is similar in magnitude. Accordingly, low reliability is found for all of the analyzed variables obtained from the probabilistic questions related to students' earnings expectations. This is a severe situation, but it also provides an important perspective on previous results in the literature.

6. Conclusion

The interest to elicit expectations has increased in the literature. Apart from asking direct questions on expectations it is common to add probabilistic questions to evaluate the uncertainty that the respondents attached to their own answers. The literature has put an important effort to evaluate the validity of the answers, and to create an environment to reduce logical inconsistencies in the answers. Very little attention is on the reliability of these measures. This study focus on students' log earnings expectations and a test-retest evaluation is performed. It turns out that students can provide a quiet different answer on their earnings expectations only about 14 days later. Using data from 2015 reliability for log earnings expectations was found to be about 0.59, and the corresponding result in 2016 is 0.67. The results indicate that students, in general, do not have well formulated earnings expectations. The measures of reliability are fairly similar to what Krueger & Schkade (2008) found for subjective well-being. They suggest that a reason for the fairly low reliability is that "answering life satisfaction questions explicitly invokes a non-systematic review of one's life, which leaves such measures vulnerable to transient influences that draw attention to arbitrary or incomplete information [...]. It seems as earnings expectations also are equally vulnerable, and this problem of measurement error should be taken into account in any empirical analysis using these variables.

We strongly recommend each study to perform its own evaluation of the reliability. In cases when this is practically not possible, we suggesting using a measure of reliability from a study where this problem can be supposed to similar. In this paper we also show how measurement error is related to working with a homogeneous sample. We suggest a way to adjust the reliability to make it more suitable to the homogeneity found in the study. In an application we study temporal stability in log earning expectations. The observed correlation of log earnings in 2014 and 2015 was about 0.38 at URV. Using the reliability (collected in 2015 and 2016) in another Spanish University (UIB) as a rule of thumb-adjustment means a correlation of about 0.57-0.65. If we, in addition, consider that the sample in URV is more homogeneous, the adjusted correlation would be between 0.88 and 1.00! Once the measurement error, in combination with a homogenous sample, is taken into account, the persistence in log earnings expectation is very, very strong. This conclusion is far from the originally observed correlation coefficient, and it shows how measurement error can provide an important distortion of the analysis.

The reliability is found to be fairly low and the problem of measurement error is quite severe. While the sample is considered homogenous due to being collected for a particular group of students, it does not stand out to be particularly homogeneous compared to other studies in the area. This, together with the already low reliability gives a discouraging impression on using students' log earnings expectation in empirical analysis. It is, however, important to consider what is the population of interest? If it actually is students in secondary schooling this would probably introduce more heterogeneity in log earnings expectations for the sample. Of course, the assumption of equal variance of the measurement error could be unfulfilled. A recommendation for future research is to expand outside the comfort zone of using

"own" students and actually collect a random sample from a relevant population. In this respect, we are as guilty as others in the area. Hopefully measures of reliability will be collected in different situations which will provide a more complete picture of measurement error in the variable of students' log earnings expectations.

The analysis is particularly focused on students' earnings expectations and the reliability of questions on other expectations and probabilistic reasoning are still to be evaluated. While the magnitude of the problem can vary substantially depending on the area, assessing the reliability should a natural first step in the analysis. If this step never is taken, any conclusion could be severely misleading.

References

Brunello G, Lucifora C, Winter-Ebmer R. 2004. The wage expectations of European business and economics students. *The Journal of Human Resources* **39**: 1116-1142. DOI: 10.3368/jhr.XXXIX.4.1116

Carmines EG, Zeller RA. 1979. *Reliability and Validity Assessment*. Sage University Papers series on Quantitative Applications in the Social Sciences, 07-017. Beverly Hills and London: SAGE Publications.

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? *Journal of Applied Econometrics* **26**: 479-497. DOI: 10.1002/jae.1233

Dominitz J. 1998. Earnings expectations, revisions, and realizations. *The Review of Economics and Statistics* **80**: 374-388. DOI: 10.1162/003465398557618

Dominitz J, Manski CF. 1996. Eliciting student expectations of the returns to schooling. *The Journal of Human Resources* **31**: 1-26. DOI: 10.2307/146041

Gottschalk P, Huynh M. 2010. Are earnings inequality and mobility overstated? The impact of nonclassical measurement error. *Review of Economics and Statistics* **92**: 302-315. DOI: 10.1162/rest.2010.11232

Gouret F, Hollard G. 2011. When Kahneman meets Manski: Using dual systems of reasoning to interpret subjective expectations of equity returns. *Journal of Applied Econometrics* **26**: 371-392. DOI: 10.1002/jae.1224

Hartog J, Diaz-Serrano L. 2013. Schooling as a risky investment: A survey of theory and evidence. *Foundation and Trends in Microeconomics* **9:3-4**: 159-331. DOI: 10.1561/0700000053

Krueger AB, Schkade DA. 2008. The reliability of subjective well-being measures. *Journal of Public Economics* **92**: 1833-1845. DOI: 10.1016/j.pubeco.2007.12.015

Majeske KD, Lynch-Caris T, Berlin-Fornari J. 2010. Quantifying R^2 bias in the presence of measurement error. *Journal of Applied Statistics* **37**: 667-677. DOI: 10.1018/02664760902814542

Manski CF. 2004. Measuring expectations. *Econometrica* **72**: 1329-1376. DOI: 10.1111/j.1468-0262.2004.00537.x

Manski CF. Molinari F. 2010. Rounding probabilistic expectations in surveys. *Journal of Business & Economic Statistics* **28**: 219-231. DOI: 10.1198/jbes.2009.08098

Nicholson S, & Souleles NS. 2001. Physician income expectations and specialty choice. NBER Working Paper 8536.

Revelle W. 2015. An introduction to psychometric theory with applications in R, Chapter 7, E-book.

Schweri J, Hartog J, Wolter SC. 2011. Do students expect compensation for wage risk? *Economics of Education Review* **30**: 215-227. DOI: 10.1016/j.econedurev.2010.12.001

Solon G. 1989. Biases in the estimation of intergenerational earnings correlations. *The Review of Economics and Statistics* **71**: 172-174. DOI: 10.2307/1928066

Wiswall M, Zafar B. 2015. How do college students respond to public information about earnings? *Journal of Human Capital* **9**: 117-169. DOI: 10.1086/681542

Van Santen P, Alessie R, Kalwij A. 2012. Probabilistic survey questions and incorrect answers: Retirement income replacement rates. *Journal of Economic Behavior & Organization* **82**: 267-280. DOI: 10.1016/j.ebo.2012.02.007

Zafar B. 2011. Can subjective expectation data be used in choice models? Evidence on cognitive biases. *Journal of Applied Econometrics* **26**: 520-544. DOI: 10.1002/jae.1236

Appendix

Clarifications on data and variables

The randomly assigned studies included the following degrees; Medicine, Biology, Law, Psychology, Sociology, History, Mathematics, Philology (Spanish or Catalan) and Art History.

Each skill (mathematical, verbal, social and commercial) were defined for the students, and an introduction clarified the scale of the ordering; from 1 (lowest) to 10 (highest). The introduction and an example of a question are included below.

Now we ask you to classify yourself in relation with others regarding four skills:

Now we ask you to classify yourself in relation to other people your age who have graduated high school. Imagine all those people are classified by skill level. At the bottom of the line-up is the person with the lowest capacity, on top is the person with the highest capacity. Now let's cut this line in groups of equal size. In group 1 are the first 10% of people with the lowest skill levels. In group 2 are the following 10%, the next skill level. Then the group 3, with the next skill level, and so on, until the group of 10, consisting of 10% with the highest skill levels. What group do you consider yourself?

What is your position on Mathematical ability in relation to other people your age who have graduated high school?

Another question was on how they perceive themselves considering risk taking.

How did you see yourself? Are you a person who is usually fully prepared to take risks? Or do you try to avoid taking risks? Please select an option on the following scale where (1) not at all willing to take risks and (10) fully prepared to take risks.

Table A1. Correlation coefficients in Zafar (2011)

	Correlation	
	age 30	age 40
eng	0.6064 (51)	0.5866 (51)
nat	0.3273 (40)	0.5008 (40)
math	0.2661 (32)	0.4058 (32)
soc1	0.4416 (66)	0.6320 (66)
soc2	0.5649 (41)	0.5637 (41)
eth	0.7210 (23)	0.2123 (23)
area	0.0233 (43)	0.1606 (43)
lit	0.6180 (52)	0.6353 (52)

Note: sample size is included in parenthesis.

Tables

Table 1. Descriptive statistics for data collected in 2015.

	wave 1			wave 2			wave 3		
	mean	std.dev	n	mean	std.dev	n	mean	std.dev	n
Earnings expectations, yourself, current	1625.70	776.20	467	1641.12	735.24	421	1621.74	624.93	312
ln(earnings expectations), yourself, current	7.31	0.41	467	7.33	0.38	421	7.32	0.39	312
ln(earnings perception), 45 years, current	7.58	0.40	461						
ln(earnings perception), "a student", current				7.30	0.61	420			
ln(earnings perception), yourself, random	7.15	0.49	467	7.18	0.44	420			
ln(earnings perception), 45 years, random	7.41	0.48	468						
Mathematical skill	6.62	1.80	470						
Verbal skill	6.56	1.63	448						
Social skill	6.79	1.56	432						
Commercial skill	6.45	1.62	441						
Willingness to take risks	6.59	1.75	471						

Notes: Earnings expectations are measured in Euros per month. n refers to the number of valid answers.

Table 2. Correlation matrix on log earnings expectation from different surveys.

	Wave 1	Wave 2	Wave 3
Wave 1		0.5893 (401)	0.5892 (303)
Wave 2	0.5998 (281)		0.6056 (284)
Wave 3	0.6330 (281)	0.6107 (281)	

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 found in the two waves analyzed, but not necessarily in the third. The sample size is included in parenthesis.

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		0.3415	0.9259		0.7225	0.7629
		(401)	(303)		(401)	(303)
Wave 2	0.6919		0.7092	0.1018		0.9560
	(281)		(284)	(281)		(284)
Wave 3	1.0000	0.7006		0.0961	0.9776	
	(281)	(281)		(281)	(281)	

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 analyzed, but not necessarily in the third. The same idea concerns the variance. The sample size is included in parenthesis.

Table 4. Correlation on log earnings expectation and other variables

Log earnings expectations	Mathematical skill	Verbal skill	Social skill	Commercial skill	Willingness to take risks
Wave 1	0.0423	0.1047	0.0915	0.1153	0.0974
Wave 2	0.1018	0.0854	0.1389	0.1781	0.1813
Wave 3	0.0810	0.0808	0.1411	0.1168	0.1349

Table 5. Correlation on log earnings expectations and log earnings perceptions

	Log earnings perceptions for "…" on "… study", asked on "wave …".			
Log earnings expectations	"a person 45-years old" "current study" "wave 1"	"a student" "current study" "wave 2"	"yourself" "random study" "wave 1"	"yourself" "random study" "wave 2"
Wave 1	0.5390	0.5274	0.4005	0.3900
Wave 2	0.4693	0.8227	0.3831	0.6647
Wave 3	0.4419	0.4711	0.3353	0.4287

Table 6. Reliability and descriptive statistics for log earnings expectation in 2015.

Description	Notation	Value in current study (using log earnings)
Reliability	$r_{EE'}$	0.5893
Variance of measurement error	$s_E^2(\upsilon)$	0.0594
Mean of log earnings expectation	$\overline{E}$	7.3195
Variance of log earnings expectation	$s^2(E)$	0.1446
Variance of true log earnings expectations	$s^2(E^*)$	0.0852

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.

Table 7. Pair wise correlation matrix on subjective earnings risk.

	Wave 1	Wave 2	Wave 3
Wave 1	1		

	0.4900	1	
Wave 2	(333)		
	0.4266	0.4326	1
Wave 3	(245)	(235)	

Note: Subjective earnings risk is measured as the sum of stated probability to end up either below 50% or above 150% of the wage expectation.

Table 8. Reliability and descriptive statistics for several variables in 2016

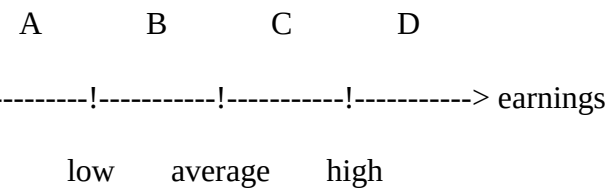
Description	Notation	log earnings expectations	log earnings perception	Difference	Maths skills	Verbal skills	Social skills	Commercial skills	Willingness to take risks
Reliability	$r_{XX'}$	0.6738	0.6714	0.4878	0.7778	0.6153	0.6975	0.5750	0.6406
Variance of measurement error	$s_X^2(\upsilon)$	0.0542	0.0611	0.1888	0.8677	0.9772	0.8445	1.1020	1.2585
Mean	$\overline{X}$	7.3652	7.5815	-0.2204	6.5850	6.6328	6.8151	6.5815	6.3977
Variance	$s^2(X)$	0.1663	0.1858	0.3686	3.9050	2.5401	2.7916	2.5930	3.5016
Variance of true X	$s^2(X^*)$	0.1121	0.1247	0.1798	3.0373	1.5629	1.9471	1.4910	2.2431
Sample size	n	386	387	383	394	369	357	368	391

Notes: The data collected in 2016 is used for the table. Log earnings expectations refers to the current studies for the student while log earnings perceptions refers to earnings for someone at age 45 years that had graduated in the studies that the individual is pursuing. "Difference" is the difference of between these variables. The definition of the variables on skills and risk can be found in Appendix. The average and variance refers to the joint average and the joint variance for the two waves (and the sample size is accordingly twice what is specified in the table).

Online Appendix

The students were asked to divide the amount specified as earnings perception by two and then to sum the first and second answers. Hence, three limits are identified, low (50% of earnings perception), average (perception on average earnings) and high, (150% of earnings perception). A scheme was included to clarify the information that would be used in the following questions. This scheme is included below.

Figure 1. Scheme to help students to understand questions.



The students were given the additional information that

"Of course, not all individuals of 45 years old will earn the average income. Wages may fall into four possible intervals: Below "Low", between "low" and "average" between "average and high" and "above high", as indicated by the figure above.

So think of 100 people aged 45 that graduated in the study that you currently are pursuing. How many do you think will have earnings in each of the intervals? Remember that the sum of your responses should equal 100."

Four questions were then included, where the first is included as an example below:

Of the 100 people aged 45 that graduated in the study that you currently are pursuing, how many do you think have their earnings below "low" (area A in the graph)?

Some clarifications are necessary. Based on previous surveys on students we prefer to ask for the average instead of the median, because using the median conditions that answers to areas A+B = 50 and C+D=50, which most students do not pay attention to. Hence, using the median invites to have logical inconsistencies. Notice that we ask for 50% and 150% of the mean instead of, for example 75% and 125% of the mean. Since the questionnaire is not interactive the students have to perform the calculations, and choosing 50% simplifies the calculation for the students. Another issue concerns using four areas, despite that the primary interest is on A and D. A common mistake when only asking for the tails of the distribution is that students think that they should sum to 100% which causes a very strange distribution with no probability mass in between “low” and “high”. Adding the questions on B and C corrects this mistake. Finally, notice that the formulation uses 100 people to be distributed in the four areas, hence we do not ask for a probability, but the students are introduced to the idea of a probability.

The same structure was used to ask students about their earnings expectations and how uncertain these expectations were considered. After providing their expectation students were asked to calculate half of the earnings expectation and the sum of calculated half and the stated expectations. The additional information was included to help the students to answer the next questions.

Of course, you do not know if your earnings will be equal to the expectation that you have answered above. Your salary can fall into four possible areas: “Below low”, “between low and average”, between "average and high” and “above high”, as indicated by the figure above.

In the following four questions, we want you to indicate the probability (in %) that your gross monthly earnings will be in each of the intervals. Remember that the sum of your responses should equal 100.

The first question is included below as an example,

What do you think is the probability (in %) that you will obtain earnings below "low"?: (area A in the graph)

Table A1. Descriptive statistics for subjective risk, collected in 2015.

		wave 1		
		mean	std.dev	n
Earnings expectations, yourself, current	Subjective risk	33.05	13.17	414
Earnings perception, 45 years, current	Subjective risk	34.97	12.07	415
Earnings perception, yourself, random	Subjective risk	34.58	15.47	410
Earnings perception, 45 years, random	Subjective risk	34.73	13.36	414

Notes: Subjective earnings risk is measured as the sum of stated probability to end up either below 50% or above 150% of the wage expectation. n refers to the number of valid answers. These samples are restricted to students with a correct calculation of 0.5*wage expectation and 1.5*wage expectation.