THE PROMISES AND PERILS OF DIAGNOSTIC LABELS

E. Abdurakhmanova

National Research University Higher School of Economics (RUSSIAN FEDERATION)

Abstract

People rely on diagnostic labels to identify and solve problems, and psychological and sociological research has documented adverse consequences of diagnostic labels on individual outcomes. This study examines three fundamental questions: (1) To what extent does the introduction of diagnostic labels affect students' academic performance towards the end of the school year? (2) How do the effects of diagnostic labels for groups and individuals depend on the proportion of other individuals who are also labelled in the group? (3) Is this effect mediated by the length of time spent by the teacher for a particular student? To answer these questions, a large-scale cluster randomized controlled trial among 4460 first-grade school children and 186 schools in one region of Russia was conducted. Schools were randomly assigned to one of two treatment arms, where teachers received information about the performance of their students (a) without a corresponding diagnostic label for their students’ performance or (b) with a label for students as being “basic” / “proficient” in their academic or “developing” / “mature” in their behavioral performance. Results demonstrate that diagnostic labels do not directly affect students' academic performance. However, diagnostic labels have an indirect effect on the academic performance of students from a specific group. It was also found that the lower a student's academic performance was, the more time the teacher devoted to the student.

Keywords: Diagnostic labels, teachers expectations, students achievement, START, experiment.

# INTRODUCTION [Arial, 12-point, bold, upper case and left alig.]

Diagnostic labels are used across hospitals, courts of law, schools, and other key domains of human life. These labels are designed to coordinate social action, enabling people to distribute scarce resources or apply common strategies to address problems. Of particular interest are diagnostic labels that are applied toward individuals: such as diagnoses of children as having “autism,” or diagnoses of students as displaying “proficient” or “basic” academic performance ([1], [2], [3]). Research across sociology and psychology suggests that diagnostic labels can stigmatize individuals by transforming a situation that could otherwise be perceived as normal into something problematic ([4], [2], [5], [6]). Diagnostic labels may also lead to self-fulfilling prophecies, such as when students who are labelled as “behind grade level” are expected to perform worse and therefore not challenged enough – what some have called the “soft bigotry of low expectations” ([7]~~,~~ [8]). As such, diagnostic labels do not merely identify problems but often also crystallize and even exacerbate differences across individuals, thus creating problems of their own.

This study intervenes in this literature by investigating three fundamental, questions:

1. To what extent does the introduction of diagnostic labels affect students' academic performance towards the end of the school year?

2. How do the effects of diagnostic labels for groups and individuals depend on the proportion of other individuals who are also labelled in the group?

3. Is this effect mediated by the length of time spent by the teacher for a particular student?

# METHODOLOGY

## Research Design

In order to answer the research questions posed, a cluster randomized controlled trial (RCT) was organized with the participation of 4460 first grade schoolchildren (mean age of children = 7.4 years, proportion female = 0.50) across 186 schools in one of the regions of Russia. The selected schools were asked to select classes for participation in the study. In the vast majority of schools, the study selected one class with one teacher, but some schools had multiple first grade classes, each with its own teacher.

The experiment took place in four stages.

The first stage took place in October 2019 (closer to the beginning of the academic year). At this stage, the design of the study was developed, the approval of the ethical committee was obtained, a sample of schools participating in the study was selected, and the consent of the parents of students to participate in the study was obtained. Next, a baseline assessment of the academic performance and social and emotional skills of first graders was carried out, and a survey of teachers of first graders was conducted.

The second stage (November 2019) included the randomization of schools according to different treatment conditions. Schools were randomized to one of two RCT groups - control or experimental. Fifteen tests comparing average variable values among the treatment and control arms were conducted to test for balance on baseline observables across the treatment arms. Out of the 15 tests (e.g. baseline math score, baseline reading score, class size), only one was statistically significant (different from zero) at the 5% level (a binary for whether the student reports that the teacher knows how well they are learning).

The third stage (December 2019) is the conduct of the very treatment (intervention) in randomly selected schools that fell into the experimental group. In the control group schools, teachers were provided with numeric measures of student academic and behavioral performance (on a scale between 0 to 100) on the baseline assessments. In the treatment group schools, teachers were provided with the numerical measures of student cognitive and non-cognitive performance and an additional report with diagnostic labels of student performance as well. Specifically, students across the sample who scored below 50 in their academic assessment were labeled as “basic” rather than “proficient”, and students in the sample who scored below 50 in their behavioral assessment were labeled as “developing” rather than “mature”. These assessments and labels were designed and pre-tested by Russian academics to be easily interpretable by teachers. For instance, teachers were invited to focus groups to offer feedback on different labels, and these labels were the most easily understood.

Thus, four groups were allocated.

* Group 1 included children with proficient academic performance and mature social and emotional skills.
* Group 2 – children with proficient academic performance and developing social and emotional skills.
* Group 3 is a group of children with basic academic performance and mature social and emotional skills.
* Group 4 included children with basic academic performance and developing social and emotional skills.

An important feature of this design is as follows. The teachers from the schools in the control group received the same information about their students' test scores as the teachers in the experimental group, but did not receive information from the diagnostic labels of the students. It follows that any potential differences in the results of students from the two groups may be due to the submission of additional reports to the teachers from the experimental group by the diagnostic labels of the students, in other words, the consequence of treatment.

The fourth stage consisted in conducting a final survey of teachers and a final assessment of students from the same schools. This phase was scheduled for April 2020, but was postponed to September 2020 due to the coronavirus pandemic. Teachers were asked to estimate the time they directed toward each student. The teacher's self-reported effort toward individual students was collected as an important mediator to explore in the study.

## Assessment of the academic and behavioral performance

Assessment of academic and behavioral performance was carried out using the START tool. START is a tool for starting diagnostics of students at the school entrance and assessing their individual progress after finishing the first grade, developed at the HSE Institute of Education based on the iPIPS tool. The tool has good psychometric properties and validity ([9]). This tool allows you to assess the mathematical and reading performance and social and emotional skills of the child.

Mathematic Performance (Math). The mathematic performance test included five types of tasks: geometric sequences, arithmetic sequences, number line, the concept of part and whole, calculation skills.

Reading Performance (Reading). The reading performance test includes four types of tasks: letter recognition, word decoding, reading decoding, and comprehension.

Personal social and emotional development (PSED) questionnaire was used to assess the student's social and emotional skills, which is completed by the teacher. The questionnaire has high psychometric characteristics ([10]). PSED includes 11 parameters that form two scales "Behavior at school" and "Communication" ([10]).

To scale all of the performance tests the dichotomous Rasch model was used ([11]). The analysis of the scales was carried out using the Winsteps software package ([12]). All of the scales had the features of unidimensional, with items fitted to the model, and sufficiently high-test reliability (Cronbach’s alpha) was obtained. The reliability of the test (Cronbach's alpha and the reliability of the test subjects) ranged from 0.79 to 0.98. Assignments do not show the effects of the floor or ceiling.

## Statistical Approach

In order to answer the first and the second questions a two-level multilevel regression analysis (Models 1-4), which is appropriate for analysing data that have a hierarchical structure, was performed (e.g., [13]). The obtained data had a hierarchical structure as follows: the students were nested within classes, and the classes were nested within schools. A two-level hierarchy was used (students on the first level and classes on the second level) because many schools in the sample had only one class (75%); thus, the variability between classes within these schools was could not estimate.

To test if diagnostic labels have a direct effect on academic performance or this effect mediated by the time spent by the teacher on a particular student (TS) a two-level multilevel mediation analysis was done (Models 5-10). TS was measured using an ordinal scale from 1 to 3, where 1 is much less than with my other students, 2 is not more but not less than with other students, and 3 is much more than with my other students. The mediation (indirect) effect was computed as a product of the coefficients for variables diagnostic labels and TS. The standard errors for direct, indirect, and total effects were estimated by using the Delta method [14].

The dependent variables were mathematic and reading performance at the end of grade 1.

The independent variables were:

-treatment variable was considered as "the information about a specific diagnostic label". This variable contains five levels:

* Label 0 means that the teacher did not receive information about the diagnostic labels;
* Label 1 - the teacher received information about the diagnostic label of the student and the student belongs to group 1;
* Label 2 - the teacher received information about the student's diagnostic label and the student belongs to group 2;
* Label 3 - the teacher received information about the student's diagnostic label and the student belongs to group 3;
* Label 4 - the teacher has received information about the student's diagnostic label and the student belongs to group 4.

- “the proportions of students with a specific diagnostic label”.

* Proportion 1 is a proportion of students with a diagnostic label 1 in class.
* Proportion 2 is a proportion of students with a diagnostic label 2.
* Proportion 3 is a proportion of students with a diagnostic label 3.
* Proportion 4 is a proportion of students with a diagnostic label 4 in class. Proportion 4 was considered as a reference group.

As a covariate in the regression models, it was decided to include variables class size (centered variable grand mean), teacher experience (centered variable grand mean), teacher education (0 - no higher education, 1 - with higher education) at the class level; gender of the child (0 - boy, 1 - girl), children's scores at the beginning of the school year on all four scales (mathematics, reading, behavior, and communication) at the student level.

# RESULTS

Descriptive statistics are presented in Tab. 2.

Table 2. Descriptive statistics of performance tests.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean | SD | Min. | Max. |
| Math (baseline) | 50.26 | 9.81 | 11.69 | 93.91 |
| Reading (baseline) | 50.40 | 9.70 | 25.21 | 68.17 |
| Behavior (baseline) | 50.50 | 9.77 | 23.91 | 69.39 |
| Communication (baseline) | 50.40 | 9.86 | 17.17 | 73.88 |
| Math (endline) | 60.98 | 11.21 | 25.89 | 94.56 |
| Reading (endline) | 57.52 | 8.67 | 25.22 | 72.63 |

## To what extent does the introduction of diagnostic labels affect students' academic performance towards the end of the school year?

On the first step, to what extent the introduction of diagnostic labels affects students' academic performance towards the end of the school year was examined. The results in Tab. 3 show that diagnostic labels do not affect students' academic performance towards the end of the school year.

Table 3. The results of models with predictors for Math and Reading (endline).

|  |  |  |
| --- | --- | --- |
|  | Model 1  *B (s.e.)* | Model 2  *B (s.e.)* |
|  | *Math* | *Reading* |
| (Intercept) | 13.16\*\*\* (1.17) | 19.0351\*\*\* (0.88) |
| *Student-level variables* | | |
| Label 1 | 0.27 (0.60) | 0.60 (0.46) |
| Label 2 | 0.06 (0.64) | 0.13 (0.49) |
| Label 3 | 0.31 (0.64) | -0.58 (0.49) |
| Label 4 | 0.05 (0.59) | -0.27 (0.45) |
| Math (baseline) | 0.73\*\*\* (0.01) | 0.10\*\*\* (0.01) |
| Reading (baseline) | 0.09\*\*\* (0.02) | 0.58\*\*\* (0.01) |
| Behavior (baseline) | 0.06\*\*\* (0.02) | 0.03\*\* (0.01) |
| Communication (baseline) | 0.09\*\*\* (0.01) | 0.03\*\* (0.01) |
| Gender | -2.05\*\*\* (0.22) | 0.20 (0.17) |
| *Class-level variables* | | |
| Class size | -0.08\* (0.03) | -0.12\*\*\* (0.02) |
| Teacher experience | 0.02 (0.02) | 0.01 (0.02) |
| Teacher education | -0.10 (0.69) | 0.52 (0.52) |

*Notes: 1) \*\*\* p<.001, \*\* p < .01, \* p<.05*

*2) Label 0 is a referent group*

## How do the effects of diagnostic labels for groups and individuals depend on the proportion of other individuals who are also labelled in the group?

The results of analysis of how the effects of diagnostic labels for groups and individuals depend on the proportion of other individuals are presented in Tab. 4. The results demonstrate that the effects of diagnostic labels for groups and individuals do not depend on the proportion of other individuals with the particular label.

Table 4. The results of models with interactions for Math and Reading (endline).

|  |  |  |
| --- | --- | --- |
|  | Model 3  *B (s.e.)* | Model 4  *B (s.e.)* |
|  | *Math* | *Reading* |
| (Intercept) | 14.71 (1.60)\*\*\* | 20.64\*\*\* (1.21) |
| *Student-level variables* | | |
| Label 1 | -0.39 (2.13) | -0.86 (1.61) |
| Label 2 | -1.89 (2.23) | -2.01 (1.68) |
| Label 3 | 1.09 (2.29) | -2.04 (1.72) |
| Label 4 | -1.56 (1.70) | -0.73 (1.28) |
| Math (baseline) | 0.73 (0.01)\*\*\* | 0.10\*\*\* (0.01) |
| Reading (baseline) | 0.09 (0.02)\*\*\* | 0.58\*\*\* (0.01) |
| Behavior (baseline) | 0.06 (0.02)\*\*\* | 0.04\*\* (0.01) |
| Communication (baseline) | 0.09 (0.01)\*\*\* | 0.04\*\* (0.01) |
| Gender | -2.06 (0.22)\*\*\* | 0.18 (0.17) |
| *Class-level variables* | | |
| Class size | -0.08\* (0.03) | -0.12\*\*\* (0.03) |
| Proportion 1 | -1.55 (2.11) | -2.03 (1.59) |
| Proportion 2 | -2.33 (2.66) | -1.67 (2.01) |
| Proportion 3 | -4.21 (2.40) | -4.06\* (1.81) |
| Teacher experience | 0.01 (0.02) | 0.00 (0.02) |
| Teacher education | -0.06 (0.68) | 0.54 (0.52) |
| *Interaction terms* | | |
| Proportion l 1 \* Label 1 | 0.27 (3.17) | 0.82 (2.39) |
| Proportion 1 \* Label 2 | 1.54 (3.41) | 1.85 (2.57) |
| Proportion 1 \* Label 3 | -3.53 (3.92) | 1.65 (2.95) |
| Proportion 1 \* Label 4 | 4.07 (3.25) | 0.96 (2.45) |
| Proportion 2 \* Label 1 | 1.26 (4.33) | 3.27 (3.26) |
| Proportion 2 \* Label 2 | 3.32 (4.38) | 3.93 (3.30) |
| Proportion 2 \* Label 3 | -0.28 (5.62) | 1.67 (4.23) |
| Proportion 2 \* Label 4 | 2.47 (4.34) | 0.81 (3.27) |
| Proportion 3 \* Label 1 | 2.21 (4.18) | 2.85 (3.14) |
| Proportion 3 \* Label 2 | 5.01 (4.98) | 3.76 (3.75) |
| Proportion 3 \* Label 3 | 0.72 (3.99) | 2.36 (3.01) |
| Proportion 3 \* Label 4 | 0.48 (3.73) | -1.05 (2.81) |

*Notes: 1) \*\*\* p<.001, \*\* p < .01, \* p<.05*

*2) Label 0 is a referent group*

*3) Proportion 4 is a referent group*

## Is the effect of diagnostic labels mediated by the length of time spent by the teacher for a particular student?

The results of the effects of TS on academic performance are presented in Tab. 5. It was found that a student's academic performance negatively associated with a teacher's time spending with the student than with other students. This also could be interpreted that teachers were more likely to report spending more time with a particular student if the student had lower academic performance.

*Table 5. The effects of time spent (TS) by the teacher for a particular student on Math and Reading (endline)*

|  |  |  |
| --- | --- | --- |
|  | Model 5  *B (s.e.)* | Model 6  *B (s.e.)* |
| Effect of TS on academic performance | *Math* | *Reading* |
| much less than with my other students [95% CI] | 0.31 [-0.36; 0.97] | 0.05 [-0.45; 0.55] |
| much more than with my other students [95% CI] | -1.86\*\*\* [-2.42; -1.29] | -1.13\*\*\* [-1.55; -0.70] |

*Notes: 1) \*\*\* p<.001, \*\* p < .01, \* p<.05*

*2*) *TS - not more but not less than with other students is a referent group*

The results of the effect of diagnostic labels on TS are showed in Tab. 6. The results show that if a teacher received information that a student had a Label 2 (student was from group 2), then, compared with the control group, such teachers were less likely to report that they spent more time with such children.

*Table 6. The effects of diagnostic labels on the time spent by the teacher on a particular student (TS)*

|  |  |  |
| --- | --- | --- |
|  | Model 7  *B (s.e.)* | Model 8  *B (s.e.)* |
|  | *Math* | *Reading* |
| *TS - much less than with my other students* | | |
| Effect of Label 1 on TS [95% CI] | -0.01 [-1.33; 1.30] | -0.01 [-1.33; 1.31] |
| Effect of Label 2 on TS [95% CI] | -0.75 [-2.11; 0.60] | -0.75 [-2.11; 0.61] |
| Effect of Label 3 on TS [95% CI] | -0.43 [-1.80; 0.93] | -0.43 [-1.80; 0.94] |
| Effect of Label 4 on TS [95% CI] | -1.14 [-2.50; 0.23] | -1.12 [-2.49; 0.24] |
| *TS - much more than with my other students* | | |
| Effect of Label 1 on TS [95% CI] | 0.06 [-0.35; 0.47] | 0.09 [-0.33; 0.50] |
| Effect of Label 2 on TS [95% CI] | -0.48\* [-0.88; -0.08] | -0.47\* [-0.87; -0.07] |
| Effect of Label 3 on TS [95% CI] | -0.35 [-0.78; 0.09] | -0.33 [-0.76; 0.10] |
| Effect of Label 4 on TS [95% CI] | -0.28 [-0.63; 0.06] | -0.27 [-0.62; 0.09] |

*Notes: 1) \*\*\* p<.001, \*\* p < .01, \* p<.05*

*2) Label 0 is a referent group*

*3) TS - not more but not less than with other students is a referent group*

The results of the mediation analysis are presented in Tab. 7. The obtained results show that a specific diagnostic label 2 (group of children with proficient academic performance and developing social and emotional skills) indirectly affects students' performance in both math and reading. This effect is mediated by the time spent by the teacher for a particular student.

*Table 7. The mediating effects of time spent by the teacher for a particular student (TS) on the relationship between the effect of diagnostic labels and Math and Reading (endline)*

|  |  |  |
| --- | --- | --- |
|  | Model 9  *B (s.e.)* | Model 10  *B (s.e.)* |
|  | *Math* | *Reading* |
| *Mediator - much less than with my other students* | | |
| *Label 1* | | |
| Indirect effect [95% CI] | 0.00 [-0.41; 0.40] | 0.00 [-0.06; 0.06] |
| Direct effect [95% CI] | 0.26 [-0.91; 1.44] | 0.63 [-0.27; 1.53] |
| Total effect [95% CI] | 0.26 [-0.98; 1.50] | 0.63 [-0.27; 1.53] |
| *Label 2* | | |
| Indirect effect [95% CI] | -0.23 [-0.88; 0.42] | -0.04 [-0.42; 0.35] |
| Direct effect [95% CI] | -0.12 [-1.37; 1.13] | 0.05 [-0.90; 1.01] |
| Total effect [95% CI] | -0.35 [-1.76; 1.05] | 0.01 [-1.01; 1.04] |
| *Label 3* | | |
| Indirect effect [95% CI] | -0.13 [-0.64; 0.38] | -0.02 [-0.25; 0.21] |
| Direct effect [95% CI] | 0.15 [-1.11; 1.41] | -0.65 [-1.60; 0.31] |
| Total effect [95% CI] | 0.02 [-1.34; 1.37] | -0.67 [-1.65; 0.32] |
| *Label 4* | | |
| Indirect effect [95% CI] | -0.35 [-1.21; 0.52] | -0.05 [-0.62; 0.51] |
| Direct effect [95% CI] | 0.00 [-1.16; 1.16] | -0.27 [-1.15; 0.62] |
| Total effect [95% CI] | -0.35 [-1.78; 1.09] | -0.32 [-1.37; 0.72] |
| *Mediator - much more than with my other students* | | |
| *Label 1* | | |
| Indirect effect [95% CI] | -0.11 [-0.89; 0.65] | -0.10 [-0.57; 0.37] |
| Direct effect [95% CI] | 0.26 [-0.91; 1.44] | 0.63 [-0.27;0 1.53] |
| Total effect [95% CI] | 0.15 [-1.25; 1.55] | 0.53 [-0.48; 1.54] |
| *Label 2* | | |
| Indirect effect [95% CI] | 0.90\* [0.11; 1.69] | 0.53\* [0.03; 1.02] |
| Direct effect [95% CI] | -0.12 [-1.37; 1.13] | 0.05 [-0.90; 1.01] |
| Total effect [95% CI] | 0.78 [-0.69; 2.25] | 0.58 [-0.49; 1.65] |
| *Label 3* | | |
| Indirect effect [95% CI] | 0.64 [-0.18; 1.47] | 0.37 [-0.14; 0.88] |
| Direct effect [95% CI] | 0.15 [-1.11; 1.41] | -0.65 [-1.60; 0.31] |
| Total effect [95% CI] | 0.79 [-0.71; 2.29] | -0.27 [-1.36; 0.81] |
| *Label 4* | | |
| Indirect effect [95% CI] | 0.53 [-0.14; 1.19] | 0.30 [-0.11; 0.71] |
| Direct effect [95% CI] | 0.00 [-1.16; 1.16] | -0.27 [-1.15; 0.62] |
| Total effect [95% CI] | 0.53 [-0.81; 1.86] | 0.03 [-0.94; 1.01] |

*Notes: \*\*\* p<.001, \*\* p < .01, \* p<.05*

# CONCLUSIONS

The results show that diagnostic labels have no direct effect on student academic performance in both math and reading. Similarly, the effect of diagnostic labels not associated with the proportion of children in the group with a specific label. It was also found that the lower the student's academic performance, the more time the teacher spent on the student. At the same, two-level mediation analysis showed that diagnostic labels have an indirect effect on a specific group of children, namely children with proficient academic performance and developing social and emotional skills. However, this result could be only a statistical artefact. Thus the additional studies need to conduct to examine it is truly mediation or not.

ACKNOWLEDGEMENTS

The reported study was funded by RFBR, project number 19-29-14110. I thank Prashant Loyalka, an associate professor, Stanford University, and James Y. Chu, a postdoctoral fellow, Stanford University, for assistance for discussing the experimental study design.

REFERENCES

1. B. Algozzine, and J. Sutherland, "The “Learning Disabilities” Label: An Experimental Analysis." *Contemporary Educational Psychology* 2(3): 292-297, 1977.
2. B. G. Link, and J. C. Phelan, "Conceptualizing Stigma." *Annual Review of Sociology* 27(1): 363-385, 2001.
3. J. P.Papay, R. J. Murnane, and J. B. Willett, "The Impact of Test Score Labels on Human-Capital Investment Decisions." *Journal of Human Resources* 51(2): 357-388, 2016.
4. P. Conrad, *Deviance and Medicalization: From Badness to Sickness*. Philadelphia: Temple Univ. Press, 1992.
5. P. W. Corrigan, "How Clinical Diagnosis Might Exacerbate the Stigma of Mental Illness." *Social Work* 52(1): 31-39, 2007.
6. B. G. Link, Understanding labeling effects in the area of mental disorders: An assessment of the effects of expectations of rejection. *American Sociological Review, 52*(1), 96–112, 1987. [https://doi.org/10.2307/2095395](https://psycnet.apa.org/doi/10.2307/2095395)
7. R. Rosenthal, and L. Jacobson, "Pygmalion in the Classroom." *The Urban Review* 3(1): 16-20, 1968.
8. R. Rosenthal, and D. B. Rubin, "Issues in Summarizing the First 345 Studies of Interpersonal Expectancy Effects." Behavioral and Brain Sciences 1(3): 410-415, 1978.
9. E. Orel, I. Brun, E. Kardanova, and I. Antipkina, "Developmental Patterns of Cognitive and Non-Cognitive Skills of Russian First-Graders." *International Journal of Early Childhood* 50(3): 297-314, 2018.
10. .E. Orel, and A. Ponomareva, Patterns of social and emotional development of a first grader at the entrance to school // Psychology. Journal of the Higher School of Economics. Vol. 15, No. 1, S. 107-127, 2018.
11. G. Rasch, *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche, 1960.
12. J. M. Linacre, Winsteps (Version 3.73) [computer software]. Winsteps.com., 2011
13. J*.* Hox, *Multilevel analysis: Techniques and applications,* Mahwah, NJ: Lawrence Erlbaum, 2002. [http://dx.doi.org/10.1037/0022-0663.97.2.184](https://psycnet.apa.org/doi/10.1037/0022-0663.97.2.184).
14. G. King, M. Tomz, and J. Wittenberg, Making the Most of Statistical Analyses: Improving Interpretation and Presentation. American Journal of Political Science, 44(2), 347, 2000, doi:10.2307/2669316.