Organisation for Economic Co-operation and Development

For Official Use **English - Or. English**

The Use of Test Scores in Secondary Analysis

A dialogue between data users and data producers

Paris, 14 June 2019

Contact: marco.paccagnella@oecd.org

1. Introduction

1. The empirical literature on human capital has been deeply marked in the last couple of decades by the increasing availability of individual measures of cognitive skills. As the concept of human capital is central in a variety of social sciences (such as economics, education, sociology and psychology), a large number of researchers are now making use of these data.

2. Measures of cognitive skills are most commonly available for students. Test scores derived from students' performance in standardised tests such as PISA, TIMMS, PIRLS, NAEP (just to name the most well-known) are routinely used by researchers in academia and in more policy-oriented institutions, and in some countries are even directly used for accountability purposes.

3. The release of data from the Programme of International Assessment of Adult Competencies (PIAAC) has further increased the use of this kind of data, by allowing researchers to look at the relationship between cognitive skills of adults and labour market outcomes (Hanushek et al., 2015[1]). Previous work on the returns to cognitive skills has relied on data from national longitudinal studies, such as the British Cohort Study or the National Longitudinal Survey of Youth (Heckman, Stixrud and Urzua, 2006[2]; Cunha, Heckman and Schennach, 2010[3]).

4. Large-scale assessments from which cognitive skills measures are collected are typically low-stakes assessments designed by psychometricians to provide an estimate of the distribution of one (or more) cognitive trait in the target population. They are not designed to provide accurate or unbiased measure of individual proficiencies.

5. The correct use of test scores data from large scale assessments requires some knowledge on the process leading to their collection and production. A debate in this sense has already started between economists and psychometricians (Jacob and Rothstein, 2016[4]; Braun and von Davier, 2017[5]). The OECD is organising a Methodological Seminar whose objectives are to give wider publicity to these issues and to foster a dialogue between producers and users of these data.

6. This note aims at setting the stage for the Seminar by highlighting what we believe to be the most important issues at stake. The note is based on our reading of the literature on the topic, as well as on our own practical experience in working with data from large-scale assessments.

2. The use of data from large-scale assessments

7. A distinction is often made between primary and secondary analysis of assessment data. Primary analysis denotes the production of descriptive statistics (mean, standard deviations, and percentiles) that characterise the distribution of proficiency in the target population, or in various subgroups. This is the kind of analysis that is often found in reports that accompany the release of survey and assessment data. This is often the main objectives of the organisations promoting large-scale assessments, and therefore the main focus of the designers of the assessments (which can be denoted as "data producers").

8. Secondary analysis typically takes the form of estimation of statistical models (e.g. linear regressions), with measures of cognitive skills used either as independent or dependent variables.

9. In the former case, cognitive skills are put in relation with a number of inputs that are supposed to contribute to their production (this is often referred to as estimation of the educational production function). An important application of this approach is the estimation of so-called value-added models that aim to quantify the contribution of schools or teachers to students' proficiency. In the latter case, the interest is typically to establish the impact (or the "returns") of cognitive skills on some outcomes of interest. A common implementation of these exercises is the estimation of a Mincerian wage regression aimed at estimating the economic returns to skills.

10. The release of data is often accompanied by technical documentation and guidelines that instruct researchers on how to use the data. These guidelines are however: (a) often overlooked by researchers that engage in secondary analysis; and (b) often too much focused on the primary analysis.

11. The latter point is a direct consequence of the fact that the primary analysis – i.e. an accurate description of the level and distribution of cognitive skills in the population of interest – is typically the primary reason why the assessment was undertaken in the first place.

12. However, the fact that the production of cognitive measures is geared towards what we labelled "primary analysis" can sometime impose limitations on the use of data in secondary analysis.

13. Primary and secondary analysis have different goals, and the methods that are best suited for the former are not necessarily best suited for the latter. At the risk of over-simplifying the issue, primary analysis is mainly about solving a problem of inference, i.e. the efficient estimation of the main parameters (mean and variance) of the distribution of a latent trait (the cognitive skills) in the population of interest. Secondary analysis, on the other hand, is mainly focused on the correct identification of the parameters of a statistical model that aims at uncovering the causal relationship between different variables.

14. Communication between data users interested in secondary analysis and data producers mainly interested in primary analysis is often made difficult by the fact that users and producers typically belong to different disciplines, have few occasions to interact with each others, and often speak slightly different jargon. Still, such a dialogue can be fruitful not only to improve current practices in the use of the data, but also to stimulate a reflection on how current production practices can be modified to increase ease of use and flexibility in the analysis.

3. Features of large scale assessments

15. Cognitive skills measures are produced by administering a cognitive assessment on a given subject to survey participants. Respondents are also typically asked to fill in a more or less extensive background questionnaire, in which ancillary information on the respondents is collected.

16. Different assessments might differ in their sampling design, as well as in the way the assessment is constructed and scored, and in how the final score is computed. Nevertheless, the best known international large scale assessments (LSA) do share many features and make use of state-of-the-art psychometric techniques to generate the test scores of interest. Below we briefly describe the main characteristics of such LSA, such as TIMMS, PIRLS, PISA and PIAAC.

17. Two features of international large scale assessments have an important bearing on the statistical use of the data they produce: (a) a complex sampling design; (b) a psychometric model used to estimate individual proficiency.

18. Both features are instrumental to the primary objective of international LSA, that is to provide estimates of the distribution of cognitive skills in the target population. In particular, they both concur in increasing the overall efficiency of the survey, by maximising the amount of information collected and minimising the costs of data collection.

19. A complex sampling design is needed to limit the number of locations where the assessment is taken, therefore minimising the cost of data collection. School-based surveys sample schools first, and then students (PISA) or classes (TIMMS and PIRLS) within schools. In the case of household-based surveys like PIAAC, each country adopts its own sampling design, which is typically a multi-stratified design with two, three or four stages of sampling. There is often cross-country variation in sampling frames, which can be based on population registers, social security archives or other sources.

20. Similarly, psychometric models are used to maximise the efficiency of the assessment, by allowing efficient use of the information extracted from a necessarily limited pool of items that can be administered to any single respondent. The current state-of-the-art consists in having adaptive tests, where the difficulty of the items each respondent is administered depends on expected proficiency and/or observed performance in the first stages of the assessment. Items targeted to the respondent's proficiency are efficient because they are more discriminating than excessively easy or excessively difficult items.

21. As different respondents answer different items, the actual proficiency of respondents on a common scale needs to be estimated. To do so, IRT models are used for item calibration. IRT models relate the probability of giving a correct answer to a given item to a latent (unobserved) parameter which represents the ability of the respondent, as well as on parameters that characterise the item (in the most common cases the model contains three parameters that denote the difficulty of the item, its ability to discriminate between high- and low-ability respondents, and the extent to which a correct response could be given by random guessing).

22. As tests contain relatively few items, individual (posterior) estimates of the latent construct are often imprecise. One popular way to address this problem is to draw several

"plausible values" from the posterior distribution in order to better characterise the uncertainty associated with the estimate. This approach is closely linked to methods of multiple imputation for missing data (Little and Rubin, 2014[6]; Rubin, 1987[7]).

23. In order to further increase the precision of the population estimates, the prior distribution is often derived from a "population model" that links the latent ability of respondents to background characteristics (see Von Davier, Gonzalez and Mislevy (2009[8]). The posterior distribution from which plausible values are drawn will therefore depend on individual characteristics that are correlated with latent ability. This approach ensures unbiased estimation of group differences for those characteristics that are part of the imputation model (Mislevy, 1991[9]). For this reason, the population model typically contains a very large number of conditioning variables.

4. Implications for primary and secondary analysis

24. The way test scores are estimated require some care in their use for primary and secondary analyses. Psychometricians (Carstens and Hastedt, 2010[10]; Braun and von Davier, 2017[5]; Von Davier, Gonzalez and Mislevy, 2009[8]) typically recommend that:

- provided final survey weights are used, both for computing descriptive statistics and for running regressions, in order to obtain correct estimates of the population statistics taking into account the complex survey design;
- the entire set of plausible values is used, together with replication weights, in order to correctly take into account the measurement and the imputation error associated with test scores.
- Disregarding such practices will lead to errors in the computations of descriptive statistics such as the average test scores of the populations of interest.

25. In the case of secondary (regression) analysis, issues become more complex and often not so clear-cut. The issues can be broadly classified in three categories: identification, scaling, and inference.

4.1.1. Identification of the parameters of interest

26. Jacob and Rothstein (2016[4]) intend to inform and warn economists about the implications of using test scores in regression analyses. They stress in particular the fact that test scores estimated via psychometric models are not intended to be unbiased estimates of the latent construct of interest at the individual level. As plausible values are random drawn from the posterior distribution estimated by the population model, individual values are typically shrunk towards the average value of the group they belong to, where the group is defined by the (often very large) number of characteristics that were used in the population model.

27. As a result, the measurement error that surrounds individual estimates of cognitive skills cannot be interpreted as "classical" measurement error, uncorrelated with other observables and unobservable characteristics of the respondent. The consequence is that the use of test scores in regression analysis is likely to lead to biased estimates of the relationships of interest between test scores and other variables.

28. When test scores are used as dependent variable, the estimated coefficients of the model the researcher is interested in are unbiased only if all the explanatory variables in the model were also included in the population model used to estimate the test scores.

29. When test scores are used as independent variable, the conditions to obtain unbiased coefficient estimates are more restrictive: the population model should include all the variables used in the research model, but no other variables that are correlated with the dependent variable of the research model.

30. In this sense, the choices made by psychometricians in the estimation of test scores through the population model constrain the secondary analysis that researchers might want to undertake.

31. A possible solution to these issues would be to use Mixed Effects Structural Equations Models (Schofield et al., 2015[11]; Schofield, 2014[12]; Schofield, 2015[13];

Junker, Schofield and Taylor, 2012[14]), estimating jointly the research model and the test scores. This however would require information on the responses given by test-takers to each item of the assessment, something that is either not always publicly available or that is in any case not taken into account by most researchers performing secondary analysis.

32. Different approaches in labour economics have relied on the use of factor models. Identification of the effects of cognitive and non-cognitive latent traits on labour market outcomes is discussed for instance in Heckman, Stixrud and Urzua (2006[2]), while Cunha, Heckman and Schennach (2010[3]) use dynamic factor models to estimate the production function of cognitive and non-cognitive skills, accounting for measurement error, the relationship between the different latent traits, and the endogeneity of inputs such as parental investments.

4.1.2. Scaling

33. Jacob and Rothstein (2016[4]) also point out that test scores cannot typically be interpreted as being scaled on an interval scale. This is an issue that is common to all scales that attempt to estimate a latent (unobservable) construct. In this respect, Bond and Lang (2013[15]) observe how arbitrary scaling decisions can affect secondary analysis based on test scores. They suggest that scales can only be interpreted in an ordinal sense, and that researchers should verify that their results are robust to arbitrary monotonic transformations of the scores. A possible solution is to anchor test scores in terms of future, policy-relevant outcomes, such as years of schooling (Bond and Lang, 2017[16]). Anchoring functions are also used by Cunha and Heckman (2008[17]) and by Cunha, Heckman and Schennach (2010[3]).

4.1.3. Inference

34. Jerrim et al. (2017[18]) note that economists typically do not use the provided survey weights in regression analysis using test scores data, and illustrates how this choice can change the estimated coefficients and standard errors. They argue that, in international databases, a pondered choice should be made between the use of final student weights or of senate weights. The former scale the sample up to the population size of each country, with the results that larger countries will contribute more to the analysis. Senate weights, on the contrary, rescale the weights so that each country contribute equally to the analysis. There is no right or wrong decision in this respect, but the researcher should make a conscious choice and decide which is the underlying population of interest for the analysis.

35. The issue of weighting, however, is just another example of the differences between primary and secondary analysis. While there is no doubt that survey weights should be used to obtain correct statistics to describe the distribution of skills in the population, the issue is much less clear-cut in the context of a regression analysis (Solon, Haider and Wooldridge, 2015[19]).

36. Survey weights do help in the computation of standard errors, as they take into account the fact that the stratified sampling design induce correlation between units sampled from the same stratum. On the other hand, when it comes to identification, a choice about the appropriateness of using weights should consider: (a) whether or not the weights help to correct for endogenous sampling (i.e. whether the sampling probabilities are conditionally independent of the error term in the regression equation); and (b) whether or not the statistical model allows for heterogeneous effects (in which case weights would be needed to correctly estimate the population average of the heterogeneous effects).

References

Bollinger, C. (2003), "Measurement Error in Human Capital and the Black-White Wage Gap", *Review of Economics and Statistics*, Vol. 85/3, pp. 578-585, http://dx.doi.org/10.1162/003465303322369731. [20]

Bond, T. and K. Lang (2017), "The Black-White Education Scaled Test-Score Gap in Grades K-7", *Journal of Human Resources*, pp. 0916-8242R, http://dx.doi.org/10.3368/jhr.53.4.0916.8242R. [16]

Bond, T. and K. Lang (2013), "The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results", *The Review of Economics and Statistics*, Vol. 95/5, pp. 1468-1479, https://www.mitpressjournals.org/doi/pdf/10.1162/REST_a_00370 (accessed on 21 September 2018). [15]

Braun, H. and M. von Davier (2017), "The use of test scores from large-scale assessment surveys: psychometric and statistical considerations", *Large-scale Assessments in Education*, http://dx.doi.org/10.1186/s40536-017-0050-x. [5]

Briggs, D. (2008), "Using explanatory item response models to analyze group differences in science achievement", *Applied Measurement in Education*, http://dx.doi.org/10.1080/08957340801926086. [21]

Carstens, R. and D. Hastedt (2010), *The effect of not using plausible values when they should be: An illustration using TIMSS 2007 grade 8 mathematics data*. [10]

Cunha, F. and J. Heckman (2008), "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation", *Journal of Human Resources*, Vol. 43/4, pp. 738-782, http://dx.doi.org/10.3368/jhr.43.4.738. [17]

Cunha, F., J. Heckman and S. Schennach (2010), "Estimating the Technology of Cognitive and Noncognitive Skill Formation", *Econometrica*, http://dx.doi.org/10.3982/ECTA6551. [3]

Hanushek, E. et al. (2015), "Returns to skills around the world: Evidence from PIAAC", *European Economic Review*, Vol. 73, pp. 103-130, http://dx.doi.org/10.1016/j.euroecorev.2014.10.006. [1]

Heckman, J., J. Stixrud and S. Urzua (2006), "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior", *Journal of Labor Economics*, Vol. 24/3, pp. 411-482, http://dx.doi.org/0734-306X/2006/2403-0003$10.00. [2]

Jacob, B. and J. Rothstein (2016), "The Measurement of Student Ability in Modern Assessment Systems", *Journal of Economic Perspectives*, Vol. 30/3, pp. 85-108, http://dx.doi.org/10.1257/jep.30.3.85. [4]

Jerrim, J. et al. (2017), "What happens when econometrics and psychometrics collide? An example using the PISA data", *Economics of Education Review*, http://dx.doi.org/10.1016/j.econedurev.2017.09.007. [18]

Junker, B., L. Schofield and L. Taylor (2012), "The use of cognitive ability measures as explanatory variables in regression analysis", *IZA Journal of Labor Economics*, http://dx.doi.org/10.1186/2193-8997-1-4. [14]

Little, R. and D. Rubin (2014), *Statistical Analysis with Missing Data.*, Wiley, https://books.google.fr/books?hl=fr&lr=&id=AyVeBAAAQBAJ&oi=fnd&pg=PT8&ots=uyTV6xnQgH&sig=KiZlYYnY6na-QpEHfZ4h4lDtQtk&redir_esc=y#v=onepage&q&f=false (accessed on 05 November 2018). [6]

Mislevy, R. (1991), "Randomization-based inference about latent variables from complex samples", *Psychometrika*, Vol. 56/2, pp. 177-196, http://dx.doi.org/10.1007/BF02294457. [9]

Rubin, D. (ed.) (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Inc., http://dx.doi.org/10.1002/9780470316696. [7]

Schofield, L. (2015), "Correcting for measurement error in latent variables used as predictors", *Annals of Applied Statistics*, http://dx.doi.org/10.1214/15-AOAS877. [13]

Schofield, L. (2014), "Measurement error in the AFQT in the NLSY79", *Economics Letters*, http://dx.doi.org/10.1016/j.econlet.2014.02.026. [12]

Schofield, L. et al. (2015), "Predictive Inference Using Latent Variables with Covariates", *Psychometrika*, http://dx.doi.org/10.1007/s11336-014-9415-z. [11]

Solon, G., S. Haider and J. Wooldridge (2015), "What Are We Weighting For?", *Journal of Human Resources*, Vol. 50/2, pp. 301-316, http://jhr.uwpress.org/content/50/2/301.full.pdf (accessed on 18 October 2018). [19]

Von Davier, M., E. Gonzalez and R. Mislevy (2009), "What are plausible values and why are they useful?", in *Issues and Methodologies in Large-Scale Assessments*. [8]