CEPEO Working Paper No. 25-14

READ THE DAMN DOCUMENTATION (CAREFULLY). A case study using the PISA data.

John Jerrim
UCL

Maria Palma Caravajal
UCL

Jake Anders
UCL

María Ladrón de Guevara Rodríguez
UCL

Oscar Marcenaro-Gutierrez
Málaga

When you get access to a new dataset, do you always carefully read the documentation first? We all know we should. But – let's be honest – it's a lot more fun to just start playing with the data. This can however be a dangerous game to play. This paper presents a case study of this matter using the OECD's Programme for International Student Assessment (PISA). A survey question included in this study attempts to measure student truancy across countries over time. The international survey documentation suggests an identical question has been used across countries and cycles. Yet the national documentation illustrates how a subtle – yet important – change to the wording was made in some countries in 2015. We demonstrate how researchers could easily miss this change and how this would impact inferences in changes in truancy rates before and after the COVID-19 pandemic. Attempts to use artificial intelligence and large language models to spot this problem resulted in overconfidently incorrect advice. The findings thus serve as a reminder to even the most experienced data analysts (including ourselves) – ALWAYS READ THE SURVEY DOCUMENTATION CAREFULLY.

VERSION: November 2025

Suggested citation: Jerrim, J., Palma, M., Anders, J., Ladrón, M. & Marcenaro-Gutierrez, O. (2025). *READ THE DAMN DOCUMENTATION (CAREFULLY). A case study using the PISA data.* (CEPEO Working Paper No. 25-14). UCL Centre for Education Policy and Equalising Opportunities. https://EconPapers.repec.org/RePEc:ucl:cepeow:25-14.

Centre for Education Policy & Equalising Opportunities

Disclaimer

Any opinions expressed here are those of the author(s) and not those of UCL. Research published in this series may include views on policy, but the university itself takes no institutional policy positions.

CEPEO Working Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Highlights

- This paper presents a case study of how small changes in data definitions can have substantial consequences for the conclusions drawn from analyses built upon them.

- We present evidence on the effect of changing a single word in a survey item in a major international survey in some countries, but not others, and the implications this has on the interpretation of the resulting statistic.

- With the rapidly growing use of artificial intelligence (AI) and large language models (LLM), we investigate whether such tools would be able to help researchers spot such problems. This includes how specific prompts to these LLMs would need to be to do so.

- We hope these findings provide a cautionary tale for novice survey methodology and statistics practitioners, helping the next generation of researchers understand the importance of carefully studying the data documentation – and the traps that await them once they are let loose in the wild. Although, realistically, we think these lessons are relevant for experts, too!

Why does this matter?

Failing to spot even subtle changes in data definitions can result in serious errors in the conclusions drawn from analysis.

READ THE DAMN DOCUMENTATION (CAREFULLY). A case study using the PISA data.

John Jerrim[1] (UCL Contact author)

Maria Palma Carvajal (UCL Social Research Institute) [2]

Jake Anders (UCL Centre for Education Policy & Equalising Opportunities) [3]

María Ladrón de Guevara Rodríguez[4]

Oscar David Marcenaro-Gutierrez[5]

When you get access to a new dataset, do you always carefully read the documentation first? We all know we should. But – let's be honest – it's a lot more fun to just start playing with the data. This can however be a dangerous game to play. This paper presents a case study of this matter using the OECD's Programme for International Student Assessment (PISA). A survey question included in this study attempts to measure student truancy across countries over time. The international survey documentation suggests an identical question has been used across countries and cycles. Yet the national documentation illustrates how a subtle – yet important – change to the wording was made in some countries in 2015. We demonstrate how researchers could easily miss this change and how this would impact inferences in changes in truancy rates before and after the COVID-19 pandemic. Attempts to use artificial intelligence and large language models to spot this problem resulted in overconfidently incorrect advice. The findings thus serve as a reminder to even the most experienced data analysts (including ourselves) – ALWAYS READ THE SURVEY DOCUMENTATION CAREFULLY.

Key words: PISA; AI; data documentation; survey methodology; truancy; absences; COVID-19.

Funding statement: Oscar David Marcenaro-Gutierrez received funding from Fundacion BBVA – Prismas y Problemas – 2023 ("La inequidad socioeconómica derivada de la ineficiencia de los sistemas educativos (INESOCEF)").

[1] Social Research Institute, University College London, 20 Bedford Way London, WC1H 0AL.

[2] Social Research Institute, University College London, 20 Bedford Way London, WC1H 0AL.

[3] UCL Centre for Education Policy & Equalising Opportunities, 20 Bedford Way London, WC1H 0AL.

[4] *Departamento de Economía Aplicada (Estadística y Econometría). Facultad de Ciencias Económicas y Empresariales. Universidad de Málaga. Plaza de El Ejido s/n, 29013, Málaga (España)*. E-mail: marialadron@uma.es. Tel.: +34 952131206. ORCID: 0000-0002-5087-422X

[5] *Departamento de Economía Aplicada (Estadística y Econometría). Facultad de Ciencias Económicas y Empresariales. Universidad de Málaga. Plaza de El Ejido s/n, 29013, Málaga (España)*. E-mail: odmarcenaro@uma.es. Tel.: +34 952137003. ORCID: 0000-0003-0939-5064

1. Introduction

When buying a new gadget, we know the first thing we should do is to read the instructions carefully. In reality, all anyone wants to do is get their new toy out of the box and have a play. Most of the time, all will be well and good. But, somethings, things can go wrong. Our haste leading us to break our shiny new toy, sinking our money down the drain.

The same can happen in data analysis too. After waiting months – if not years – to get our hands on some exiting new data, the temptation is to just dive straight in. Why bother to read pages and pages of boring survey documentation, when the exciting world of data cleaning, descriptive statistics and statistical modelling awaits? The reason - as all good, experienced data analysts know – is that we might break our new "toy" (dataset) in the process, putting significant time and effort to waste. But – let's be honest – do we always practise what we preach?

The overarching aim of this paper is to illustrate how subtle issues tucked away in the depths of survey documentation can make a substantial difference to one's results. It focuses on a change to a single word in a survey question within a single survey wave, impacting a subset of participants in a major international study (the OECD's Programme for International Student Assessment - PISA). The documentation to such studies is – to the survey organisers credit - extensive. It consists of lengthy international technical reports (over 500 pages), multiple different questionnaires, codebooks and much more. Each of these are freely available to download across several survey cycles, with some in different languages. Given this complexity and level of detail – along with the time pressure researchers work under – important nuances can be easily missed.

The specific issue we focus on is changes in school truancy rates - i.e. pupils intentionally skipping school – over time, with particular interest in how this has risen following the COVID-19 pandemic. This is an important policy issue facing several countries (Andres, 2024; Mokhtarian, 2024; Nathwani et al., 2021), with our recently published companion paper illustrating how – since the pandemic – this has become an acute challenge amongst girls within English-speaking countries (Anders et al., 2024). Within this paper, we provide a case study demonstrating how easily this important finding could be missed.

In doing so, we attempt to make three contributions to the existing evidence base. First, part of the survey methodology literature has studied how nuances in the wording of survey items can change the meaning of the construct being measured (Kalton and Schuman, 1982), and how this can impact the inferences made (Bulut and Bulut, 2022; DiStefano and Motl, 2006; Zeng et al., 2020). We add to this literature by presenting new evidence on the effect of changing a single

word in a survey item to our understanding of school truancy rates. Second, with the rapidly growing use of artificial intelligence (AI) and large language models (LLM), we investigate whether such tools would be able to help researchers spot such problems. This includes how specific prompts to these LLM need to be to do so. Finally, our hope is that the paper will become a key resource used in introductory survey methodology and statistics courses, helping the next generation of researchers understand the importance of carefully studying the data documentation – and the traps that await them once they are let loose in the wild.

2. Data and methodology

Data

The data we use are drawn from the 2012, 2015, 2018 and 2022 rounds of PISA. This is a high-profile study of 15-year-olds skills across the globe. Around 80 countries now participate in PISA, though our analysis focuses on data from OECD countries with English as the official language[6] (though, in Appendix G, we replicate parts of our analysis for all OECD nations with data available). PISA uses a two-stage sample design, with schools selected as the primary sampling unit, with around 30 to 40 pupils then randomly selected within each. Response rates to the survey are generally high (≈80-90% of randomly selected schools and pupils participate) though with some important variation across countries. The international database includes a set of student and balance-repeated-replication weights which adjust estimates to account for all aspects of this complex sample design (including the clustering of pupils within schools). These weights are applied throughout our analysis via the STATA svy command (OECD, 2022).

As part of PISA, pupils respond to a background questionnaire. According to the international survey documentation this includes the following question:

"In the last two full weeks of school, how often: I skipped a whole school day"

With four possible answers: Never, one or two times, three or four times, and five or more times.

Appendix F provides the precise wording included in the international documentation of the questionnaire. This indicates that the *exact* same wording of this question was used in PISA 2015, 2018 and 2022, with a slightly different format in 2012 (see Appendix F for further details).

Figure 1 illustrates the percent of pupils indicating they had skipped school at any point over the last two weeks (i.e. the percent selecting one of the top three categories) in the United States and

[6] Nine OECD countries with English as official language participate in PISA: Australia, Canada, England, Northern Ireland, Wales, Scotland, Ireland, New Zealand and USA.

Australia between 2012 and 2022[7]. The results are mixed. In Australia there is clear evidence of a sharp increase in the PISA 2022 cycle (post-pandemic) in comparison to PISA 2012-2018 (pre-pandemic). But, in the United States, there is zig-zag pattern, with lower levels in 2012 and 2018 but higher levels in 2015 and 2022. Taken at face value, the increase in truancy experienced in the US in the 2022 cycle (following the COVID-19 pandemic) does not stand out as particularly unusual. Indeed, it looks like an even bigger change occurred between 2012 and 2015. Moreover, unlike Australia, there is little clear evidence of an upward trend.

Figure 1: Proportion of students reporting skipping/missing school at least once in the past two weeks for USA (blue) and AUS (red).

All, however, is not as it first seems. Within international surveys, countries are allowed to make adaptions to the questions, usually to aid interpretation in their setting and cross-national comparability. Looking at the international questionnaires (see Appendix F) one would not initially think this to be an issue here. The term "skipped" appears – in the international questionnaires – consistently across the 2012 to 2022 waves. The same holds true in the labelling of the variables within the downloadable datasets from the OECD website[8]. Moreover, given our focus on English-

[7] Appendix Figure A1 and Table A1 show proportions for all countries in the analysis.
[8] The label of this variable is "Truancy- Skip whole school day" for in PISA 2012, "In the last 2 weeks, how often have you <skipped> a whole school day." In PISA 2015, 2018 and 2022.

speaking countries, it is not clear why any national or linguistic adaption would be necessary – particularly one that changes across survey cycles.

Yet, bizarrely, the wording of the question seems to have changed in a subset of PISA countries over time. Take the United Kingdom, for example. In PISA, Scotland takes part on its own as a separate sub-national entity, distinct from England, Northern Ireland and Wales. This means it oversees its own national adaptions to its questionnaires, separate from the rest of the UK. To find out the precise wording of the questions pupils *actually* answered, one must download the national versions of the questionnaires from the PISA website. These are provided in Appendix E for reference.

It is only by closely studying these additional documents that a key issue is revealed. In Scotland, the question was worded consistently between 2012 and 2022 (always using the term "skipped"). But – in the rest of the UK – the 2015 edition changed the wording to "In the last two full weeks of school, how often: I missed a whole school day?". Then – in 2018 and 2022 – the phrasing altered back to "skipped". Similar changes to the wording were made in other countries in PISA 2015, which we document for predominantly Anglophone nations in Table 1 (and for a broader array of countries in Appendix G).

Table 1: Exact word used in truancy question in each country and PISA round.

Country	PISA 2012	PISA 2015	PISA 2018
Australia	Skipped	skipped	skipped
Canada	Skipped	skipped	skipped
Ireland	Skipped	skipped	skipped
New Zealand	Skipped	skipped	Skipped
England	Skipped	missed	Skipped
Northern Ireland	Skipped	missed	Skipped
Wales	Skipped	missed	Skipped
Scotland	Skipped	skipped	Skipped
United States	Skipped	missed	Skipped

Note: The complete question is "In the last two full weeks of school, how often: I <skipped>/<missed> a whole school day". Information gathered from each country's questionnaire.

This change of wording is very easy to miss. There is no hint of any such alteration in the international versions of the questionnaire, or in the more than 500-page PISA technical reports (OECD, 2022). We can also find no evidence of it ever being reported by the organisation that conducts PISA – the OECD. In-fact, when the OECD released results from PISA 2015, they included the following statement in the country note for England "*Between 2012 and 2015, the percentage of students in the United Kingdom who had skipped a day of school in the two weeks*

prior to the PISA test increased by eight percentage points (the OECD average is an increase of five percentage points), signalling that students' engagement with school has deteriorated during the period" (OECD, 2016, p. 7). This illustrates how even the organisation running the PISA study did not spot this problem – potentially leading to an erroneous conclusion.

This important point is also likely to have led to challenges in prior academic studies using PISA to study trends in school truancy. For instance, Fredriksson et al. (2023) used PISA to compare school truancy across four countries over time – Germany, Japan, Sweden and the UK. The first three of these countries used the word "skipped" consistently across cycles while – as noted above – it changed to "missed" in 2015 across most of the UK. The headline conclusion reported in the abstract - "*The UK is the only country where the changes between 2012 and 2015 as well as between 2015 and 2018 were significant*" (Fredriksson, 2023, p2) – may therefore be brought into question. The same authors have since conducted a similar follow-up study (Fredriksson et al., 2024) where they also fail to note the important changes made to the wording of the question in the UK.

It is this change of wording that is the focus of our analysis. In particular, we are interested in the impact this wording change has had on the estimated percent of pupils playing truant from school.

Methodology

Our analysis begins by presenting a selection of descriptive plots illustrating the trend in truancy rates across Anglophone countries over time, with and without data from 2015 in instances where the question wording changed. This is then followed by a series of linear probability models of the form:

$$SKIP_{ij} = \alpha + \beta.WAVE_{ij} + \delta.D_{ij} + \varepsilon_{ij} \ \nabla K \qquad (1)$$

Where:

$SKIP_{ij}$ = A binary variable coded zero if the young person indicated they are not skipped/missed school at any point over the last two weeks and coded one if they indicated once or more.

$WAVE_{ij}$ = A variable capturing the PISA survey wave, entered as a binary term that compares 2022 to previous years (2012-2015-2018).

D_{ij} = A vector of variables for pupil's demographic characteristics, including gender, socioeconomic status and immigrant status.

ε_{ij} = Random error term. The clustering of pupils within schools is accounted for via the application of the BRR replication weights.

∇K = Indicates that the model is estimated separately for each of the K countries.

i = Pupil i.

j = School j.

These models will be estimated on the pooled PISA 2012, 2015, 2018 and 2022 samples, with separate estimates produced for each of the K countries. The estimated β parameter from these models capture the annual increase in the truancy rate in the country over the 2012 to 2022 period. For those countries impacted by the 2015 wording change, we estimate this model with and without the 2015 data included. The difference between these two estimates will then illustrate how the wording change in these countries impacts one's inferences about trends in truancy rates over time.

We then consider heterogeneity across selected demographic groups; how were estimates of gaps in truancy rates impacted by the wording change? Here we estimate models of the form:

$$SKIP_{ij} = \alpha + \beta.WAVE_{ij} + \delta.D_{ij} + \gamma.WAVE_{ij} * D_{ij} + \varepsilon_{ij} \ \nabla K \qquad (2)$$

Where:

$WAVE_{ij}$ = A set of dummy variables for survey wave, with 2015 as the reference group.

D_{ij} = One of the demographic variables of interest (e.g. gender).

With all other variables as specified above. These models are again estimated on the pooled sample, separately by country. The parameters of interest are γ, particularly amongst those nations impacted by the change of wording. These will reveal whether differences in truancy across groups (e.g. between genders) was bigger or smaller in the 2012, 2018 and 2022 PISA waves (when the wording "skipped" was used) compared to 2015 (when the word "missed" was used) as the reference group. In particular, one might anticipate these γ parameters to be larger in countries that were impacted by the wording change than those that were not affected.

Next, we more formally test the impact of the wording change via estimation of the following model:

$$SKIP_{ij} = \alpha + \beta.WAVE_{ij} + \delta.MISSED_{ij} + \gamma.WAVE_{ij} * MISSED_{ij} + \varepsilon_{ij} \qquad (3)$$

Where:

$MISSED_{ij}$ = A dummy variable. Coded zero in instances where the pupil was asked the question about "skipping" school and coded one where the pupil was asked about "missing" school.

This model is estimated on the pooled 2012-2022 sample including all countries at the same time. The specification has clear similarities with a standard difference-in-difference model, with γ being the parameter of interest. This in essence captures the percentage point impact on the truancy rate amongst those exposed to the wording change.

To conclude we conduct a series of robustness and placebo tests. As part of the same question, pupils were also asked to respond to the statement:

"*In the last two full weeks of school, how often did the following things occur:*

- *I skipped some classes"*

Critically, the same issue with the change of wording impacted this question as well (i.e. "skipped" was changed to "missed" in 2015 in a subset of countries). One can therefore also explore the impact that the wording change had on this conceptually similar question.

As part of the same battery of questions, pupils were also asked about whether they had arrived late for school:

"*In the last two full weeks of school, how often did the following things occur:*

- *I arrived late for school"*

Critically, the national questionnaires suggest that the wording of this question remained identical between 2015 and 2022, and very similar in 2012 (see Appendix E for further details). While school truancy and tardiness are distinct issues, they are also conceptually related. Our anticipation is thus that we will observe much less change in response to this "lateness" question in 2015 than for the school truancy question in the subset of countries where the wording changed to "miss". Key parts of our analysis will be reproduced for this placebo outcome and the results compared. (Detailed results will be provided for this placebo outcomes in Appendix B and C, with a summary presented in the results section that follows).

<u>Could AI be used to spot this problem?</u>

Given increasing use of AI and LLMs we also investigate whether one of the most widely used tools could help researchers spot the problems documented above. In particular, we draw on Chat-GPT 5.0, utilising its "deep research" mode (Open AI, 2025). This is described by Open AI – the company that owns Chat GPT – as being "built for people who do intensive knowledge

work.....and need thorough, precise and reliable research", going on to state that it "marks a significant step toward our broader goal of developing AGI [Artificial General Intelligence], which we have long envisioned as capable of producing novel scientific research" [Authors additions]. As such, this should be one of the most powerful AI tools available to spot the problems we have encountered.

Our investigations proceed by giving Deep Research a set of prompts, followed by answers to various clarification questions it returned in response. The full set of prompts and responses used are provided in Appendix D. We begin with somewhat broader prompts asking for general methodological advice in answering our broad research question (how truancy rates have changed across countries over time) using PISA data, before making our prompts regarding possible problems more specific. For instance, our first prompt was phrased:

"*I am thinking of writing an academic paper. I am planning to use data from the OECD PISA study. I want to use these data to investigate changes in school truancy rates over time. Please could you highlight to me any important methodological points I should consider in my analysis, including issues of data quality. (In conducting your search and making recommendations, please ignore anything that you find that has been published by John Jerrim).*"

Note that we have included the final sentence in parenthesis as – in our sister paper discussing change in truancy before and after the COVID-19 pandemic (Anders et al., 2024) – we included the following paragraph:

"*However, our preliminary investigations of the data – and of the national adaptions made to the PISA questionnaire – have highlighted an issue that seems to have affected responses to this question in the 2015 wave in a small number of countries. Specifically, for some countries, the wording of the question differed in PISA 2015, when students were asked whether they missed a school day (rather than skipped). This materially alters the question, so that encompasses all forms of absence rather than just truancy. We have therefore excluded the PISA 2015 data for the following countries that were affected by this issue: England, Northern Ireland, Wales, United States and Finland".*

We were therefore concerned that the fact we have briefly mentioned the change of wording previously might get picked up by the AI, making our investigations somewhat circular. As it turns out, the responses the AI provided were similar whether this final sentence in our prompt was included or not (see prompt 1 and prompt 4 in Appendix D for further details).

Following these results, we then prompted the AI more directly (in a new chat) about possible issues with the truancy question in PISA over time:

"*I am thinking of writing an academic paper. I am planning to use data from the OECD PISA study. I want to use these data to investigate changes in school truancy rates over time. Please could you highlight to me any important issues regarding the comparability of these data over time and across countries, particularly with respect to the truancy question*"

An overview of our key findings from these prompts will be presented at the end of our results section.

3. Results

Table 2 begins by presenting estimates from our regression models comparing the truancy rate before (2012-2018) and after (2022) the COVID-19 pandemic. For those countries where there was a change of wording (shaded in green), we present estimates with and without the 2015 PISA wave. Figures reflect the percentage point difference in the truancy rate.

The first key point to note is that, in those countries impacted by the wording change, not spotting this problem would lead one to underestimate the increase in school truancy rates following the COVID-19 pandemic. Take the United States, for instance. When PISA 2015 is included in the analysis, one concludes that the post-pandemic increase in the truancy rate is relative small at 3.4 percentage points. This is in-fact a significantly smaller increase than in almost all the other Anglophone countries included in the analysis (the exception being Northern Ireland). A Donald Trumpian interpretation of this result would be that no country has performed significantly better than America in limiting the impact of the pandemic on the school truancy rate – the country has truly been made great again. This includes its nearest neighbour Canada, where levels of truancy appear to have increased by more than double the figure for the United States.

But – once the erroneous PISA 2015 data have been excluded - a rather different story emerges. The estimate of the post-pandemic increase in the truancy rate in the United States is now almost three times higher (8.9 percentage point increase rather than 3.4 percentage points). Moreover, the point estimate of the increase in the United States is now actually above that of Canada (8.9 versus 7.6 percentage points). The inferences one has reached have thus dramatically changed.

A similar story holds in other countries as well, though admittedly not by the same magnitude. For instance, in Northern Ireland, the estimate more than doubles once the 2015 data have been removed (2.6 to 5.5 percentage points), while in England and Wales it is inflated by 40%. It is also worth noting that, while this study focuses on English-speaking countries, some other countries

also experienced a similar change to the question wording (e.g. Finland) and thus whose data – and analogous estimates – are likely to have been impacted as well.

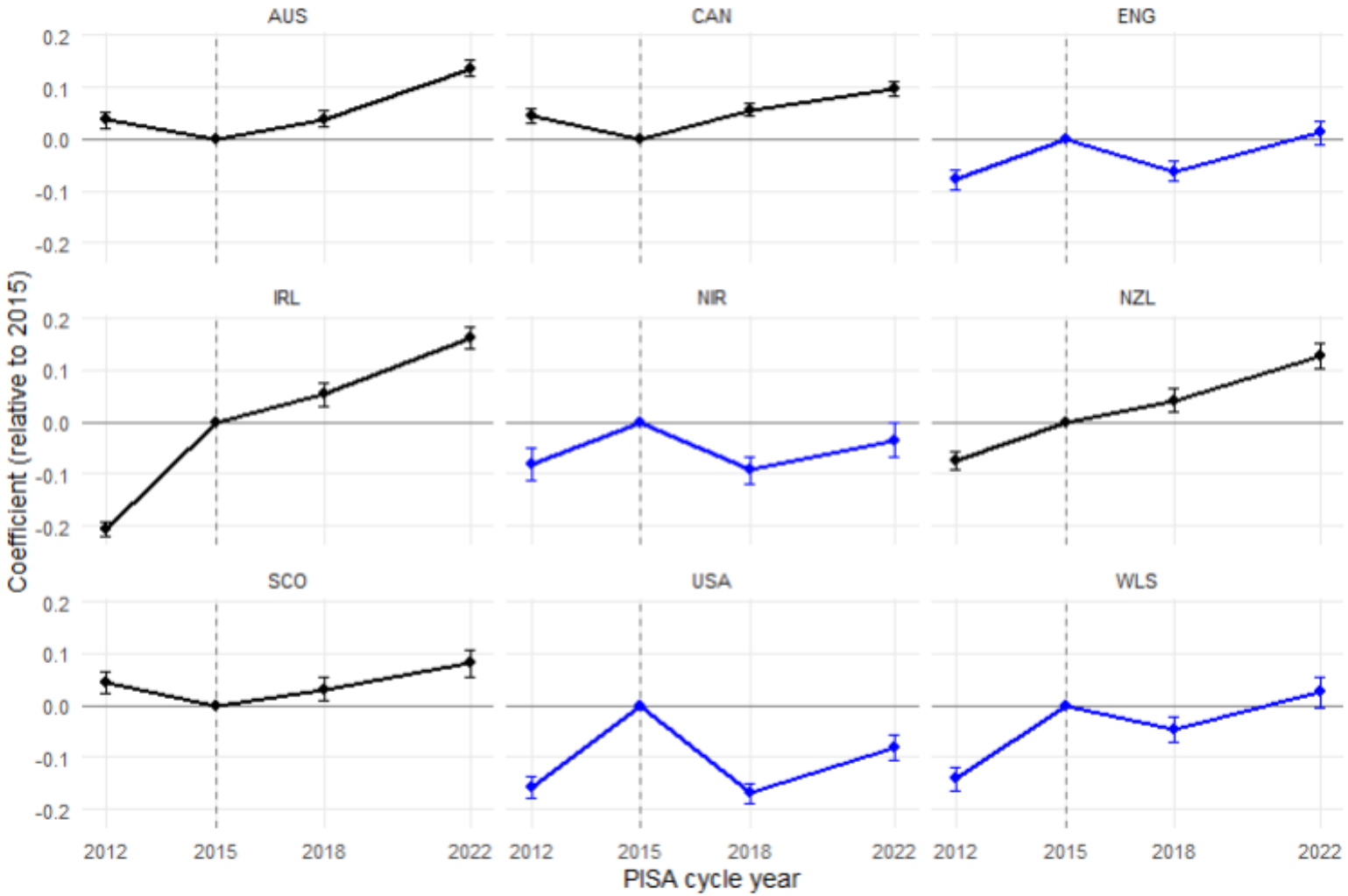
Table 2: Estimates of the association between PISA year and truancy. Pre-pandemic (PISA 2012 to 2018) compared to post-pandemic (PISA 2022).

	All years			Excluding 2015		
Country	**Coef.**	**SE**	**N**	**Coef.**	**SE**	**N**
AUS	11.8%	0.6%	49,517			
CAN	7.6%	0.7%	77,561			
ENG	6.0%	1.0%	16,713	8.4%	0.9%	11,993
IRL	21.8%	0.8%	20,139			
NIR	2.6%	1.5%	7,921	5.5%	1.6%	5,732
NZL	15.6%	1.0%	17,339			
SCO	7.3%	1.1%	11,092			
USA	3.4%	1.1%	18,707	8.9%	1.1%	13,372
WLS	9.5%	1.3%	10,392	13.0%	1.4%	7,261

Note: Year was included as a binary term equal to 0 in years 2012, 2015 and 2018, and 1 in 2022. The estimation adjusts for gender, migration status and socioeconomic status quartiles. Countries where the wording was changed from "skipped" to "miss" are highlighted in green shading.

Figure 2 then illustrates results from another regression model, where survey year is entered as a set of dummy variables, with 2015 used as the reference year. Figures refer to the percentage point change in the truancy rate, relative to this reference year. The countries in blue are those where there was a wording change. In each these countries, the truancy rate in 2022 (post-pandemic) is similar or even below the apparent level in 2015 – driven by the wording change. If taken at face value, it would be hard to claim there is much evidence of an increase over time, or a clear effect of the pandemic (given the similar values observed in 2015). For those countries without the wording change (plotted in black) the upward trajectory in truancy from school is much more apparent.

Figure 2: Estimates of the association between year and truancy for each country



Note: Year was included as a categorical term, with 2015 as the reference year. The estimation adjusts for gender, migration status, PV1 Math score quartiles and socioeconomic status quartiles. Coefficients for each country/year are shown in Appendix Table A2.

Appendix Figures A2-A5 investigate potential differences by gender, immigrant background, socio-economic status and mathematics achievement. Across countries, the year effects and their broad movements are robust to these interactions; while some subgroup contrasts are statistically significant, none overturn the central conclusion that 2015 stands out anomalously in the treated countries.

To conclude, we pool data from the nine Anglophone OECD countries together. A dummy variable is included in the model contrasting "treated" (i.e. those with a wording change – England, Wales, Northern Ireland, United States) to "untreated" countries. This treatment dummy is then interacted with dummy variables for survey year (reference = 2015), thus providing difference-in-difference style estimates to capture the effect of the wording change. Appendix Table A3 shows sample sizes across the years. These results are presented in Table 3.

The main parameter of interest is the treatment-by-year interactions. These reflect how much smaller the estimated truancy rate is when the word "skipped" is used rather than "missed". In each year, the coefficients are sizeable, negative and statistically significant. Specifically, they imply that using the word "missed" rather than "skipped" in the question increases estimates of the truancy rate by around 15 to 20 percentage points. In other words, the truancy rate in treated countries looks unusually high in all years relative to 2015, consistent with a wording-induced discontinuity rather than a true shift in truancy behaviour.

Appendix Table A4 extends this analysis to explore potential differences across subgroups. The wording change effect appears consistently across gender, immigrant background, mathematics achievement, and socioeconomic status.

Table 3: Pooled estimates for the effect of the change in wording

	Percentage point change in the truancy rate
Countries with wording change = 1	9pp***
	(1.0pp)
Year (ref=2015)	
year = 2012	1pp*
	(0.5pp)
year = 2018	4.9pp***
	(0.4pp)
year = 2022	12.7pp***
	(0.5pp)
1.WordingChange#2012.year	-16.2pp***

	(1.0pp)
1. WordingChange#2018.year	-20.7pp***
	(0.9pp)
1. WordingChange#2022.year	-19.4pp***
	(1.1pp)
Constant	37.2pp***
	(0.6pp)
Observations	229,381
R-squared	0.035

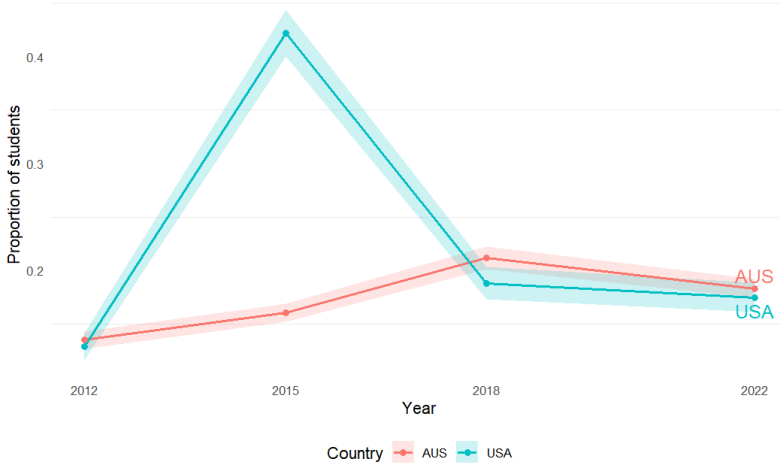
Note: Estimated using OLS. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Estimation adjusts for country fixed effects, gender, migration status and socioeconomic status quartiles.

Robustness and placebo tests

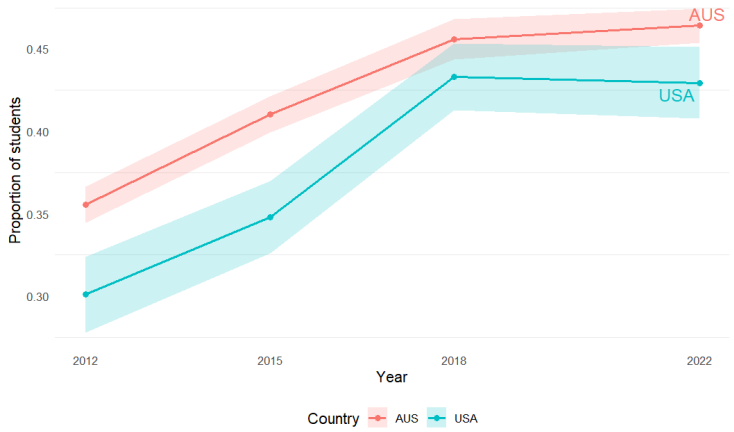
Figure 3 presents results from our robustness and placebo tests for Australia and the United States – the two countries for which we presented results for skipping school in Figure 1. Panel (a) presents equivalent results for missing classes, while panel (b) captures the percentage of pupils arriving late to school. While there was also a change of wording in the former (missing classes) in 2015, in the latter (skipping classes) the wording remained unchanged between 2012 and 2022.

Figure 3. Proportion of students reporting skipping/missing classes within a school day and arriving to school late at least once in the past two weeks for USA (blue) and AUS (red).

(a) Missing classes



(b) Late arrivals



Note: Figures refer to the percentage of students reporting they have missed/skipped classes within the school day (panel a) or had arrived late for school (panel b) at any point over the last two weeks.

It is clear in panel (a) there is a big spike in the results for the US but not for Australia. The wording change in the United States has had a major impact on the results. In contrast, in panel (b) – where the wording of the question was consistent across the two countries and over time – the lines for Australia and the United States run broadly parallel to one another throughout the period. Together, this provides further evidence of just how big the impact the wording change had on the PISA 2015 data for the United States.

Table 4 continues by presenting results across all Anglophone countries. For the countries where there was a wording change, it presents results for three closely related outcomes: skipping school (wording change 2015), skipping classes (wording change 2015), late arrival at school (consistent wording over time). Figures refer to the difference compared to 2015. Pink shading indicates where the percentage is significantly lower than in 2015, with green shading where it is significantly higher.

The first notable feature of Table 4 is that the results for skipping classes – where there was also a wording change in 2015 – are consistent with the results presented above for skipping whole school days (repeated in the left-hand column of Table 4 for ease of comparison). Indeed, the magnitude of the differences compared to 2015 are even starker. For instance, in England and the United States, the percent reporting they skipped classes in 2012/2018/2022 is around 20 to 30 percentage points lower than the percent reporting that they missed classes in 2015. This thus illustrates the robustness

of our substantive conclusion that this change of wording has had a major impact of the PISA 2015 results.

Table 4. Change compared to 2015 for alternative measures of school truancy and tardiness

	Skipped/missed school		Skipped/missed classes		Late	
	PP difference	SE	PP difference	SE	PP difference	SE
England						
2012	-8%	1%	-23%	1%	0%	1%
2015 (Ref)						
2018	-7%	1%	-20%	1%	7%	1%
2022	1%	1%	-18%	1%	10%	1%
Northern Ireland						
2012	-6%	2%	-34%	1%	1%	2%
2015 (Ref)						
2018	-9%	1%	-27%	1%	14%	2%
2022	-2%	2%	-33%	1%	15%	2%
Wales						
2012	-13%	1%	-26%	1%	0%	1%
2015 (Ref)						
2018	-6%	1%	-15%	2%	12%	1%
2022	3%	2%	-17%	2%	14%	2%
United States						
2012	-16%	1%	-30%	1%	-5%	2%
2015 (Ref)						
2018	-17%	1%	-24%	1%	8%	1%
2022	-8%	1%	-25%	1%	8%	2%

Notes: Figures refer to percentage point difference compared to 2015 as the reference group. The estimation adjusts for gender, migration status and socioeconomic status quartiles. Pink shading indicates where there has been a statistically significant decline, green shading a significant increase. In 2015 the wording changed from skipped school (left column) and skipped classes (middle column) to missed school and missed classes. The wording of the question about late to school has remained consistent between 2015 and 2022 (with minor difference in 2012).

The results in the final column of Table 4 essentially act as our placebo test. This asked pupils about late arrivals to school, with the wording remaining the same in each country over time. These results – measured for a conceptually linked outcome – are hence unaffected by the change of wording. Interestingly, the estimates here for the 2018 and 2022 waves (relative to 2015) are positive and statistically significant. They are, in other words, in the opposite direction to those for skipping/missing school that were affected

by the change of wording. There are two plausible explanations for this result. One is that the percentage of pupil's arriving late for school has genuinely increased since 2015 – potentially adding weight to the evidence that the change of wording to missed in 2015 has led to underestimation of increases in truancy and related behaviour. The other is that the change of wording in 2015 could have led to some spillover effect on the question about arriving late for school (as it is asked directly after the questions about truancy). In other words, the fact that more pupil's indicated in 2015 that they had missed school could have led them to have been less willing to indicate that they arrive late for school. While we cannot tease these potential explanations apart, our placebo test has clearly been passed, in that we do not observe the same negative effect in these countries on a question where the wording has remained consistent over time.

Truancy results for all OECD countries:

To assess whether changes in the wording of the truancy question affected responses across OECD countries, we collected the exact phrasing from each national questionnaire and translated it (see Appendix G for details on how this was done).

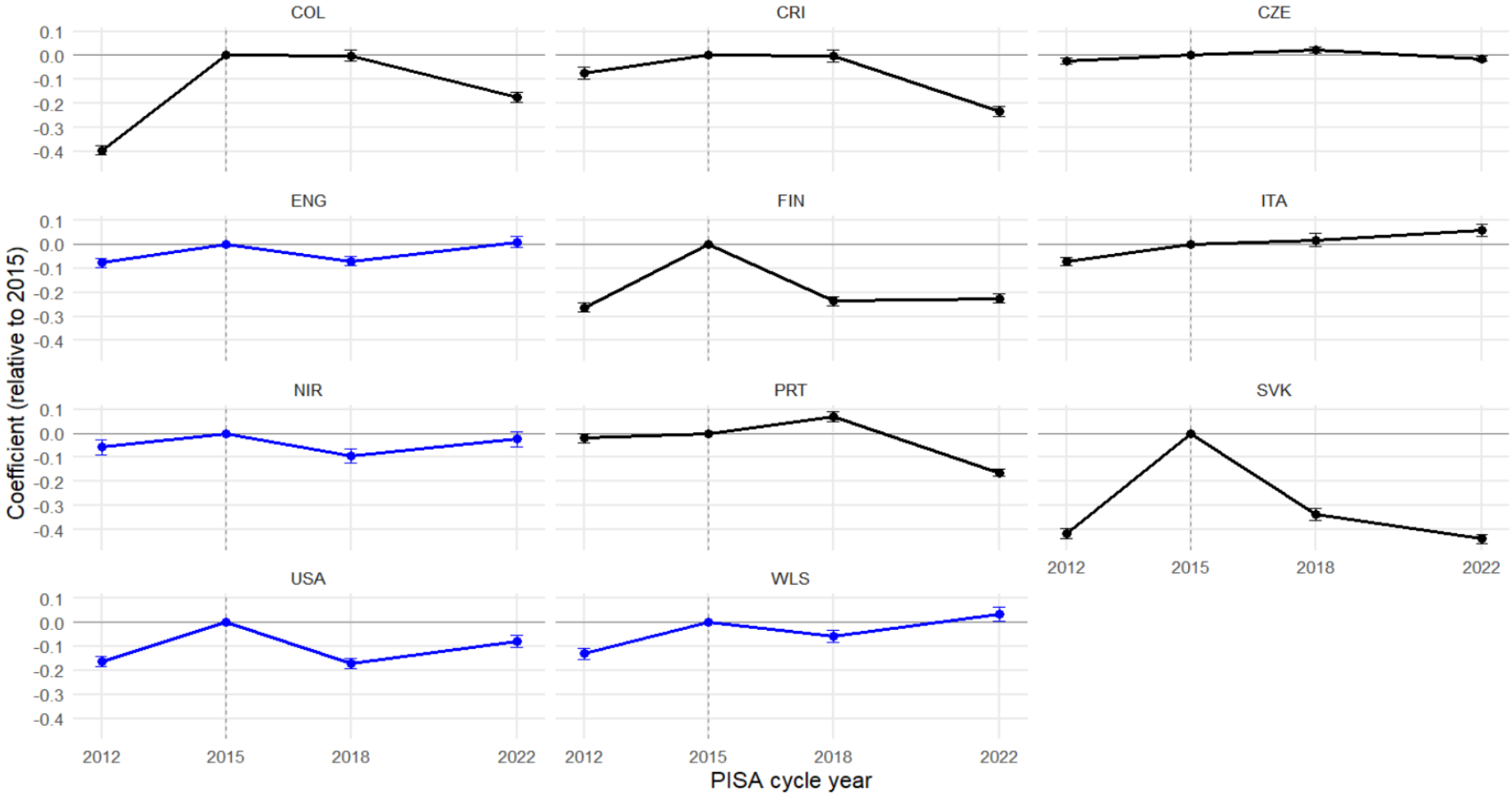
Our analysis showed that changes to the wording of the truancy question were neither exclusive to anglophone countries nor limited to the 2015 cycle (see Appendix Table G1). For example, in the Czech Republic, the question reads "I was absent from school all day" in each PISA round. In Colombia, the wording was "I missed a whole school day" in 2015 and 2018 but changed to "I missed a whole day of school without any justification" in 2022 (aligning more closely with the concept of truancy than with general absenteeism). A similar change was observed in Costa Rica. Portugal is the only country where the wording referred to "missed" only in 2018.

We also identified differences in meaning across languages within the same country. The English translations of the wording used in Estonia, Finland, and the Slovak Republic refer to truancy (using *skipped* rather than *missed*) in the Estonian, Swedish, and Hungarian versions of the questionnaires. Yet, in the Russian, Finnish, and Slovak versions, the corresponding terms translate as *missed*, with no mention of "without permission" or "without justification."

Figure 4 replicates the estimates presented in Figure 2 for all countries (i.e. not only anglophone) where the wording referred to absenteeism rather than truancy. As before, survey year is included as a set of dummy variables, with 2015 serving as the reference year. The figures indicate the percentage point change in the truancy rate relative to 2015. From the figure, we observe that countries such as the Czech Republic (CZE), where the term *missed* was used consistently across all survey years, show no noticeable jumps relative to 2015. In contrast, Finland (FIN), the Slovak Republic (SVK), Colombia (COL), Costa Rica (CRI) and Portugal (PRT) display patterns similar to those found for the nine anglophone OECD countries: the year(s) in which the wording changed from *skipped* to *missed* correspond to those with a higher proportion of students reporting having missed or skipped school at least once in the previous two weeks.

Finally, Appendix Table G6 shows pooled estimates for this broader array of countries. Similarly to what we observed for anglophone OECD countries, results show that the use of the word “missed” rather than “skipped” in the question increases estimates of the truancy rate. For this broader set this increase is slightly smaller, ranging between 9 and 16 percentage points.

Figure 4: Estimates of the association between year and truancy for each country



Note: Year was included as a categorical term, with 2015 as the reference year. For all countries but Japan (JPN) and Costa Rica (CRI) the estimation adjusts for gender, migration status, PV1 Math score quartiles and socioeconomic status quartiles. For JPN it does not adjust for migration status, and for CRI it does not adjust for ESCS quartile because they were not available for all years. The countries in blue are those already included in the main analysis. Coefficients and standard errors for all OECD countries are presented in Appendix Table G5.

Would AI be able to spot this problem?

To conclude, we consider whether the use of AI and LLMs might help researchers spot this problem.

As noted in the methodology section and detailed in Appendix D (see prompts 1 and 4) the initial wording used in our prompts was reasonably general. The responses that the AI returned did not raise any particular issue regarding any change of wording, and in fact would lead one to believe that the measures are directly comparable. Indeed, it told us that "*each cycle's questionnaire asks exactly the same question about skipping a day, so responses are directly comparable from 2012 onward*" and that "*fortunately, PISA has used the same skip-question wording in 2012, 2015, 2018 and 2022. Thus, one can directly compare these cycles*". Out of the 1,000 to 2,000-word responses received, the most useful sentence was the following generic statement: "*PISA's international datasets are generally high quality, but translation errors or data entry mistakes (rare for a simple question) could occur*".

Consequently, as noted in the methodology question, our next prompt was much more specific, in that it directly asked about comparability of the data (and the truancy question in particular) across PISA cycles. It again returned statements that repeatedly reassured us of the data's comparability, and its suitability for studying changes in truancy rates across countries over time (see responses under prompts 2 and 5 in Appendix D for further details). Some of the key points it said were that "*the* ***question wording and response scale have been unchanged in 2012, 2015, 2018 and 2022*** *so the basic measure is directly comparable across those years*", that "***trends 2012→2015→2018→2022 can be compared directly"*** and how the "*PISA 2018 Technical Report explicitly confirms that the same skip questions were used in 2012, 2015, 2018*" (emboldened text was provided in the original AI response). Again, across the 1,500-words the AI provided, it was only the following sentence that could have feasibly led to a researcher investigating this problem (if they were willing to ignore the reassurances the AI gave regarding comparability of the measure above) "*the PISA Technical Reports (available from OECD) provide the official wording and any country-specific notes for each cycle. These should be consulted for any subtle changes*."

One thing we noticed in response to our prompts was that the AI seemed to be drawing heavy on the two papers published by Fredriksson (2023, 2024) investigating truancy using PISA data. As noted earlier, these studies included analysis of trends in truancy in the United Kingdom but did not account for the change of wording in 2015. We consequently concluded by investigating the responses provided by the AI after we explicitly told it to ignore the Fredriksson papers (see prompt 3 in Appendix D for further details). This prompt again asked very directly about potential comparability issues with the truancy question. The response started once more by reassuring us of the data's comparability "*The wording has been essentially the same from 2012–2022, which supports comparability*". But it now also provided the following text, which may have been some assistance to researchers in discovering this problem: "*It is important to check that this holds for all national language versions (PISA uses standardized translations). Any minor wording differences or translation issues could affect answers in some countries. PISA reports do not highlight any major wording changes in 2012–2022, so the measure itself is stable – however, researchers should verify the exact phrasing in each cycle's questionnaire if concerned about a particular country's wording*". We note, however, that this has only been mentioned by the AI after using a very specific prompt asking about the methodological problem we are studying – as well as telling it to ignore certain pieces of evidence (the Fredriksson 2023 and 2024 studies).

4. Conclusions

Whenever we get access to new data, the temptation is to dive straight in. The finer details in the survey documentation can wait – we want to get to work straight away. After all, as experienced analysts, we have probably handled hundreds of datasets in our time. But it's easy to become complacent. Data is never one-size fits all, and unexpected traps may await. Unless we are careful, this could cause us to come a cropper, leading to erroneous inferences that undermine all our hard work.

The aim of this paper has been to present a case study illustrating some of the subtle challenges analysts face when handling real world data. Taken at face value, the task we have presented is relatively simple – essentially plotting trends in truancy rates across a relatively small set of countries over four time points. It does not involve particularly sophisticated statistical techniques or complex modelling that are at the heart of many

quantitative social science papers (which bring even further challenges). Yet important nuances with the data mean that it would be incredibly easy for incorrect inferences to be drawn. The change of wording we document – from "skipping" to "missing" school – clearly has a big impact on the meaning of the question, and thus cross-national and temporal differences in the truancy rate. This has led previous research – including work conducted by the OECD who run the survey – to reach some erroneous conclusions (OECD, 2016; Fredriksson, 2023; Fredriksson, 2024). If not spotted, this would also lead one to miss the affect the COVID-19 pandemic has had on unexcused absences in the affected countries (e.g. England, Wales, Northern Ireland, United States). Clearly, this subtle alteration to a single word in a single survey wave can – and in some instances has – impact the results.

This subtle change was not straightforward to spot. The international survey documentation suggests that the question has remained consistent over time. It is also hard to see why some English-speaking countries would change the wording in one single cycle, while others would not (and – in the case of the UK – for this to even differ across its four constituent countries). It this serves as a reminder to users of international datasets to also carefully study the *national* data documentation, including the questionnaires. The devil is often in the detail and – in complex surveys conducted across many countries – unexpected issues may arise.

Our investigations of whether AI could be used to help researchers spot such issues have uncovered some particularly interesting results. Overall – and despite our use of deep research mode drawing on one of the most sophisticated models - it was overconfidently incorrect. The advice the AI gave was that the data were comparable over time and that trends in truancy rates could be reliably estimated. It seems the advice the AI gave drew heavily on the work of Fredriksson (2023, 2024) – where the change in the wording of the truancy question was not discussed – but did not pick-up our sister paper which included a short paragraph briefly noting the question-wording problem (Anders et al., 2025). Analysts relying too heavily on AI would therefore be unlikely to spot the change of wording - and simply reinforce existing claims being made. The AI did however provide some useful general advice regarding the research question we presented to it, including how to handle some of the more unusual features of the PISA data (see Appendix D for

further details). We thus believe that AI can be a useful aid to researchers when they are embarking on a new research question or analysing a new dataset, but only when used with care.

The case study we have presented does have some limitations, with two key issues standing out. First, we have presented a very specific example. It is useful in that it serves as a good illustration as to the dangers that lurk when conducting data analysis, and how even minor, hard-to-spot changes can have a substantial impact on results. But it is admittedly a somewhat unusual situation – a particular idiosyncrasy with these data that is unlikely to occur again. Second, our investigations using AI have used a small set of prompts within one tool. It is thus essentially a qualitative investigation, with the possibility of different prompts or alternative large language models producing different results.

Our case study nevertheless serves as a useful reminder to data analysts, both junior and experienced. There is simply no substitute for carefully studying data documentation and – when first getting into the data – producing a set of simple descriptive statistics. It was through this combination of efforts – along with gut instinct – that led us to identify the change of wording in the PISA 2015 truancy question (and then understanding the implications for our analysis). While AI and large language models are revolutionising the ways we work, it cannot yet replace the care and attention needed when preforming analysis of complex data. Robots may one day come to replace our data analytic jobs – but that time is not yet here.

References

Anders, J., Jerrim, J., Ladron de Guevara Rodriguez, M., Marcenaro-Gutierrez, O., 2024. The rise in teenagers skipping school across English-speaking countries. Evidence from PISA. Unpublished manuscript expected as working paper before publication.

Andres, J., 2024. Understanding the Experiences of Secondary Students Identified as In-School Truant and Their Perceived Contributing Factors to Truancy: When and What Caused the Disconnection. Dissertations, Theses, and Projects.

Bulut, H.C., Bulut, O., 2022. Item wording effects in self-report measures and reading achievement: Does removing careless respondents help? Studies in Educational Evaluation 72, 101126.

DiStefano, C., Motl, R.W., 2006. Further investigating method effects associated with negatively worded items on self-report surveys. Structural Equation Modeling 13, 440–464.
Fredriksson, U., Rasmusson, M., Backlund, A., Isaksson, J., Kreitz-Sandberg, S., 2024. Which students skip school? A comparative study of sociodemographic factors and student absenteeism using PISA data. PLoS One 19, e0300537. https://doi.org/10.1371/journal.pone.0300537
Fredriksson, U., Rasmusson, M., Backlund, Å., Isaksson, J., Kreitz-Sandberg, S., 2023. School absenteeism among students in Germany, Japan, Sweden, and the United Kingdom:: A comparative study using PISA data. Nordic Journal of Comparative and International Education 7. https://doi.org/10.7577/njcie.5034
Kalton, G., Schuman, H., 1982. The effect of the question on survey responses: A review. Journal of the Royal Statistical Society Series A: Statistics in Society 145, 42–57.
Mokhtarian, N., 2024. Roll Call: A Scoping Review for School Attendance Problems Among Youth. Université d'Ottawa | University of Ottawa.
Nathwani, G., Shoaib, A., Shafi, A., Furukawa, T.A., Huy, N.T., 2021. Impact of COVID-2019 on school attendance problems. J Glob Health 11, 03084. https://doi.org/10.7189/jogh.11.03084
OECD, 2022. PISA 2022 Technical Report.
OECD, 2016. Programme for International Student Assessment (PISA). Results from PISA 2015. United Kingdom country note.
Open AI, 2025. Introducing deep research. Retrieved from https://openai.com/index/introducing-deep-research/.
Zeng, B., Wen, H., Zhang, J., 2020. How does the valence of wording affect features of a scale? The method effects in the undergraduate learning burnout scale. Frontiers in Psychology 11, 585179.

ucl.ac.uk/ioe/cepeo