THE IMPACT OF VOCATIONAL TRAINING FOR THE UNEMPLOYED IN TURKEY

Pre-Analysis Plan

Rita Almeida, Sarojini Hirshleifer, David McKenzie, Cristobal Ridao-Cano,
Ahmed Levent Yener

February 6, 2012

1. Introduction

This plan outlines the hypotheses to be tested and specifications to be used in the analysis of the impact of the Turkish National Employment Agency's vocational training programs. Since the authors completed the plan before the follow-up data was collected and analyzed, the plan can provide a useful reference in evaluating the final results of the study.

The plan is outlined as follows: Section 2 reviews the motivation for the study, the sample selection and data sources; Section 3 enumerates the hypotheses to be tested as part of the study; Section 4 outlines the specifications to be used in analyzing the data. Appendices 1-4 provides additional details on the evaluation sample, the data sources used, the balance of the evaluation sample at baseline, and the demographic profile of the evaluation sample. Appendix 5 is a table that summarizes the hypotheses and indicators.

2. Overview of the Study

2.1 Motivation and Program Description

This study offers the first rigorous evaluation of a national vocational training program for the unemployed in a middle-income country. There are few existing experimental evaluations of active labor market programs in developing countries and they are typically focused on smaller pilot programs or training programs for youth. This evaluation will improve our understanding of whether job training programs are effective, and why, in a range of labor market contexts across a large middle-income country.

The subject of this evaluation is the primary vocational training program of the Turkish Employment Agency (ISKUR), which is currently the single most important active labor market program in Turkey.[1] In response to an employment reform in 2008 and the economic crisis in 2009, the number of people trained by ISKUR-supported programs increased rapidly from 25,000 in 2008 to over 100,000 in 2009, and 200,000 in 2010. In 2009, the Government of Turkey requested support from the World Bank in evaluating ISKUR-supported vocational training programs.

The World Bank worked closely with ISKUR to design and implement a large-scale randomized impact evaluation of its vocational training program. ISKUR provides training by contracting with public and

[1] Turkey has a population 75 million and GDP per capita of $8,200.

private providers to offer training courses in a range of occupations. Unemployed individuals can apply to any training courses for which they met the basic requirements. The courses are free and trainees receive a small daily stipend intended to cover transportation and lunch while attending a course. Individuals are only allowed to take one ISKUR-supported course in a five year period. Since ISKUR-supported training courses are frequently over-subscribed, the evaluation design relied on randomly selecting some applicants to receive training and comparing their employment outcomes with those who were not selected to receive training. The trainings that were included in the evaluation began between October and December 2010 and were completed by April 2011.

2.2 Sample Selection

The evaluation sample consists of 5,700 individuals, with approximately equal numbers selected into taking vocational training (treatment group) and selected out of training (control group). The 130 courses included in the evaluation took place in 23 provinces representing a range of labor market conditions from across Turkey.

2.2.1 Province Selection

The primary considerations in selecting provinces were: i) ensuring that the province would have oversubscribed courses, and ii) that the evaluation captured the effect of training under a range of labor market characteristics. The provinces in the evaluation sample were chosen from the 39 provinces which had at least two significantly oversubscribed training courses in 2009.[2] In 2010, the unemployment rate in Turkey was 11.9 percent, with significant regional variation. Approximately half of the provinces with oversubscribed courses had unemployment rates that were above 10 percent. Thus, provinces were first stratified by whether they had an unemployment rate above 10 percent; then 10 provinces were then randomly selected from each strata with probability proportionate to total size of ISKUR vocational training programs in that province. Three additional provinces (Antalya, Gaziantep, and Diyarbakir) were included in the sample at the request of ISKUR because of their importance in representing varying labor market conditions across Turkey.

2.2.2 Course Selection

The evaluation team worked closely with the regional ISKUR offices to determine the final set of courses to be included in the evaluation.[3] The vocations that were taught in the courses selected for the evaluation reflected a range of the most popular vocations taught in ISKUR-supported courses (See Appendix 1). The course capacities ranged from 12 to 100 trainees; however, most of the courses in our sample had between 20 and 25 trainees. Several criteria dictated the final selection of courses, the most important of which were:

[2] Oversubscription in the training courses was defined as having at least twice the number of applicants as available spaces. In a significant number of provinces, the demand for and provision of ISKUR vocational training is quite low, largely as the result of the local economic context.

[3] ISKUR provides a range of training programs, but the majority of its trainees are trained through the general vocational training program, thus the courses in that program were the focus of evaluation.

- **Oversubscription:** The evaluation only included training courses which were likely to be sufficiently oversubscribed in order to allow for the random selection into treatment and control among individuals in the evaluation sample.
- **Course providers:** The evaluation prioritizes diversity in the training providers for the same vocation. This will enable the analysis of the effectiveness of the trainings for providers of the same courses with different qualities. A combination of public and private service providers is also explicitly targeted.[4]
- **Timing of the courses:** Evaluation sample covers courses that started between October and December 2010 and finished by April 2011. The timing of the start of the evaluation was determined by the fact that it tends to be a time of year when people in Turkey are more likely to seek training through ISKUR.

2.2.3 Assignment of Treatment

Potential trainees applied to courses that they were interested in and were interviewed by training providers according to standard procedures. At no time were trainees or prospective trainees informed of the evaluation. For the purposes of the impact evaluation, training providers selected twice as many applicants for training than usual and submitted the list to ISKUR's Management Information System (MIS). The MIS then randomly selected participants into treatment (after stratifying on age and gender) using code written by the evaluation team and returned the list of trainees to course providers. There was little difficulty in convincing training providers to participate in the evaluation, since most courses in the evaluation sample were heavily oversubscribed, and the interview process for selecting potential trainees was as a rule not especially intensive.[5]

2.3 Key data sources

The primary sources of data are a baseline survey conducted immediately before trainings began and a follow-up survey that will be collected approximately nine months after the end of trainings. Additional sources of data (including: a brief survey of course providers, ISKUR's Management Information System (MIS), and the Turkish Labor Force Survey) are outlined in Appendix 4.

2.3.1 Baseline survey

There is typically a short period of time (around five days on average) between the time a training providers interviews prospective trainees and the time that a given course begins. In order to allow time to conduct the baseline survey before trainees were informed of whether they had been accepted into the program, training providers were asked extend this period to approximately 10 to 15 days.[6] Thus, the baseline data collection took place on a rolling basis (as courses began) between September and December 2010.

[4] Inevitably, the timing of the evaluation affected the types of the training providers that are part of our sample. For example, in some provinces, the courses provided by MoNE open only in summer.

[5] In Istanbul trainings were advertised somewhat more widely than usual in order to ensure over-subscription. A note on the baseline report describes in more detail the randomization procedure.

[6] Further extending this period had the potential to significantly lower take-up rates of the treatment group, in part because they may have simultaneously applied to multiple trainings.

Of the 5,700 individuals in our evaluation sample, 5,318 individuals responded to our baseline survey for a 6.7% refusal rate. The evidence from the baseline data indicates that there are few statistically significant differences across treatment and control on a range of co-variates (See Annex 3).

2.3.2 Follow-up survey

The follow-up survey will be conducted in January-March 2012.

3. Hypotheses

We have collected a relatively rich dataset, which will be used to test a number of hypotheses with regards to the impact of training and the determinants of whether training has an impact. We also assess evidence of market failures that may prohibit the take-up of training in the absence of publicly funding. The hypotheses are presented in groups, which are summarized in Table 1.

Table 1: Hypothesis Groups

A.	**Impact on Outcomes:** ISKUR training may have positive average impacts on employment and non-employment outcomes for trainees and their households. We group these outcomes into four domains: employment, individual well-being, household well-being, and empowerment and attitudes.
B.	**Causal Chain of Process and Mechanisms:** The training will have more effect if individuals attend and complete it, if the training is longer and higher quality, if it teaches them useful skills and job seeking tips, and if labor demand is strong.
C.	**Heterogeneity of Impacts:** The individual characteristics of trainees will determine their labor market outcomes and may determine whether they differentially benefit from training; training impact may also differ with type of course and type of provider.
D.	**Market Failures:** In the absence of receiving free vocational training from ISKUR, market failures prevent individuals from paying for training—even if training has high returns.

Hypothesis Group A: **IMPACT ON OUTCOMES** *ISKUR training may have positive average impacts on employment and non-employment outcomes for trainees and their households. The likely transmission mechanism for non-employment outcomes is increased employment.*

We group our outcome measures into four domains:

Hypothesis A.1: ISKUR vocational training is likely to have a positive average impact on **employment outcomes** – both in terms of the likelihood of work, and also in terms of the formality and quality of the work obtained and expectations about future likelihood of work.

The following individual indicators will form the family of outcomes in this domain:

Outcomes obtained from follow-up survey for the individual who applied for training:

- Individual has worked at all for pay in the last 4 weeks (follow-up D.21, baseline C62)
- Individual currently works for 20 hours a week or more (follow-up D.23, baseline C66)
- Hours worked per week in last month employed (follow-up D32, baseline C77)
    - This will be coded as zero for individuals who are not currently working
    - This will be top-coded at 100 hours per week (99^{th} percentile of baseline response) to reduce influence of outliers
- Total monthly income from work in the last month (follow-up D.31, baseline C78)

    - This will be coded as zero for individuals who are not currently working
    - This will be top-coded at the 99th percentile of the control group earnings distribution conditional on working to reduce influence of outliers
- The inverse hyperbolic sine transformation of total monthly income from work in the last month $\log(y+(y^2+1)^{1/2})$ – which is similar to the log transformation, but can deal with zero income. (transform of follow-up D.31, transform of baseline C78)
    - This will be coded as zero for individuals who are not currently working
- Individual is employed in a job covered by social security (follow-up D28, baseline C75).
- Occupational status of workers. This will be coded based on the occupation in D22 and the international measures of occupational status of Ganzeboom and Treiman (1996)[7]. This is a continuous measure ranging from 16 (e.g. domestic helpers) to 90 (e.g. judges).
    - Given the small number of individuals working at baseline, we will not control for baseline occupational status when looking at this outcome.
    - This will be coded as zero for individuals who are not currently working
- Expected chance of having a job in two years time (follow-up E11, baseline E12)
    - Responses that lie outside 0 to 100 range will be coded as missing
    - This question will be dropped if more than 95% of the control group answer 100%.

Outcomes obtained from Social security record data:

- Individual currently is employed in a job which is registered in the social security system
- Income the individual earns from formal work

A standardized employment outcome impact will be obtained by aggregating these different effects as described below in our methods section. In addition, we will also explore whether income, formality of job status, and occupational quality are higher conditional on working. Since this conditions on an endogeneous outcome, these conditional outcomes will not be part of the formal analysis, but will be used for exploratory purposes only.

Hypothesis A.2: Training may have positive impacts on **individual well-being** for the trainee.

The following indicators are part of this domain:

- Mental health, as measured by the MHI-5 index Veit and Ware (1983). This is a five item scale with a maximum score of 25 and minimum score of 5. Higher scores are desirable in that they indicate the experience of psychological well-being and the absence of psychological distress during the past 4 weeks. This is the sum of responses to G3.1-G3.5 after reverse-coding questions G3.1, G3.2 and G3.4.
    - This was not asked at baseline.
- Subjective well-being today (ladder question J.0 on follow-up)
    - This was not asked at baseline.
- Expected subjective well-being in 5 years (ladder question J.10 on follow-up)
    - This was not asked at baseline

[7] Ganzeboom, H., and Treiman, D. 1996. Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations. *Social Science Research* 25(1): 201-239.

Hypothesis A.3: Receiving training may have positive impacts for the trainee's **household well-being**.

Conditional on remaining in the same that they were living in at baseline, improved employment outcomes that are the result of training may lead to small positive impacts for the trainee's household.

Indicators:

- Total household income from all sources in last 12 months (sum of follow-ups J.1-J.12, baseline C177-188)
    - Coded as zero if household has no income
    - Top-coded at 99th percentile of the control group distribution.
- The inverse hyperbolic sine transformation of total monthly household income from all sources in last 12 months.
    - Coded as zero if household has no income
- Household asset ownership (regardless of whether individual is in same household). This will be formed as an index which is the first principal component of the following indicators, using the principal component weights derived from the baseline data. Assets which had more than 95% ownership or 5% or lower ownership at baseline are not used in this index. (follow-up I1, I2, baseline I1, I2)
    - Gas or electric oven; microwave oven; dishwasher; DVD/VCD player; Camera; Digiturk/Satelite; Air Conditioner; CD player/ipod; Telephone; Computer; Internet connection; Private car; Taxi/Minibus/Commercial vehicle; Bicycle; 4 or more rooms in their house.

Hypothesis A.4: Receiving training may have impacts on **social outcomes** such as whether individuals head their own household, their decision-making power, and their gender attitudes.

Indicators in this domain

- Individual is the head of the household or spouse of the head (follow-up C7)
    - Baseline doesn't clearly measure whether they are the spouse of the head, only whether they are the head. We therefore will not include baseline value here.
- Index of number of decisions that individual is the most important decision-maker for (they answer respondent is most important) among following decisions (from module H1 in both baseline and follow-up.).
    - Clothes for yourself
    - Whether you work outside the home
    - How money earned by you is spent
    - Time you spend socializing outside the house
    - What education/training pursuits you follow
    - Selection of a spouse for you
    - Which profession you will pick for yourself
- Index of pro-gender-equality beliefs (available for baseline and follow-up)
    - Strongly agrees both husband and wife should contribute to household income (H2.1=4, vs other options)
    - Strongly disagrees that a university education is more important for a boy than for a girl (H2.2=1, vs other options)

- Strongly disagrees or disagrees that a married women should not work outside the home unless forced to do so by circumstance (H2.3=1 or H2.3=2 vs other options)
- Strongly disagrees that it is demeaning to a man for his wife to work (H2.4=1, vs other options)
- Strongly disagrees that women could express their opinions in the family but never in public (H2.5=1, vs other options)
- Strongly disagrees or disagrees that a wife must always obey her husband (H2.6=1 or H.2.6=2, vs other options).

Unpacking the causal chain and understanding for whom the training is most effective

We then want to understand why the training did or not improve employment outcomes (**process and mechanisms)**, including whether it worked better for certain types of people or courses (**heterogeneity)**, and why, if it worked, did people not undertake it without ISKUR's intervention (**market failures**). Since employment is the main goal of this program, we will focus this analysis on two outcome measures only: overall standardized employment domain outcome, and to the outcome of currently being employed for 20 or more hours per week.

Hypothesis set B: Process and Mechanisms

We envision a 5-step causal chain through which selection into a course can lead to employment outcomes:

Step 1: Individuals show up and attend courses and complete them

Step 2: Courses are taught to a reasonable standard and quality

Step 3: Training changes skills, adds credentials, and potentially job-seeking behavior of trainees

Step 4: There is sufficient labor demand that trained individuals can find jobs to apply for

Step 5: These changes lead to changes in employment outcomes (this is what the outcome measures above look at).

The training may fail to have an impact on employment outcomes if any of these steps does not occur. We investigate these steps as follows.

Hypothesis B1: (Step 1) The employment impacts will be less for courses with low attendance or completion rates

We will use the MIS tracking data to estimate the heterogeneity of treatment effects with respect to the following course-level indicators:

- Percentage of individuals assigned to treatment who attended this course
- Percentage of individuals assigned to treatment who completed this course

Since variation is at the course-level, standard errors will be clustered at the course level for this analysis.

Hypothesis B2: (Step 2) The employment impacts will be greater for courses which are longer and of higher quality.

To test this we will examine the heterogeneity of treatment effects with respect to the following course-level indicators:

- Number of hours of the course above median (MIS data and provider survey)
- Whether the course has two or more competitors, which we take as a proxy of higher quality (provider C41)
- Whether the average experience of teachers in the course is greater than 12 months (the median) (provider C61).
- Percentage of teachers for the course which are university graduates (provider C57/(provider C55+C56)). Cap this at 100%.

We will also take the first principal component of these different indicators as a summary proxy of course quality, and examine heterogeneity with respect to this measure.

Hypothesis B3: (Step 3) The training has increased skills, added credentials, and potentially changed job-seeking behavior.

Improving the skills of trainees is likely to be the primary means through which training can improve labor market outcomes—particularly increased wages and duration of employment. Unfortunately skills are difficult to measure directly, since the Turkish Government is still developing standardized vocational certifications. Others means by which training can affect labor market outcomes include: i) signaling previously acquired knowledge or innate characteristics, ii) increased knowledge of job opportunities (including through peer networks), iii) increased optimism leading to more intense job search, iv) increased knowledge about how to present oneself for a job, and other soft skills; and v) potential changes in reservation wages. Most of these mechanisms we will only have self-reported assessments for.

Self-assessed view of what course did: Follow-up B5 and E1 descriptive statistics

- Taught new technical skills
- Certified skills they already had
- Taught new strategies for finding a job
- Helped put in contact with employers
- Gave them confidence to apply for new jobs
- Made them more aware of job opportunities
- Think employers value ISKUR certification
- Think training is useful to learn a job
- Think training is useful because it improves knowledge

We will examine heterogeneity of treatment effects with the percentage of respondents in a course who think:

- Taught new technical skills (new skills mechanism)
- Certified skills they already had (signaling mechanism)
- Taught new strategies for finding a job (job matching improvement)

- Made them more aware of job opportunities (job matching improvement)

Changes in job-seeking behavior will be difficult to look at experimentally if the treatment affects the likelihood of being unemployed in the first place. With this caveat in mind, we can look, conditional on being unemployed at treatment impacts on:

- Whether they have actively sought a job in the last 4 weeks (follow-up D7)
- Number of hours spent looking for a job in the past 7 days (follow-up D10)
- Number of methods used to look for a job (follow-up D13)
- Number of jobs (truncated at 3) they have applied for in past 4 weeks (follow-up D14)
- Willingness to take a job at 660 TL per month (follow-up D18).

Hypothesis B4: The training will be more effective in labor markets with strong labor demand.

It is not clear whether additional skills should be more or less valued in a tight or competitive labor market, but on balance we hypothesize that a lack of labor demand will make it more difficult for the training to have an effect. We therefore will look at heterogeneity of treatment effects by the following labor market indicators:

- Whether provincial unemployment rate is above or below 10% (Turkey LFS)
    - This regression will have to have standard errors clustered at the province level.

Hypothesis Group C – Heterogeneity of Impacts: *The individual characteristics of trainees are likely to determine their labor market outcomes, and may determine whether they differentially benefit from training; the characteristics of the provider and course may also influence the treatment impact.*

We will examine treatment heterogeneity according to the following dimensions:

- Expected benefit: do individuals who expect larger impact of ISKUR training on likelihood of being employed have larger treatment impact. Construct as continuous variable from baseline C109-C108 (median difference is 30 percentage points)
- Gender (baseline question C1, 63% female): women are more likely to apply for the program, but have lower labor force participation rates in general. It is not clear ex ante which gender impacts will be greater for, but we hypothesize that women may be more likely to be doing training courses for reasons other than to find a full-time job, such as hobby interest or for part-time work, so impacts will be less for women.
- Education (whether or not individuals have post-high school education, based on baseline C8): direction unclear – less skilled individuals may have more trouble finding employment in general, but may also have a greater marginal impact from the training.

- Previous recent training course (baseline C12 – 26% have taken a course in last 5 years): if there are diminishing returns, the effect will be less for this group; if training requires building on past skills, the impact could be greater for this group.
- Child care responsibilities (whether or not the respondent has a child aged 6 or under in the household with them – from household roster): we expect the training impact to be lower for those with young children.
- Empowerment over working outside the home (H1.3 baseline, whether or not they say they are the main decision-maker in this): we expect the impact to be less for people who are not the main decision-maker over whether they work.
- Cognitive ability: Score on baseline Raven's test (Baseline L5) – we expect higher ability people to have a larger treatment effect
- Numeracy: whether or not an individual got all four questions right at baseline (L1-L4): we expect more numerate people to have a larger treatment effect.
- Work centrality (baseline C130, coded 1 through 5) of Mishra et al, 1990 – we expect people for whom work is more central to see larger impacts.
- Tenacity (Baum and Locke, 2004): this measures the extent to which individuals persist in difficult circumstances. It is constructed as the sum of C118 and C119 from the baseline: we expect people who are more tenacious to see larger impacts.
- Type of course provider: Private sector vs Other
- Labor market history (length of time unemployed above or below median).
- Type of course: are impacts different for either of the most common course types (computer courses, accounting courses) vs the rest?

Hypothesis Group D (Market Failures): *In the absence of receiving vocational training from ISKUR, market failures prevent individuals from paying for training—even if training has high returns.*

If we find the training has high returns, then the natural question is why individuals don't purchase it themselves even if ISKUR doesn't fund them. A variety of market failures might explain this. To examine which we market failures seem most important, we will examine heterogeneity in treatment response to particular indicators of access to different markets. If a market failure is operating, we should expect returns to training to be highest for people subject to this market failure. For example, if access to credit is a constraint for some people, then among those who are not credit constrained, we should see all those with high returns to training purchase it and those with low returns not purchase it; when offered for free it will only be some of these low returns people among those with access to credit who take the training. In contrast, the credit-constrained group will contain both high and lower return people, and should thus have a higher average return.

Question C106 in the baseline which asks directly why they might not invest in training has a most common response of cannot invest because of limited access to credit (52%).

We will look at heterogeneity of treatment effects with respect to the following:

- Missing credit market
    - Trainee thinks could obtain a loan for 1000 TL at baseline (C196) – 51.9% think they could.

    - Trainee thinks they could obtain a loan for 3500 TL at baseline (C198 and C196) – 23.5% think they could.
- Information failure
    - Individuals may not know where course can be taken privately if they don't get chosen (C99 on baseline, 33% know of alternative)
    - Individual not confident of the quality of a private provider they find on their own (C103=1 or C103=2 (not at all confident or not very confident) on baseline.)
- Supply failure
    - Course offered by ISKUR may not be offered privately: Provider survey – does the provider offer courses to individuals?
- Insurance market failure: do individuals underinvest in training because they think returns are high but risky?
    - Baseline overall willingness to take risks – 11 point scale – baseline question C113.
- Time-inconsistent preferences and high discount rates lead to people putting off training for the future since costs are now and gains are in the future
    - Not measured at baseline, so we will have to test that treatment didn't affect this measure
        - Measure as discount rate above or below median, based on F3.b on follow-up – willingness to take amount today vs 1000 TL in one month.
        - Hyperbolicity – measured as whether discount rate is higher for the one month vs today comparison than for five vs six months.

5. Estimation methodology

Estimation of Treatment Effects

For outcomes in which the same question was asked in both the baseline and follow-up surveys, our main specification will be the following ANCOVA specification:

$$Y_{i,t=1} = \beta_0 + \beta_1 T_i + \pi Y_{i,t=0} + \gamma M_{i,t=0} + X_s'\theta + \varepsilon_i \quad \text{(M1)}$$

Where $Y_{i,t=1}$ is the given outcome variable measured post-treatment, $Y_{i,t=0}$ is its baseline value and $M_{i,t=0}$ a dummy variable indicating whether or not this baseline value is missing, T_i is an indicator for being assigned to treatment, X_s is a vector of randomization strata dummy variables (course*gender*age group) and ε_i is the error term. Since randomization is at the individual-level, conditional on these strata, Huber-White standard errors will be used. β_1 will provide the intent-to-treat effect, which is the effect of being selected to participate in an ISKUR course among the experimental sample. Since not all those who were selected to participate will actually attend the course, and some of the control group may attend classes, we can also estimate the following equation

$$Y_{i,t=1} = \beta_0 + \beta_1 C_i + \pi Y_{i,t=0} + \gamma M_{i,t=0} + X_s'\theta + \varepsilon_i \quad \text{(M2)}$$

where C_i is an indicator for attending course *i*, which is instrumented by assignment to treatment status, T_i. In this case β_1 measures the treatment-on-the-treated – the impact of ISKUR training for those who take this training when selected for it and do not take it otherwise.

In cases where an outcome variable was not collected at baseline, these same specifications will be estimated without the control for baseline outcome.

Estimation of Heterogeneous Treatment Effects

Heterogeneous treatment effects will be estimated by interacting treatment status and all control variables in (M1) or (M2) with the variable of interest Z.

Dealing with Testing for Multiple Outcomes through Standardized Treatment Effects and Adjustments for Multiple Inference

We have a relatively rich set of outcome measures and characteristics with which to explore treatment-effect heterogeneity. To deal with multiple hypothesis testing we employ the two approaches employed by Finkelstein et al. (2010) in their pre-analysis plan for studying the Oregon Health experiment. First, we group our outcome measures into domains, based on the idea that items within a domain are measuring an underlying common factor. Our four domains are employment, individual well-being, household well-being, and empowerment and gender attitudes. Then we sign the outcomes within each domain so that the hypothesized effects go in the same direction, and take a standardized treatment effect within that domain. We follow Kling, Katz and Liebman in constructing this standardized treatment effect.

Secondly, to account for multiple inference within a domain we will compute and report the family-wise error rate adjusted p-values using the Westfall and Young step-down resampling method.

To control for multiple hypothesis testing with respect to the heterogeneity of treatment effects, we will follow the recommendations of Fink, McConnell and Vollmer (2010) and employ the Benjamini and Hochberg (1995) method to minimizes the false non-discovery rate (FNR). We will also limit our examination of treatment effect heterogeneity to the overall standardized employment domain outcome, and to the outcome of currently being employed for 20 or more hours per week.

Procedures for Addressing Missing Data and Questions with Limited Variation

The following sections detail the procedures for addressing the cases of survey attrition, item non-response, and questions with limited variation.

Survey attrition

Depending on response rates and budget, the follow-up survey will potentially use more expensive methods to try and get a subsample of the individuals who can be obtained through the standard survey to respond. If this is done, all data will be probability-reweighted to reflect this.

Let A_i be an indicator of whether individual *i* attrits from the study by not responding to or being able to be contacted for the endline survey. We will first estimate whether attrition is related to treatment status by means of the following regression:

$$A_i = \beta_0 + \beta_1 T_i + X_s'\delta + \varepsilon_i$$

Where X_s are dummy variables for each randomization strata *s* (consisting of course*gender*young or old age). Since randomization is at the individual-level, conditional on these strata, Huber-White standard errors will be used. We will test $\beta_1 = 0$ to determine whether attrition from the survey is related to treatment status or not.

If treatment status is found not to significantly affect attrition at the 5 percent significance level, then all estimation will proceed without any adjustment for attrition. If attrition is found to be related to treatment

status, we postulate that attrition will be higher for the control group, and will employ Lee bounds to obtain bounds on our treatment estimates which are robust to this attrition.

Missing data from item non-response

No imputation for missing data from item non-response at follow-up will be performed. Missing data on baseline variables will be dummied out of the ANCOVA specifications, as detailed above. We will check whether item non-response is correlated with treatment status following the same procedures as for survey attrition, and if it is, construct bounds for our treatment estimates that are robust to this.

Questions with Limited Variation

In order to limit noise caused by variables with minimal variation, questions for which 95 percent of observations have the same value within the relevant sample will be omitted from the analysis and will not be included in any indicators or hypothesis tests. In the event that omission decisions result in the exclusion of all constituent variables for an indicator, the indicator will be not be calculated.

Annex 2: A Successful randomization of the ISKUR vocational trainings

The high capacity of the ISKUR IT services, together with the strong implementation capacity of the regional offices, allowed that the randomization of the applicants into the ISKUR vocational trainings to be carried out by computer. The impact evaluation will compare the labor market outcomes for the trainees and the non trainees before and after the completion of the courses. At the core of this identification strategy is the capacity to randomize the individual participation within each of the courses. The randomization worked as follows: each training provider provided a list of up to 50% more applicants than they had capacity to train. These individuals were then randomly offered a position in the training course using the special information system set up to register applicants into the program. As a result, any differences between treatment and control groups in the full assigned population are due purely to chance. Because the effects of the trainings in the applicant's labor market outcomes are likely to be different depending on the vocational courses, gender, or age of the beneficiaries, the randomization was stratified by these variables.

Although the response rate for the face to face interviews was high, the 5% non-response may have affected the balance between trainees and non-trainees. Since less than 100% of those assigned to treatment and control completed the face to face baseline survey conducted between September 2010 and April 2011, it is worth checking whether treatment and control groups have balanced observable characteristics. The final sample available has a total of 5,318 individuals. Table A1 presents the means for treatment and control group and a p-value for testing the difference in means for 37 different characteristics.

The findings clearly show that the sample is well-balanced on a number of important characteristics that likely also will affect the labor market outcomes of the applicants to the ISKUR trainings. In particular, the treatment and control groups do not statistically differ in terms of gender,

educational levels, participation in prior training courses, ever having worked for pay, beliefs about the likelihood of being employed in one year with and without taking training courses, household asset ownership and household income, and numeracy and risk attitudes. However, there are statistically significant, although small in absolute magnitude, imbalances in some variables: Those in the treatment group are slightly younger on average, and get slightly higher scores on a Raven test (a measure of non-verbal IQ). Since some of the questionnaires were completed after individuals knew their treatment status, some of the other small differences might reflect changes in attitudes and knowledge as a result of treatment assignment. This could explain, for example, why the treatment group is slightly less likely to have sought work in the last 4 weeks, to say they would accept a job at 600 lira, to have worked for pay in the last 4 weeks, and to know where a similar course could be taken privately – individuals who have just found out they are taking the course might be less likely to look for work or to work, and might learn the course is offered privately also. The small significant difference in mental health could also arise from a short-term increase in stress from those starting the new course. As a result, the randomization was successful. It created two largely similar samples in terms of observable characteristics. This finding is reassuring that any differences in outcomes to be observed in the tracking (summer 2011) and in the follow up surveys (2012) will be solely driven by the ISKUR vocational trainings.

Table A1: Test of Randomization for ISKUR Sample			
	Means by Assignment Status		Diff. in Means
	Control	Treatment	p-value
Age	28.24	27.92	0.096
Female	0.63	0.63	0.741
Youth (age <25)	0.38	0.40	0.066
Male Youth	0.16	0.17	0.811
Female Youth	0.21	0.23	0.052
Male >=25	0.21	0.20	0.537
Female >=25	0.42	0.40	0.187
Less than High School Education	0.26	0.26	0.938
University Education	0.14	0.14	0.961
Number of Years Schooling	11.36	11.42	0.548
Has done prior training course in last 5 years	0.26	0.26	0.543
Household Size	4.02	4.02	0.952
Married	0.32	0.32	0.984
Has been unemployed since 2009	0.40	0.38	0.126
Receives unemployment insurance	0.05	0.05	0.907
Has sought work in last 4 weeks	0.51	0.46	0.000
Says would accept a job at 600/month	0.34	0.31	0.026
Says would accept a job at 1000/month	0.59	0.58	0.229
Worked for pay in last 4 weeks	0.14	0.11	0.008
Has ever worked for pay	0.60	0.59	0.439
Total years working for pay	3.36	3.28	0.563
Agrees that mainly interested in course for stipend	0.35	0.33	0.234
Know where course could be taken privately	0.34	0.37	0.037
Percent chance would pay for course if not selected	33.51	34.94	0.125
Percent chance will be employed if take another course	45.05	45.07	0.984
Percent chance will be employed without course	31.79	31.22	0.395
Percent chance will be employed if take ISKUR course	63.64	63.55	0.893
Risk-seeking score (higher = more risk seeking)	6.52	6.42	0.164
Financial risk-seeking score (higher=more risk seeking)	4.77	4.76	0.960
Has health problem that prevents physical work	0.02	0.02	0.720
Mental health index (higher = worse mental health)	11.72	11.98	0.003
Durable asset index	-0.00	0.00	0.978
Household has a computer	0.68	0.69	0.304
Household has internet connection	0.53	0.54	0.240
Household annual income	14,360	14,253	0.761
Gets all 4 numeracy questions right	0.57	0.58	0.420
Raven test score	5.64	5.94	0.001

Source: Authors. Based on the 5,318 observations of ISKUR applicants from the ISKUR-WB baseline survey (2011).
Note: Differences in treatment and control tend to be not statistically significance across a wide range of socio economics characteristics. There are two exceptions: age and probability of being actively looking for a job)