Annex A8. How much effort do students put into the PISA test?

Performance on school tests reflects what students know and can do. They also show how quickly students process information and how motivated they are to do well on the test.

To encourage students who sit the PISA test to do their best through to the end of the assessment, schools and students are reminded how important the study is for their country. At the beginning of the test session, the test administrator reads a script that includes the following sentence:

> *"This is an important study because it will tell us about what you have been learning and what school is like for you. Because your answers will help influence future educational policies in <country and/or education system>, we ask you to do the very best you can.*

However, many students view PISA as a low-stakes assessment: they can refuse to participate in the test with no negative consequences and do not receive any feedback on their performance. There is a risk, therefore, that students do not do their best on the test (Wise and DeMars, 2010[1]).

Several studies in the United States have found that student performance on assessments, such as the United States national assessment of educational progress (NAEP), depends on how they are administered. One study shows that students did not perform as well in regular low-stakes conditions as when students received financial rewards tied to their performance or were told their results would count towards their grades (Wise and DeMars, 2005[2]). In contrast, a study in Germany found no difference in effort or performance measures between students who sat a PISA-based mathematics test under the standard PISA test-administration conditions and students who sat the test in alternative high-stakes conditions tied to performance (Baumert and Demmrich, 2001[3]). In the latter study, experimental conditions included promising feedback on performance, providing monetary incentives contingent on performance, and letting students know that the test would count towards their grades. The difference in conclusions reached by these two studies suggests that students' motivation on low-stakes tests such as PISA differs significantly across countries. The only existing multi-country study on the effect of incentives on test performance found that offering students monetary incentives to do well on a test such as PISA – something that is not possible within regular PISA procedures – led to improved performance among students in the United States while students in Shanghai (China) performed equally well with or without incentives (Gneezy et al., 2017[4]).

Differences in student engagement in a given test often reveal important variations in test-administration conditions. For example, in 2018, students predominantly concentrated in a small number of schools in a few regions of Spain exhibited anomalous response patterns, performed below expectations, and reported low levels of engagement with the test. Further investigation revealed that the regions in which these schools were located had conducted their high-stakes exams for 10th-grade students earlier in the year than in the past. This meant that the testing period for these exams coincided with the end of the PISA testing window. Students were more negatively disposed towards PISA in schools where the PISA testing day was closer to that of high-stakes exams (OECD, 2020[5]).

Summing up, differences in countries' and economies' mean scores in PISA, and comparisons between PISA 2022 results and results from prior assessments may reflect differences not only in what students know and can do but how motivated they were to do their best. Put differently, PISA does not measure students' maximum potential but what students actually do in situations where their individual performance is monitored only as part of their group's performance.

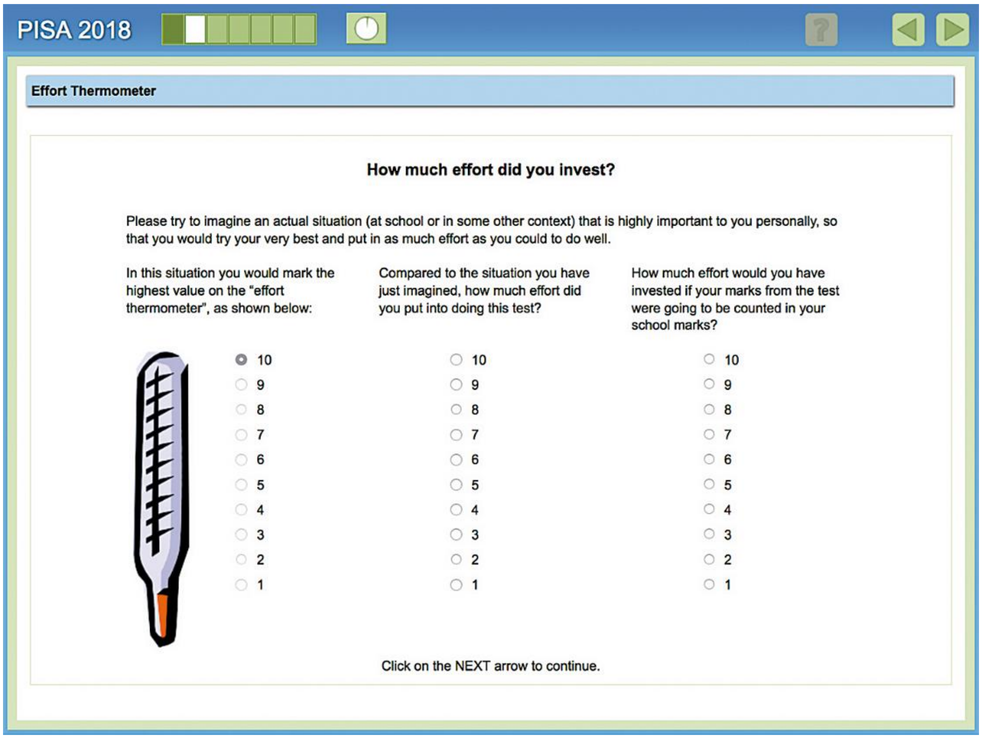
This annex computes several indicators of student engagement using PISA 2022 data to compare between countries/economies and corresponding indicators computed on 2018 data. The intention is not to suggest adjustments to PISA mean scores or performance distributions but to provide richer context for interpreting cross-country differences and trends in performance.

A number of approaches have been developed to assess differences in students' motivation in low-stakes tests (Buchholz, Cignetti and Piacentini, 2022[6]) between individuals or groups (e.g. across countries and economies). These are approaches based on self-reports (which rely on test-takers' own perceptions and reports about their effort and dispositions) and those based on behavioural indicators (which rely on observation of students' behaviour during the test). Among the latter, one can further distinguish between invasive approaches, which require dedicated resources such as human proctors, eye-tracking devices or the administration of bespoke test modules, and non-invasive approaches, which rely only on students' interactions with the test and questionnaire forms. This annex relies on self-reports and non-invasive behavioural indicators.

Self-reported effort

In PISA 2022, students were asked about the effort they invested in the test, and the effort they would have expended in a hypothetical situation if the test results counted towards their grades (see Figure I.A8.1). The same questions were also included in PISA 2018 (Figure I.A8.1).

Figure I.A8.1. The effort thermometer in PISA 2018



Effort Thermometer

How much effort did you invest?

Please try to imagine an actual situation (at school or in some other context) that is highly important to you personally, so that you would try your very best and put in as much effort as you could to do well.

In this situation you would mark the highest value on the "effort thermometer", as shown below:	Compared to the situation you have just imagined, how much effort did you put into doing this test?	How much effort would you have invested if your marks from the test were going to be counted in your school marks?
● 10	○ 10	○ 10
○ 9	○ 9	○ 9
○ 8	○ 8	○ 8
○ 7	○ 7	○ 7
○ 6	○ 6	○ 6
○ 5	○ 5	○ 5
○ 4	○ 4	○ 4
○ 3	○ 3	○ 3
○ 2	○ 2	○ 2
○ 1	○ 1	○ 1

Click on the NEXT arrow to continue.

It is paradoxical to expect that students who are disengaged and may not even read the instructions in test items would put time and effort into *this* question. Nevertheless, the self-report measure is not only widely used by scholars in this field (Wise and DeMars, 2005[2]; Eklöf, 2007[7]), it has also contributed to making PISA results more reliable. Indicators derived from student self-reporting their engagement level (OECD, 2020[5]) identified anomalies affecting Spain's data in 2018.

Self-reported effort in 2022

In 2022, more than two-thirds of students across OECD countries (71%) reported expending less effort on the PISA test than they would have done in a test that counted towards their grades (Table I.A8.1). On the 1-to-10 scale shown in Figure I.A8.1, students reported an effort of between "7" and "8" for the PISA test they just had completed, on average. They reported that they would have described their effort as "9" had the test counted towards their marks.

Students in the Dominican Republic and Uzbekistan rated their effort highest on average across all participating countries/economies. At least 75% of students completed the effort thermometer, with an average rating close to "9". Only 26% of students in Uzbekistan and 30% of students in the Philippines reported they would have invested more effort had the test counted towards their marks. At the other extreme, more than four out of five PISA students (80%) in Denmark* and Sweden (in descending order), and 71% on average across OECD countries reported they would have invested more effort if their performance on the test had counted towards their marks (Table I.A8.1).

In most countries as well as on average, boys reported investing slightly less effort in the PISA test than girls did. The effort boys reported they would have invested in the test had it counted towards their marks was also less than girls did. When the difference between the ”true” and "hypothetical" PISA effort is considered, girls are more likely than boys to report they would have worked harder on the test if it had counted towards their marks (Table I.A8.4).

Changes in self-reported effort between 2018 and 2022

Comparisons of self-reported effort across countries reflect not only actual differences in effort levels but individual and cultural differences in the use of the 1-10 rating scale as well. These differences are less likely to affect comparisons of self-reported effort across different cohorts within the same country/economy.

Students reported making less effort on the test in 2022 than in 2018 in most countries/economies: the difference corresponds to -0.2 points on the 10-point scale on average across OECD countries (Table I.A8.3). Reports about the effort students would have made had the test counted towards their grades were also lower (by 0.1 points on average across OECD countries) but the decline was more marked for reports about the actual effort students made. The proportion of students who rated their actual effort on the PISA test lower than if it had counted towards their grades increased, with only limited exceptions. Among countries where at least 75% of students completed the effort thermometer in both years, the largest increases in this proportion were in Israel (+11 percentage points), Türkiye (+10 percentage points) and Hungary (+8 percentage points). Saudi Arabia, in contrast, stands out for the opposite trend: students' self-reported effort increased by 0.3 points on the 10-point scale and the proportion reporting that their effort would have been higher if the test had counted towards their grades decreased by 12 percentage points between 2018 and 2022. It is noteworthy that there was significant improvement in mathematics performance in Saudi Arabia and students completed the test on computers in 2022 but with paper and pencil in 2018.

Sharp declines in the effort reported by students on the PISA test were observed in two of the countries with strong declines in mathematics performance: Albania (-0.6 points) and Jordan (-0.5 points) (Table I.A8.3). In both cases, the effort students would have made if the test had counted towards their grades was also significantly lower than in 2018. This suggests that lower proficiency in PISA was not just the consequence of students' lower engagement with the PISA test but with learning and school in general. In these two countries, fewer than 75% of students responded to the effort thermometer in either 2022 or 2018. The simple comparisons reported here may be affected by a lack of representativity in the sample of respondents. It is remarkable, however, that there is a strong association between the difference in effort students would have made on a regular school test and the difference in mean performance observed in PISA (Table I.A8.3 and Tables I.B1.5.4, I.B1.5.5 and I.B1.5.6) between 2018 and 2022 across all countries/economies.[1]

Behavioural indicators

There are several disadvantages to self-report measures. It is unclear whether students – especially those who may not have taken the test seriously – respond truthfully when asked how hard they tried on a test they had just taken.

And, it is unclear to what extent answers provided on subjective response scales can be compared across students, let alone across countries. The comparison between the “actual” and the “hypothetical” effort is also problematic. In the German study discussed earlier in this Annex, regardless of the conditions under which they took the test, students said that they would have invested more effort if any of the other three conditions applied; the average difference was particularly marked among boys (Baumert and Demmrich, 2001[3]). One explanation for this finding is that students are under-reporting their true effort and over-reporting their counter-factual effort, regardless of the hypothetical context of the latter: in doing so, students can attribute poor test performance to lack of effort rather than lack of ability.

In response to these criticisms, researchers have developed ways of examining test-taking effort by observing students’ behaviour during the test and questionnaire. Two sets of indicators are discussed in this section:

- indicators of endurance based on comparisons of performance on similar (or identical) tasks at different moments in the test (in particular, towards the beginning and the end of the test);
- straight-lining indicators based on the presence (or absence) of logically inconsistent responses among questions presented in close sequence;

Both types of measures are based on the idea that when respondents are disengaged they fall back on satisficing behaviour whereby they do not provide a response that reflects their best judgement or knowledge to the questions asked in the test and questionnaire. Each measure is sensitive to distinct types of satisficing behaviour and has different strengths and weaknesses.

Measures of “endurance” are sensitive to a large range of satisficing behaviours (including random or strategic guessing, skipping questions, and engaging in off-task exploration) but can be used only in cognitive tests (where the “correct” response is known by the examiner). Their interpretation as measures of engagement supposes that engagement is optimal for all students at the beginning of the test. The possibility of measuring endurance in this way also depends critically on test design.

Straight-lining indicators can be computed both for tests and questionnaires, and exploit the presence of pairs of antonyms among the items presented to the student. Antonyms are items where knowledge of a student’s answer on one item implies, logically (for semantic or psychometric properties), an opposite answer to the other item in the pair. For example, PISA questionnaire items that measure students’ sense of belonging at school ask students to what extent they agree with a number of statements, including “I make friends easily at school” and “I feel lonely at school”. Straight-lining behaviour is the use of the same response category (e.g. “strongly agree”) for all statements in a set that includes antonyms.

Endurance or the ability to sustain performance

Borgonovi and Biecek (2016[8]) developed a country-level measure of “academic endurance” based on comparisons of performance in the first and third quarter of the PISA 2012 test (the rotating booklets design used in PISA 2012 ensured that test content was perfectly balanced across the first and third quarters at aggregate levels). The reasoning behind this measure is that while effort can vary during the test, what students know and can do remains constant: any difference in performance is therefore due to differences in the amount of effort invested.[2]

The original indicator proposed for PISA 2012 can be adapted to the design used in 2022 in two ways.

A first set of indicators compares the performance of students who were administered a given test (e.g. mathematics) in the first hour to the performance of students who were administered the same test in the second hour of testing. The indicators used can be based on item-response theory (plausible values) or classical test theory (percent-correct scores) although comparisons based on the latter are only valid for students (or domains) whose tests are not adaptive and thus, under all circumstances, of identical difficulty.

A second indicator exploits the test design for mathematics in 2022, which partitions the item pool in three (mutually exclusive) sets, whose position is rotated across students. This means that items in set A were presented for one-third of students at the beginning of the mathematics test; one-third in the middle; and the remaining third at the end of the mathematics test; and similarly for sets B and C. By comparing the performance of students whose test was not adaptive (25% of all students who took the mathematics test) across different these three positions (beginning, middle, and end), it is possible to see how performance varies (and, typically, declines) over the course of the hour-long mathematics test in each country/economy.

Student performance by hour of testing

The comparison of students' performance by hour of testing shows large declines between the first and the second hour of testing in several countries and economies, in particular for reading results.

- In reading, on average across OECD countries, students who took the test in the second hour (in most cases, after completing an hour-long mathematics test) scored 14 points lower than students who took the test in the first hour – a large difference. Large performance declines during the test of between 20 and 30 score points were observed in Iceland, Israel, Latvia*, Albania, Qatar, Slovenia, Malta, Argentina and Norway (in descending order of the size of this difference) (Table I.A8.17).
- In mathematics, on average across OECD countries, the performance difference between students who took mathematics in the second hour and those who took mathematics in the first hour is only of four points. In most countries, the difference is not statistically significant; however, in Albania and Norway the decline exceeds 10 score points (Table I.A8.14).
- In science, results are between those reported above for mathematics and reading. The average decline between the first and second hour of testing is of eight points. In science, where the test was not adaptive, results based on plausible values closely match those based on percent-correct scores (the linear correlation coefficient between the two sets of estimates, a measure of their association which varies between -1 and 1, is equal to 0.95) (Table I.A8.11 and Table I.A8.20).

Overall, performance declines between the first and second hour of testing for the same country/economy across different subjects correlate only moderately. This suggests that these declines reflect both position effects (the effect of taking a test in the second hour, which is present in all subjects) and order effects (the effect of taking a reading test after a mathematics test, for example). Order effects might play out differently across subjects and depending on the country (Tables I.A8.14, I.A8.17 and I.A8.20).

Nevertheless, a few countries/economies figure consistently among those with low "endurance", meaning their second-hour results are much lower than their first-hour results regardless of the subjects. Countries/economies with low endurance in 2022 include Albania, Malta and Norway (Tables I.A8.14, I.A8.17 and I.A8.20).

The difference between the first and second hour of testing may appear large. However, similarly large declines had already been found in 2018 in most countries. In fact, on average across OECD countries, the difference between the first and second hour of testing even reduced somewhat, meaning that performance in 2022 was lower than in 2018 throughout the test but more so at the beginning of the test. The most significant exceptions to this pattern are Albania in reading, and the Dominican Republic and Greece in science, where the performance difference between the first and second hour of testing widened between 2018 and 2022 (Tables I.A8.16, I.A8.19 and I.A8.22).

Performance decline within the hour-long mathematics test

Performance declines for a given student in the hour-long mathematics test are often larger than those between students who take the mathematics test in the first and second hour of testing because students tend to perform better at the beginning of the second hour of testing (and after a break) than at the end of the first hour of testing.

On average across OECD countries, students who were assigned to a non-adaptive test in mathematics answered 47.6% of the questions correctly if they took the test in the first hour and 46.0% if they took the same test in the second hour of testing (Table I.A8.7). At the very beginning of the mathematics test, the percent-correct rate (averaged across first- and second-hour students) was 48.1% but dropped to 47.3% in the middle section, then to 44.2% in the last section – a drop of almost four percentage points (Table I.A8.23).

The largest drop in the mathematics test was observed in Israel: percent-correct rates started at levels close to the OECD average in 2022 but dropped by about seven percentage points in the third (and last) section. In contrast, performance remained at levels close to the OECD average throughout the test in France, for example. Among high-performing countries and economies, Hong Kong (China)*, Korea, Singapore and Chinese Taipei stand out for small differences (two percentage points or less) in performance between the beginning and the end of the testing hour (Table I.A8.23).

These performance declines between the first and third section of the test can modify country rankings at the margin (for example, Israel would be ranked higher if only performance at the beginning of the mathematics test was considered) but do not affect the main conclusions that can be drawn from comparisons of PISA results across countries. Around the OECD average, a 10-point difference on the PISA mathematics scale approximately corresponds to a difference of four points in the percent-correct metric.[3]

Straight-lining

Straight-lining is the tendency to use an identical response category for all items in a set (Herzog and Bachman, 1981[9]). Measures of straight-lining indicate low effort.

Patterned responses to reading-fluency tasks

The reading-fluency section introduced in the PISA 2018 test offers an opportunity to examine straight-lining behaviour in the test. Students were given a series of 21 or 22 items in rapid sequence with identical response formats ("yes" or "no"). Meaningless sentences (such as "The window sang the song loudly"), calling for a "no" answer, were interspersed among sentences that had meaning (such as "The red car has a flat tyre"), calling for a "yes" answer. It is possible that some students did not read the instructions carefully or that they genuinely considered that the meaningless sentences (which had no grammatical or syntactical flaws) had meaning. However, this response pattern (a series of 21 or 22 "yes" answers) or its opposite (a series of 21 or 22 "no" answers) is unexpected among students who demonstrated medium or strong reading competence in the main part of the reading test.

Table I.A8.25 shows that only 1.2% of all students on average across OECD countries exhibited such patterned responses in reading-fluency tasks. The proportion of patterned responses follows, in general, the proportion of students who scored below Level 2 in reading (the linear correlation coefficient between the two proportions is 0.66). However, in Korea and Türkiye, in spite of a proportion of low-performing students close to, or even below the OECD average (29% and 14%, respectively), the proportion of patterned responses in the reading-fluency test far exceeded the average proportion (5.3% and 3.5%). It is possible that the unusual response format of reading-fluency tasks triggered disengaged response behaviour and that these same students did their best in the latter parts of the test. It is also possible, however, that these students did not do their best throughout the PISA test – not only in this initial, three-minute section of the reading test.

While the content of the reading-fluency section was identical in PISA 2018 and PISA 2022, a minor change in the response format was introduced in PISA 2022: every few sentences, the position of the "yes" and "no" buttons would change slightly. This forced respondents to pay a minimum of attention in order to move forward. Comparisons between 2018 and 2022 must take this into account. Indeed, on average across OECD countries, these comparisons show a slight reduction in the proportion of patterned responses – from 1.4% to 1.2% (Table I.A8.27). It decreased even more (by 3.1 percentage points, from 3.6% to 0.5%) in Spain, where test-administration issues in 2018 limited the extent to which inferences could be drawn from the results (see the introduction to this annex, above). In contrast, the proportion of patterned responses increased significantly in Baku (Azerbaijan), the United Arab Emirates, Hong Kong (China)* and Finland (in descending order of the percentage-point increase).

Identical responses across sense-of-belonging items in the background questionnaire

The PISA questionnaire items that measure students' sense of belonging can be used to examine effort in the questionnaire and how it changed between 2018 and 2022.[4]

In most countries and economies, fewer than 5% of all students gave identical responses to all items in the sense-of-belonging set (regardless of whether the items indicated a strong sense of belonging or the opposite). Such contradictory responses were more common in Albania, Thailand, and Jordan (8%); Hong Kong (China)*, the Philippines and the United Arab Emirates (7%), the Palestinian Authority, Georgia and Qatar (6%); and in Baku (Azerbaijan) and Bulgaria (5%). These high percentages are often found in countries with large proportions of students with low reading proficiency. This suggests that some of these students did not fully understand the questionnaire items; the high percentages observed in Hong Kong (China)* stand out as anomalous in this context (Table I.A8.28).

When compared to the proportions of straight-lining students in 2018, the proportions in 2022 are, in general, lower. However, rather than reflecting increased engagement, this might reflect position or presentation effects (in 2022, every student saw, at most, five items in this set – and in all similar "matrix" questions). Among countries with large proportions of such students, this proportion increased only in Albania (Table I.A8.30).

Conclusion

Overall, the examination of various indicators of effort and motivation, and comparison with similar indicators for 2018 suggests that the conditions of administration remained similar to those observed in the past, including in terms of students' disposition towards the test. Students reported somewhat lower effort than in the past but it is unclear to what extent this phenomenon is limited to the PISA test and whether it might reflect lower engagement with learning and school more generally (in both cases, this might account for some of the negative trends observed in several countries, particularly in mathematics results).

Throughout the analysis, Albania has repeatedly been mentioned as a negative outlier: students reported spending significantly less effort on PISA and exhibited larger declines between the first and second hour of testing than in the past. There was also a larger proportion of students who used the same response category for antinomic items in the sense-of-belonging set than in 2018. These patterns suggest that the decline in performance in Albania – one of the largest ever registered in PISA – reflects, at least in part, the absence of student engagement.

Notes

[1] The linear correlation coefficient is 0.64 across all 69 countries/economies that can compare PISA 2018 and PISA 2022 results in mathematics. It is 0.55 when considering only the 57 countries/economies where at least 75% of all students completed the effort thermometer.

[2] Speed of information processing and general time management may also influence performance differences between test sections. To limit the influence of this possible confounder, Borgonovi and Biecek (2016[8]) do not use the last quarter of the test but the third (second-to-last) quarter. In the computer-based PISA 2018 and PISA 2022 assessments, the test is divided into two halves, each conducted in an hour-long session. With this design, students' time management and speed of information processing can be expected to have the same impact on both halves.

[3] This "rule of thumb" is based on the comparison of the average percentages of correct responses reported in Table I.A8.7 with the mean scores (in PISA points) reported in Table I.A8.14.

[4] The battery of items comprises six items in total; however, in 2022, only a random subset of five of these were presented to each student in countries that administered PISA on computers. Because the main focus of this analysis is on comparisons across countries and over time, questionnaire straight-lining is defined here as "providing the

same answer to at least five of the sense-of-belonging items, including at least two items loading positively and two loading negatively (i.e. indicating a lack of sense of belonging) on the scale".

References

Baumert, J. and A. Demmrich (2001), "Test motivation in the assessment of student skills: The effects of incentives on motivation and performance", *European Journal of Psychology of Education*, Vol. 16/3, pp. 441-462, https://doi.org/10.1007/bf03173192. [3]

Borgonovi, F. and P. Biecek (2016), "An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test", *Learning and Individual Differences*, Vol. 49, pp. 128-137, https://doi.org/10.1016/j.lindif.2016.06.001. [8]

Buchholz, J., M. Cignetti and M. Piacentini (2022), "Developing measures of engagement in PISA", *OECD Education Working Papers*, No. 279, OECD Publishing, Paris, https://doi.org/10.1787/2d9a73ca-en. [6]

Eklöf, H. (2007), "Test-Taking Motivation and Mathematics Performance in TIMSS 2003", *International Journal of Testing*, Vol. 7/3, pp. 311-326, https://doi.org/10.1080/15305050701438074. [7]

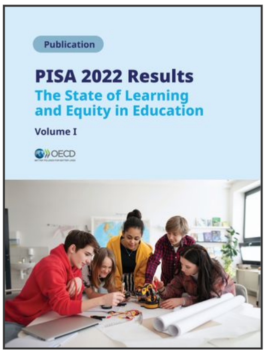
Gneezy, U. et al. (2017), *Measuring Success in Education: The Role of Effort on the Test Itself*, National Bureau of Economic Research , Cambridge, MA, https://doi.org/10.3386/w24004. [4]

Herzog, A. and J. Bachman (1981), "Effects of questionnaire length on response quality", *Public Opinion Quarterly*, Vol. 45, pp. 549–559. [9]

OECD (2020), *Annex A9. A note about Spain in PISA 2018: Further analysis of Spain's data by testing date (updated on 23 July 2020)*, https://www.oecd.org/pisa/PISA2018-AnnexA9-Spain.pdf. [5]

Wise, S. and C. DeMars (2010), "Examinee Noneffort and the Validity of Program Assessment Results", *Educational Assessment*, Vol. 15/1, pp. 27-41, https://doi.org/10.1080/10627191003673216. [1]

Wise, S. and C. DeMars (2005), "Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions", *Educational Assessment*, Vol. 10/1, pp. 1-17, https://doi.org/10.1207/s15326977ea1001_1. [2]



From:

PISA 2022 Results (Volume I)

The State of Learning and Equity in Education

Access the complete publication at:

https://doi.org/10.1787/53f23881-en

Please cite this chapter as:

OECD (2023), “How much effort do students put into the PISA test?”, in *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/cf9910a0-en

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Extracts from publications may be subject to additional disclaimers, which are set out in the complete version of the publication, available at the link provided.