An analysis of potential bias in TALIS 2024 Core in Norway

Mark White

2024-10-20

Contents

This document analyses the risk for bias in population level estimates from TALIS 2024 core (ISCED 2) in Norway. This document is being put together for review by IEA for consideration of levels of non-response bias in the Norwegian data. As laid out in the guidance provided by IEA, non-response bias is more problematic under three conditions: 1) the response rate is low; 2) larger differences between responders and non-responders; and 3) non-response is more highly related to survey data. This document will go through these three conditions in sequence. I note here, though, that the third condition is the most relevant for bias Hartman & Huang (2024a). Bias can actually be surprisingly independent of response rate Groves & Peytcheva (2008). When non-response is independent of survey data, even in cases where low response rates and high differences between responders and non-responders, bias can be very low. The problem, though, is that survey data is not collected for non-responders so one can never observe condition (3), but has to estimate it based on a set of assumptions. In the rest of this document, I go through each of the conditions that might lead to non-response bias in sequence.

1 Sample in Norway

The intended population for TALIS core is ISCED level 2 schools that provide instruction to the general population (e.g., not special education schools or schools for adults). Norway set the additional restrictions of not including very small schools (those with 4 or fewer teachers) and not including Sami-language or international schools. The sampling frame was generated in 2022 and updated in 2024. The decision was made that schools must fulfill these criteria (and so exist at all) at both time points to be included in the target population. However, one school that was included in the TALIS sample, but had only three teachers in 2024 was included in the target population.

The original sampling frame included 1206 schools that were identified at the ISCED 2 level and served the general population. Twenty-five of these schools were found to only serve special populations at the ISCED 2 level and two schools were found to only offer supplemental educational services to other schools. Fourteen schools were shut down between 2022 and 2024 and 21 new schools were found. Nine schools stopped offering classes at the ISCED 2 level. This lead the 2024 sampling frame to include 1177 total schools in the updated sampling frame. 1023 (87%) schools met the inclusion criteria at both time points (9 new schools and 13 small schools would have been added to this number if the inclusion criteria were judged only in 2024, but schools must meet the criteria from both years). 27 schools were international schools (2%). Six schools were SAMI language schools (1%). 50 schools were too small at both time points, 41 were too small only in 2024, and 13 too small only in 2022 (9%). One school seemed to be inadvertently left out of the 2022 sampling frame and so excluded here and one school was excluded because it only existed on a boat.

2 Participation Rates in Schools and Teachers

2.1 Principal Population

Table 1 shows the results of participation at the school level. Here, we see that overall participation rate is roughly 70%, but only 45% for sampled schools. This participation rate drops somewhat when using weighted numbers (see Table 2). Of the 64 refusing schools, 22 (10%) were deemed as refusing due to not returning a principal survey.

Table 1: School Participation

	Explicit Strata	**Total sampled schools**	**Ineligible schools**	**Sampled**	**1st replacements**	**2nd replacements**	**Refusing schools**	**Participating Schools**
ISCED 2 only	142	0	69	27	9	37		

ISCED 1 and ISCED 2	70	1	25	10	7	27	
Total	212	1	94 (45%)	37 (18%)	16 (7%)	64 (30%)	147 (70%)

Table 2: Weighted School Participation

Type	Sampled schools (%)	With 1st replacement schools (%)	With 2nd replacement schools (%)
Unweighted	45%	62%	70%
Weighted	41.8%	58.5%	66.7%

2.2 Teacher Population

Table 3 shows the participation at the teacher level. Here, we see that the overall participation rate is roughly 53% and close to 34% before including replacement schools. This participation rate remains roughly the same when being weighted (see Table 4). Of the 99 refusing schools, 57 (27%) were deemed as refusing due to less than 50% of teachers returning a survey.

Table 3: Table 3: Teacher Participation

Explicit Strata	Total sampled schools	Ineligible schools	Sampled	1st replacements	2nd replacements	Refusing schools	Excluded schools
ISCED 2 only	142	0	55	25	5	57	0
ISCED 1 and ISCED 2	70	1	16	10	1	42	0
Total	212	1	71 (34%)	35 (17%)	6 (3%)	99 (47%)	0

Table 4: Table 4: Weighted Teacher Participation

Type	Sampled schools (%)	With 1st replacement schools (%)	With 2nd replacement schools (%)
Unweighted	33.6	50.2	53.1
Weighted	34.1	50.8	53.9

3 Differences between responding and non-responding schools

In this section, I review differences between participating and non-participating schools. In Norway, there is a reasonable large amount of data collected at the school level (or for specific grades within schools), but very little information collected at the teacher level. The analyses will, for the most part, then be conducted at the school level. At the end of this section I look at characteristics of teachers that were collected on the teacher sampling forms.

There are several categories of schools that are useful to distinguish between: participating schools, those that were excluded due to low response rates (these schools initially agreed to participate), schools that actively refused participation, and schools that were not contacted (either due to not being selected or because they were a replacement school to a participating school). I focus on the following comparisons. (1) Comparing participating schools (principal level) to all schools, which allows us to explore if population estimates from participating schools match the population values; (2) The same comparison as (1), but for schools participating at the teacher level; (3) comparing participating schools (without replacement, at teacher level) to the population, so see if the sample of originally chosen schools can provide non-biased estimates of population values; (4) comparing participating schools to those that refused to participate; and (5) comparison (4) but with without including replacement schools. These comparisons provide a range of views into potential biases in the sampling.

Drawing on work on evaluating non-random samples within the quasi-experimental design (QED) literature (Ho et al., 2007), I do not just look at whether the population falls within the 95% confidence interval generated by the participating schools, but I also look at the standardized mean difference between the sample and population. In the QED-literature, it has been noted that non-significant differences between matched control and treatment groups can still lead to substantial bias and that establishing that standardized differences are low (ideally below 0.1 and certainly below 0.25) is vital for preventing bias in causal estimates (Ho et al., 2007). I present standardized mean differences (SMD) between groups in the graphs below, using the cited thresholds for interpretation. The differences are standardized by the population standard deviation when comparing a sample to the population and by the pools standard deviation when comparing two samples.

3.1 Comparing Schools that are participating at the principal level to the population.

This section compares schools that were *deemed as participating for the principal survey* by IEA to the set of all schools in the revised sampling frame. Here, I look at whether sample-based estimates capture population mean values and the standardized mean differences (SMD) between the sample and the population. Figure 1 shows the results for the location of the school (i.e., which county or fylke the school belongs to). The confidence interval contains the population estimate for all fifteen fylke and the average standardized mean difference between the sample and population is 0.08 on average, but is over 0.25 in one case (Telemark) and over over 0.1 in five more cases. This would suggest the sample is quite balanced across school location.

Figure 2 shows the financing for schools across Norway. There are not large differences in general in financing across schools, but financing occurs at the kommunen (municipality) level so there are some differences in funding. Figure 2 shows that the sample estimate confidence intervals cover the population score and standardized mean differences are near or below 0.1, suggesting that the sample is quite representative in terms of financing at the per student level.

Figure 3 shows the sample-based estimates of school-level national test scores. The national test had a mean of 50 and standard deviation of 10 at the population level in 2022 when the scales were reset. The mean score estimates from the sample were within 1 point of the national mean (SMD<=0.16 across subjects). The increase across years and across grades (from 8th to 9th grade) are also shown and the population values fall within sample estimates with relatively small SMD. The uncertainty (Usikkerhet) in scores is also shown. Overall, this would suggest that the sample is not problematic in terms of scores on the national test.

Figure 4 shows the sample-based estimates of various school-level resources and their match to population values. For the number of front office positions, hours for teaching assistants per student, student-teacher ratio, number of admin per student, and amount of homework help per student, the population estimates include the population values in the confidence intervals. However, the estimate for amount of homework help per student shows a quite large SMD with the population, suggesting that there may be important non-representativeness in this variable in the sample. Note though from Figure 7 that the percentage of students receiving homework help is quite balanced.

Figure 1: Comparison of estimates from participating schools to population value for school location (county/fylke)

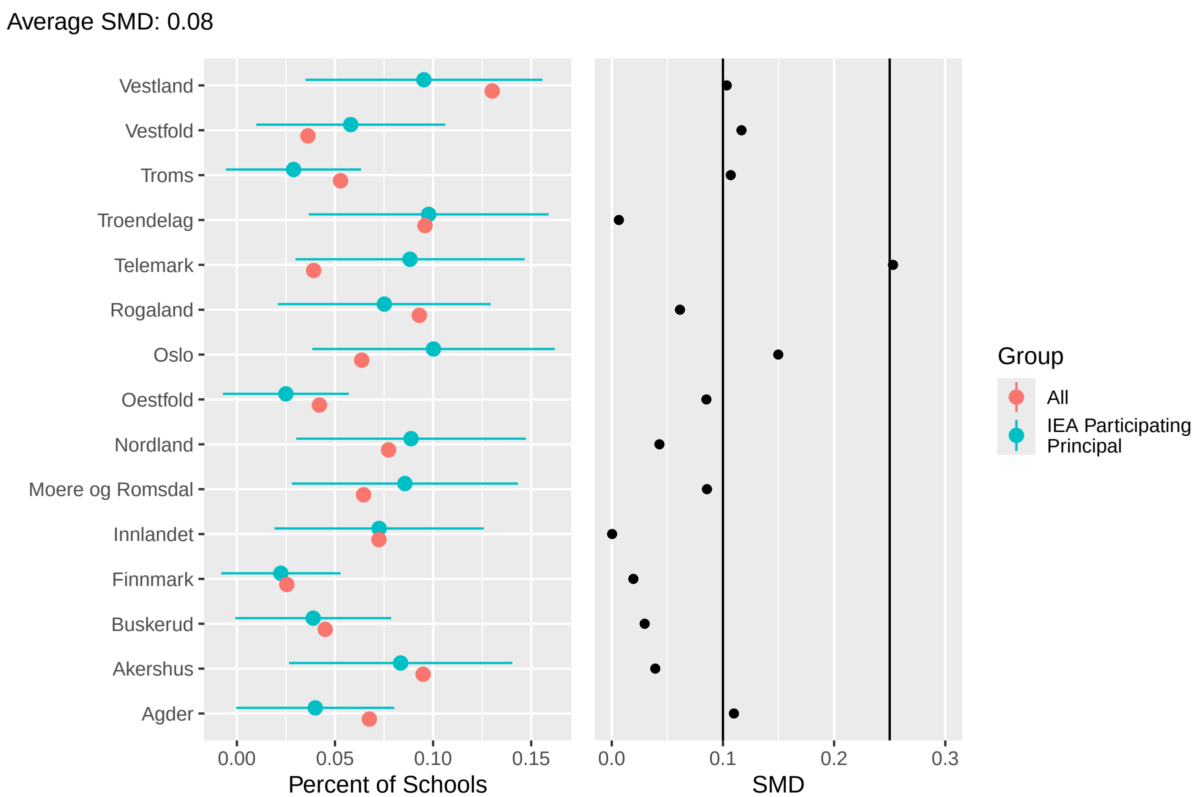
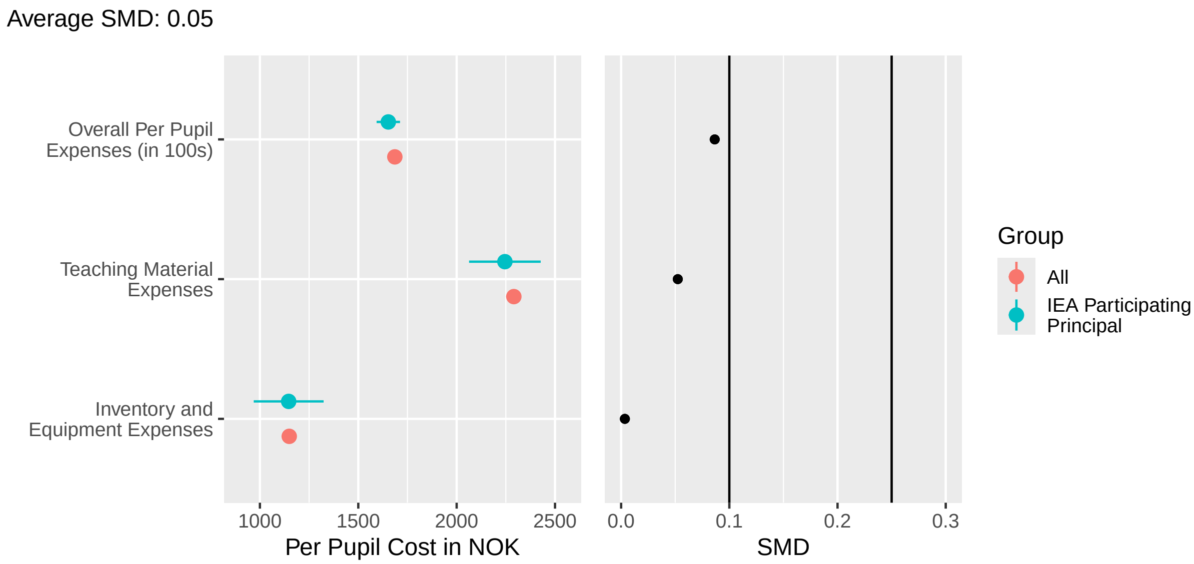


Figure 5 shows the same sample estimates for a set of variables related to school size. Here, the sample estimate do quite well and SMDs are generally low. Figure 6 shows the same sample estimates for a set of variables related to school characteristics, including number of variables related to changes in the school. Here again the sample estimates seem to function well and the SMDs are generally small. This would suggest the sample is quite representative in terms of both stable and non-stable school characteristics.

Figure 7 shows the same sample estimates for a set of variables related to teacher and student characteristics. Here again, there are no clear signs of problems in the sample. The sample provides accurate estimates for teaching by teachers without formal qualifications, percentage of special ed instruction in the school, student bullying, percentage of students transitioning immediately to high school, percentage of students receiving homework help, students in Norwegian as a second language instruction, and absences per student. Here, SMDs are all under 0.15 and generally under 0.1. This would suggest that the sample is quite representative in terms of student and teacher characteristics (as aggregated to the school level).

Figure 8 shows the same sample estimates for school-level aggregates of student survey data from the national student survey. The only population value that does not fall into the sample estimate confidence intervals is for the average score on teaching culture. This scale consists of three items (students report on the frequency of these items): There is a calmness that facilitates working during lessons; In my class, we think it is important to work with school work; and My teacher things it is ok that students make mistakes because we can learn from them. Here, though, the SMDs are a bit higher, with the teaching culture SMD over 0.25 and most of the other SMDs falling between 0.1 and 0.25. This would suggest that the sample schools have a more positive teaching culture (as observed by students) than expected (after adjusting for weights). Further, the survey score estimates are almost unanimously slightly higher than population values, which could further suggest a slight bias towards schools with slightly higher climates participating in TALIS,

Figure 2: Comparison across per pupil expenses



though the differences are quite modest on the survey scale (i.e., if took a typical class from the population and transformed it to a typical class in the sample, one would have to increase the score of 2 out of 25 students by a single point on the survey scale [while 23 students gave the same score], which is to say that student experiences in schools would not seem to be highly discrepant between sample and population based on the student survey data).

3.1.1 Summary

The overall pattern of results suggests that the sample estimates can generally recreate population values with one exception (<5% were exceptions). The SMDs were also quite small, suggesting that observed differences are unlikely to be large enough to create biases. That said, there were some larger SMDs, which might lead to some biases, but random variation would be expected to lead to some larger SMDs. Then, we have no evidence that would lead us to suspect that the participating schools at the principal level is a non-random sample.

Figure 3: Comparison across national test Scores

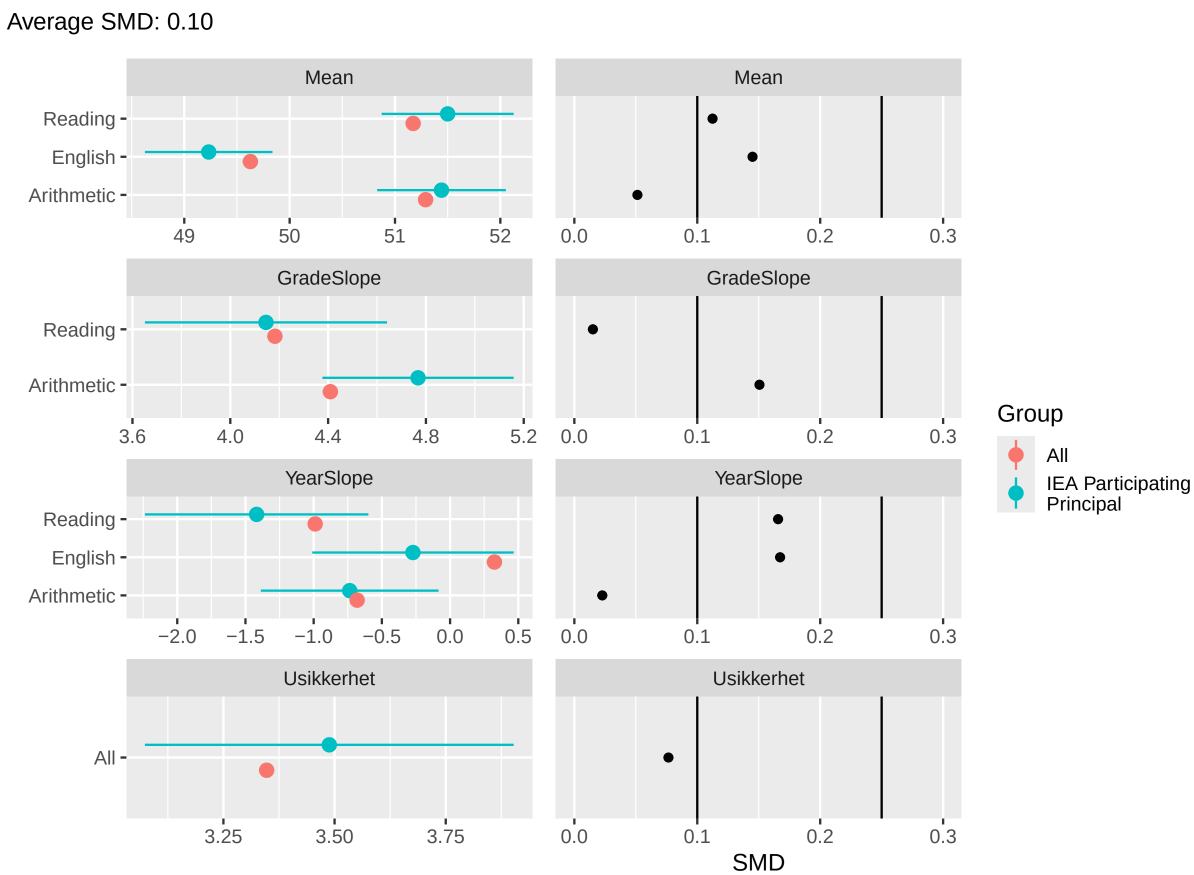


Figure 4: Comparison across amount of School Resources

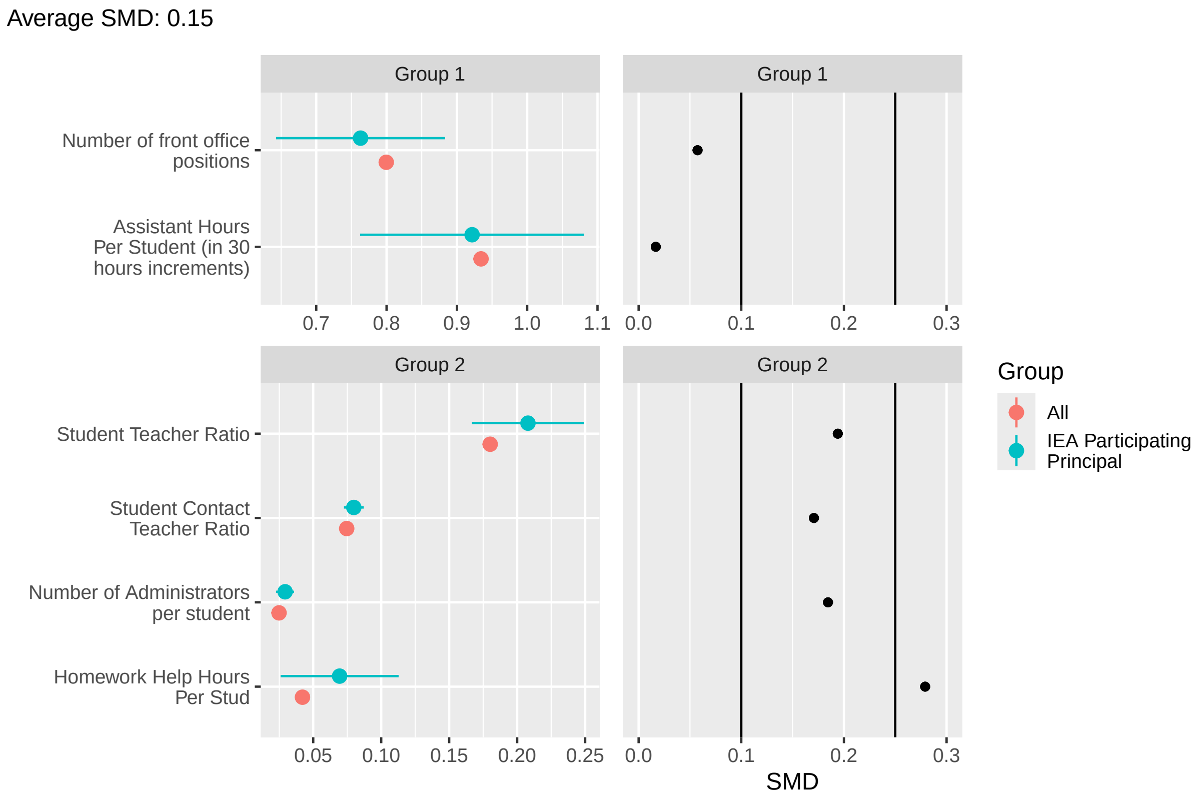


Figure 5: Comparison across school Size

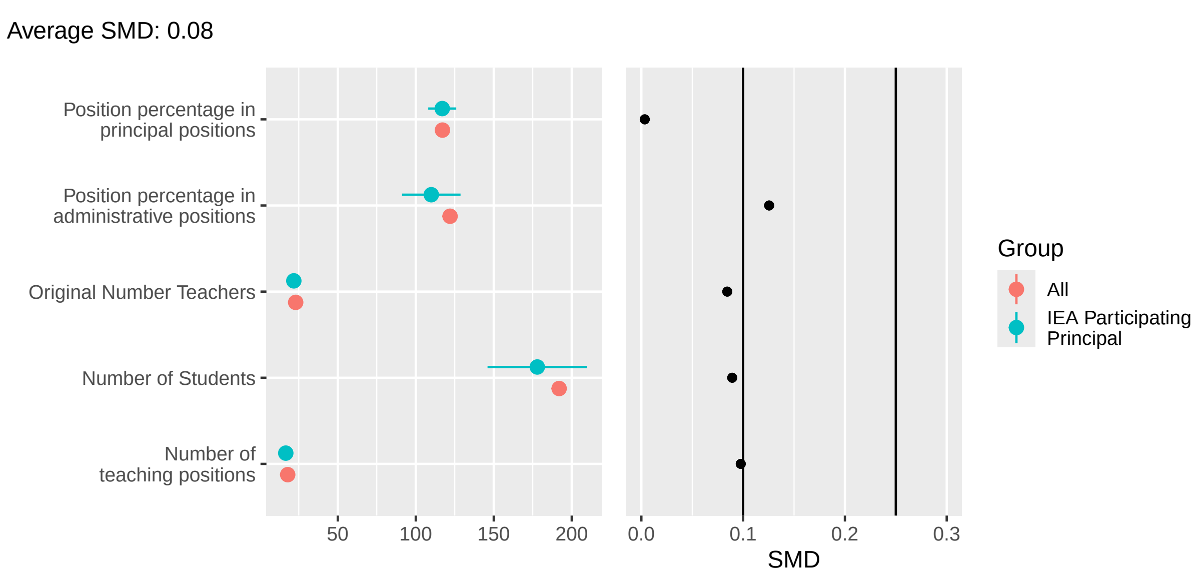
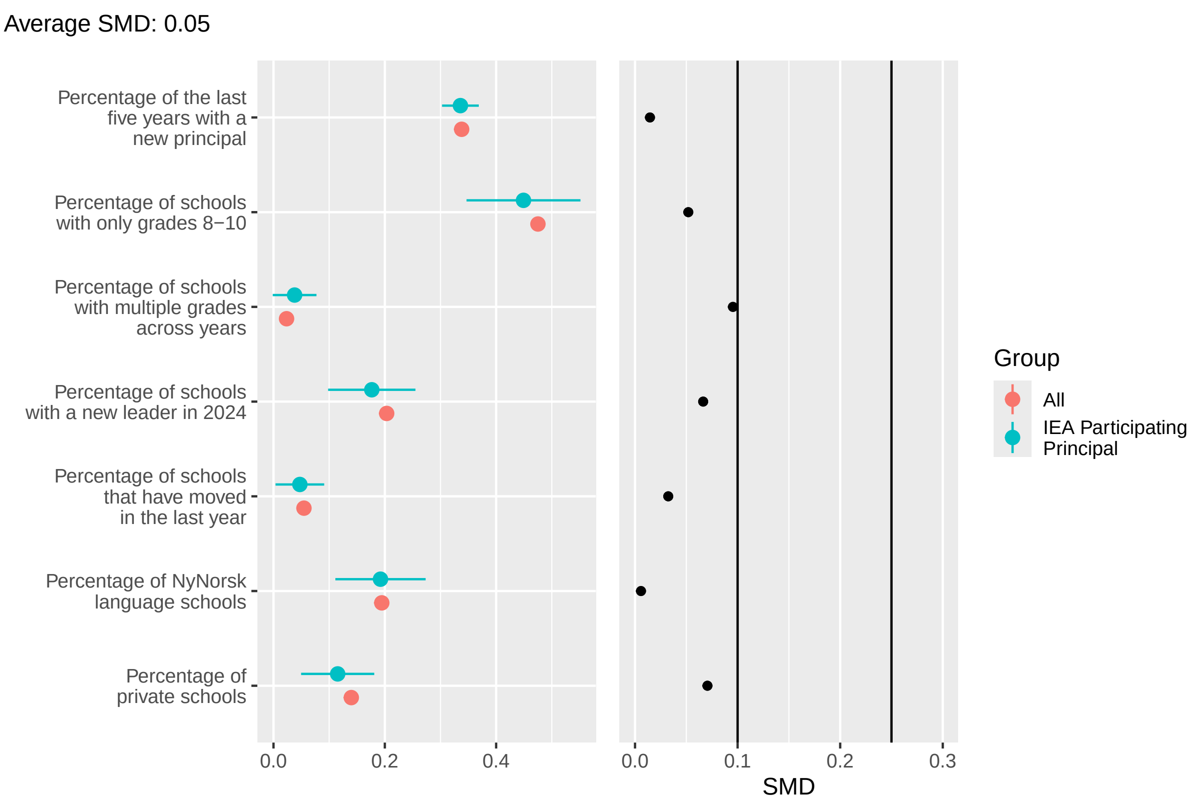


Figure 6: Comparison across characteristics of school



shows the same sample estimates for a set of variables related to instruction, bullying, and student characteristics.

Figure 7: Comparison across teacher and student Characteristics

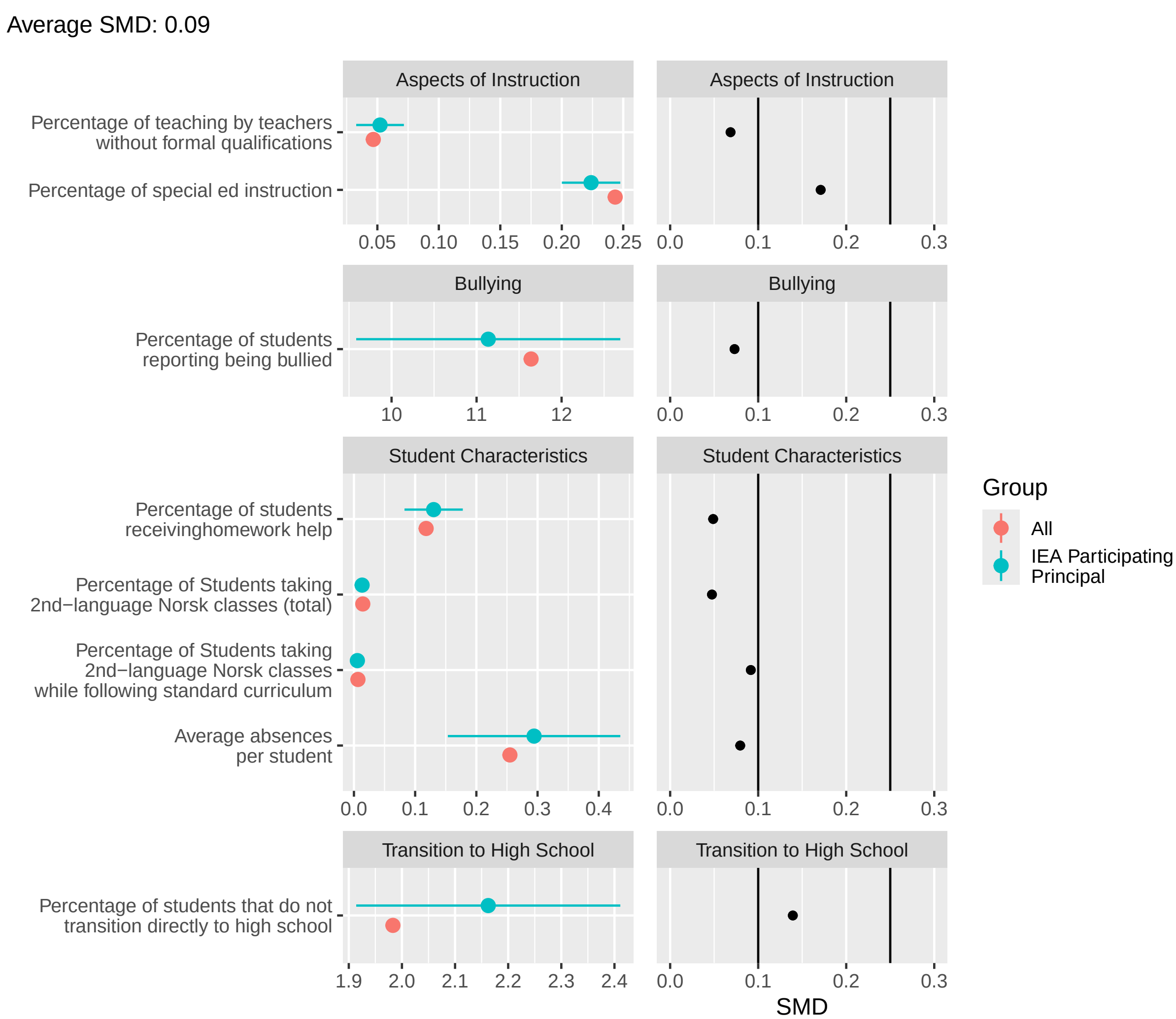
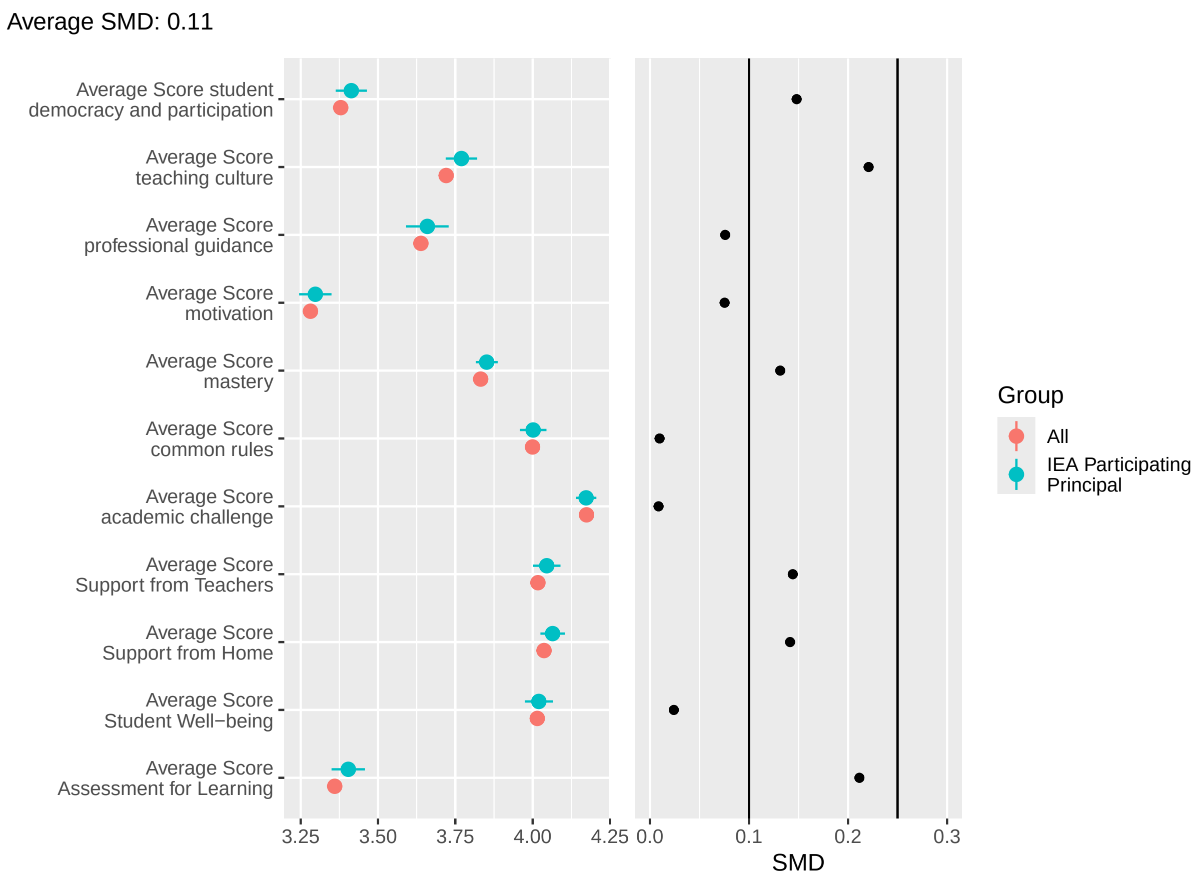


Figure 8: Comparison across Student Survey Scores



3.2 Comparing Schools that are participating at the teacher level to the population.

This section compares schools that were *deemed as participating at the teacher level* by IEA to the set of all schools in the revised sampling frame. Here, I look at whether sample-based estimates capture population mean values and the standardized mean differences (SMD) between the sample and the population. In general, the results are quite similar to the previous section for schools participating at the principal level.

Figure 9 shows the results for the location of the school (i.e., which county or fylke the school belongs to). Here, all estimates from the sample match population values, though SMDs are relatively high for Vestfold (a relatively small fylke) and Moere and Romsdal (a larger fylke). Figure 10 shows the sample-based estimates for school financing. There are not large differences in financing across schools in Norway, but financing occurs at the kommune (municipality) level so there are some differences in funding. Figure 10 shows that the sample estimate confidence intervals cover the population score. The SMDs are just over 0.1 for overall spending and spending on material expenses, with participating schools under-estimating these values (by about 150 kr per student [~15 euro] for teaching material expenses and about 5.000 kr per student [~500 euro, ~3% relative error] overall). These are small differences.

Figure 9: Comparison of estimates from participating schools to population value for school location (county/fylke)

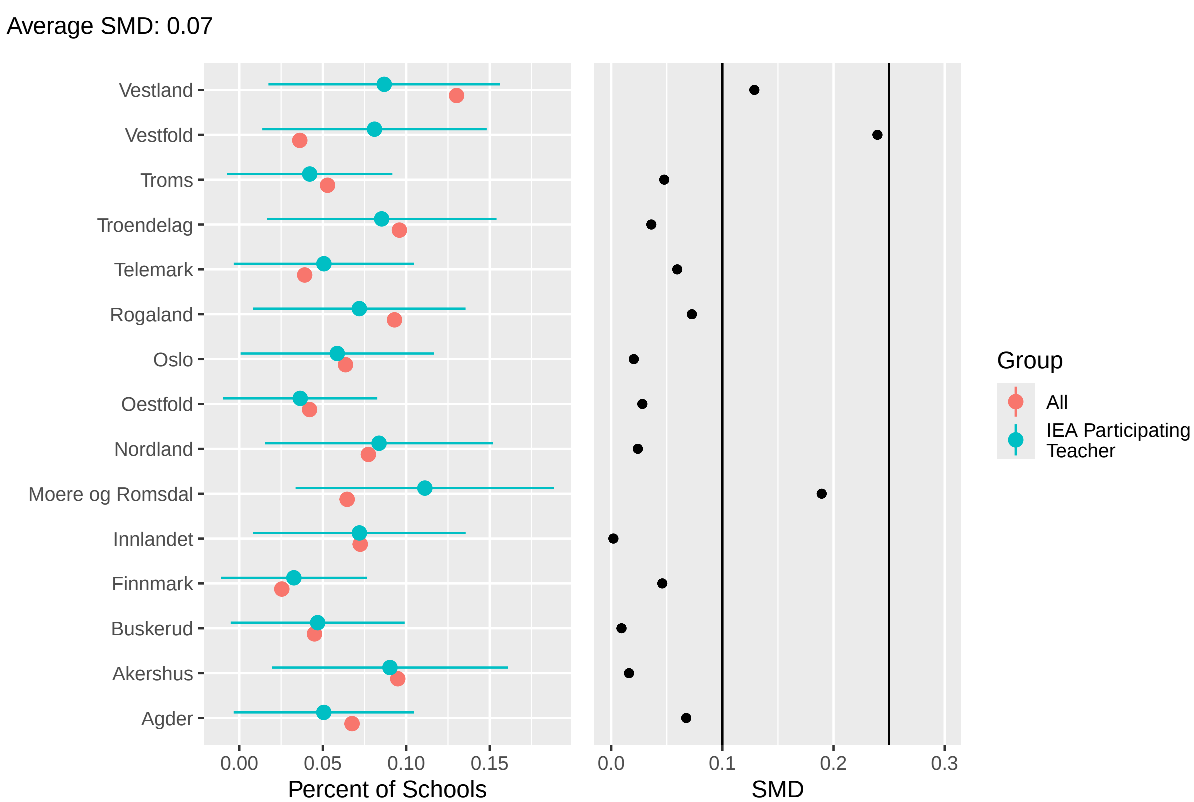
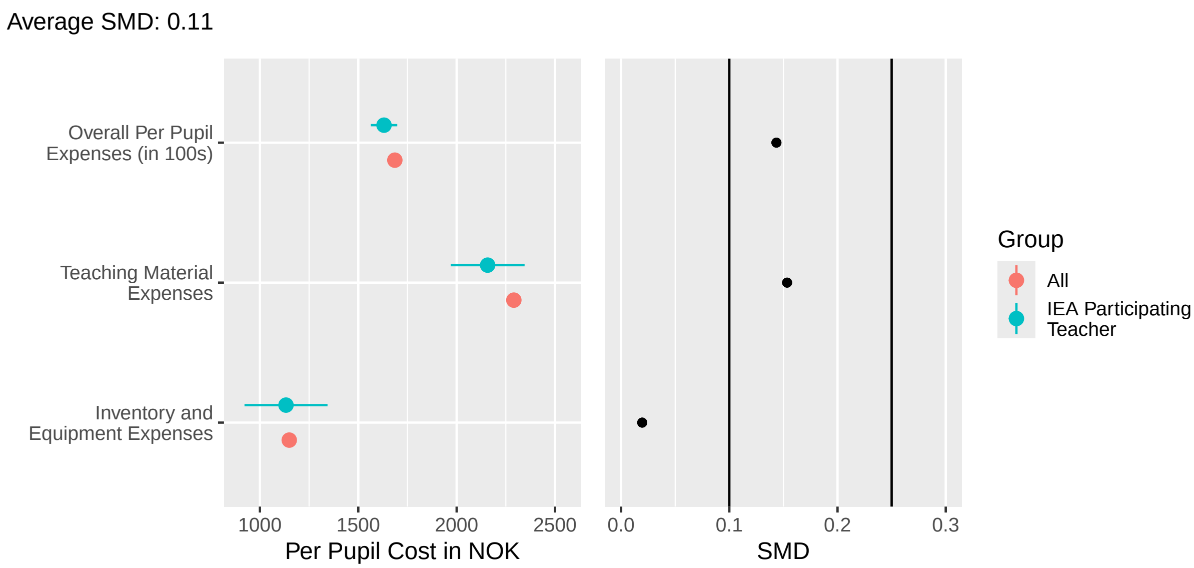


Figure 11 shows the sample-based estimates of school-level national test scores. The national test had a mean of 50 and standard deviation of 10 at the population level in 2022 when the scales were reset. The mean score estimates from the sample were within 1 point of the national mean (SMD<=0.25 across subjects). The increase across years and across grades (from 8th to 9th grade) are also shown and the population values fall within sample estimates with relatively small SMD. The year slope, which is the change in scores across years, does have higher SMD with sampled schools having slightly lower slopes (larger decreases across

Figure 10: Comparison across per pupil expenses



years), but SMDs remain below 0.25. The uncertainty (Usikkerhet) in scores is also shown. Overall, this would suggest that the sample is not problematic in terms of scores on the national test.

Figure 12 shows the sample-based estimates of various school-level resources and their match to population values. Here, most variables capture the population estimates, but the number of front office positions is lower in the sample than the population. This is an interesting point as the front office staff often helped to coordinate aspects of the survey and fill in the teacher sampling forms. The higher student-teacher and student-administrator ratios in the sample also suggest that sample schools might have traded teacher positions for administrative positions relative to the population. The SMDs across these variables are also near 0.25 for most variables, which would suggest the differences here approach meaningful levels.

Figure 13 shows the same sample estimates for a set of variables related to school size. Here, the estimate for position percentages in administrative positions (e.g., 250% position percentages implies the equivalent of two and a half full time persons, but it could be that 5 people are each 50% in the role) from the sample does not capture the population value and the SMD for this variable is over 0.25. The estimates for the other variables, though, do capture the population values. It would seem that sample schools, then, are about equal in size to the population schools, but they have more staff in administrative positions (typically assistant principals or coaches). This difference is approximately a quarter of a full time position (25%).

Figure 14 shows the same sample estimates for a set of variables related to school characteristics, including number of variables related to changes in the school. Here again the sample estimates seem to function well and the SMDs are generally very small. This would suggest the sample is quite representative in terms of both stable and non-stable school characteristics. However, the SMD for percentage of schools with a new principal is ~0.2, suggesting that sampled schools are less likely to have a new principal than population schools.

Figure 15 shows the same sample estimates for a set of variables related to teacher and student characteristics. The sample provides accurate estimates for amount of teaching by teachers without formal qualifications, student bullying, percentage of students transitioning immediately to high school, percentage of students receiving homework help, students in Norwegian as a second language instruction, and absences per student. However, the percentage of instruction dedicated to special education is underestimated by the sample with a SMD of over 0.3. This would suggest that sample schools include less special education than the population on average and the weights do not properly adjust for this gap.

Figure 11: Comparison across national test Scores

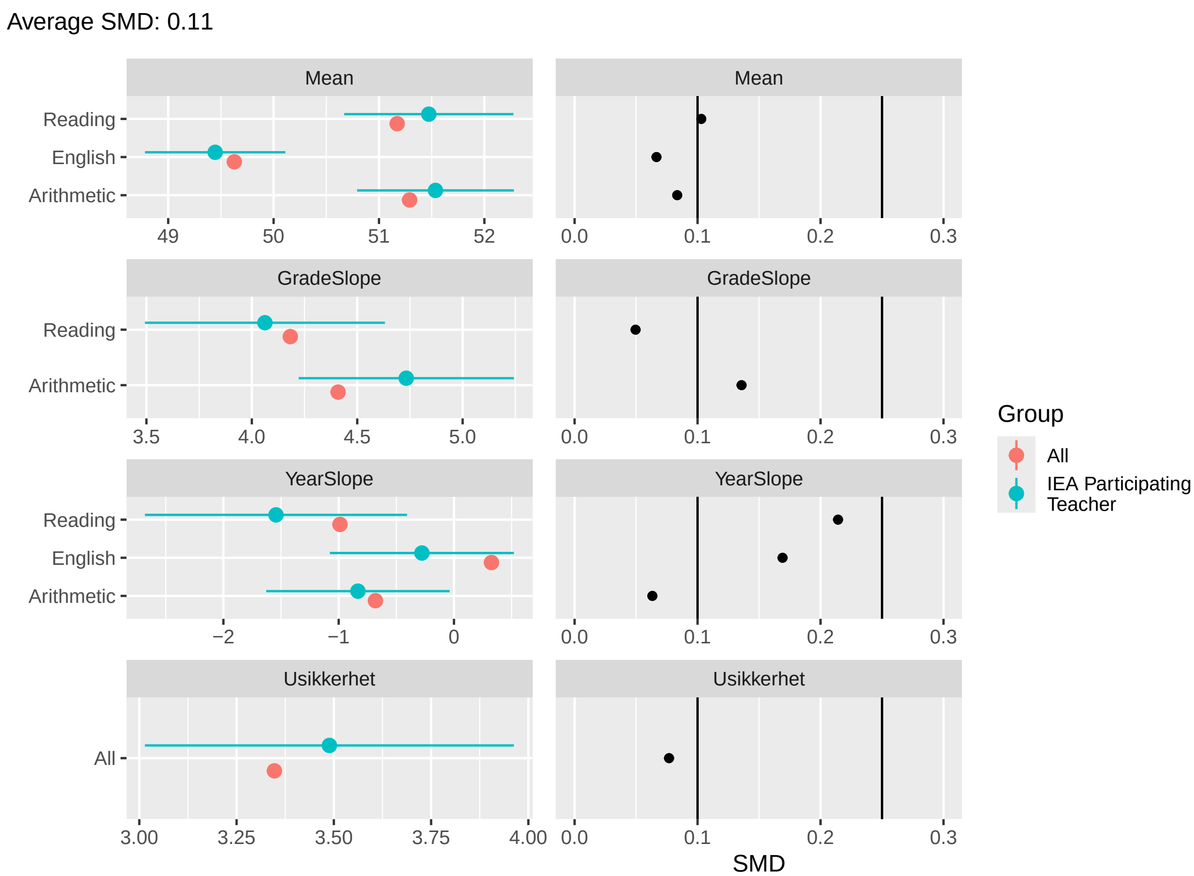


Figure 16 shows the same sample estimates for school-level aggregates of student survey data from the national student survey. All population values fall within the sample estimate confidence intervals. However, the teaching culture and assessment for learning scales have SMDs above 0.25, suggesting that these variables could contribute meaningful bias. That said, absolute differences in scores are quite small.

3.2.1 Summary

In summarizing this section, I note that the difference between the previous sample of schools that were participating at the principal level and this sample of schools participating at the teacher level is driven mostly by schools where 50% of teachers did not respond and so the school was dropped from the sample. This seems to have had a small, but noticeable impact on the accuracy of the sample estimates. Only two estimates have confidence intervals that did not include the population value (~3%) and average SMDs across for the categories of school resources, test scores, teacher/student characteristics and student survey were well above the 0.1 cutoff. Several specific variables had quite high SMDs (above 0.25). SMDs are expected to vary across variables, however, so some large values are expected. As an overall picture, I would suggest that there may be some small biases in the sample, but given the low response rate, the accuracy of sample estimates and representativeness of school characteristics (i.e., modest SMD values) gives a surprisingly positive picture.

Figure 12: Comparison across amount of School Resources

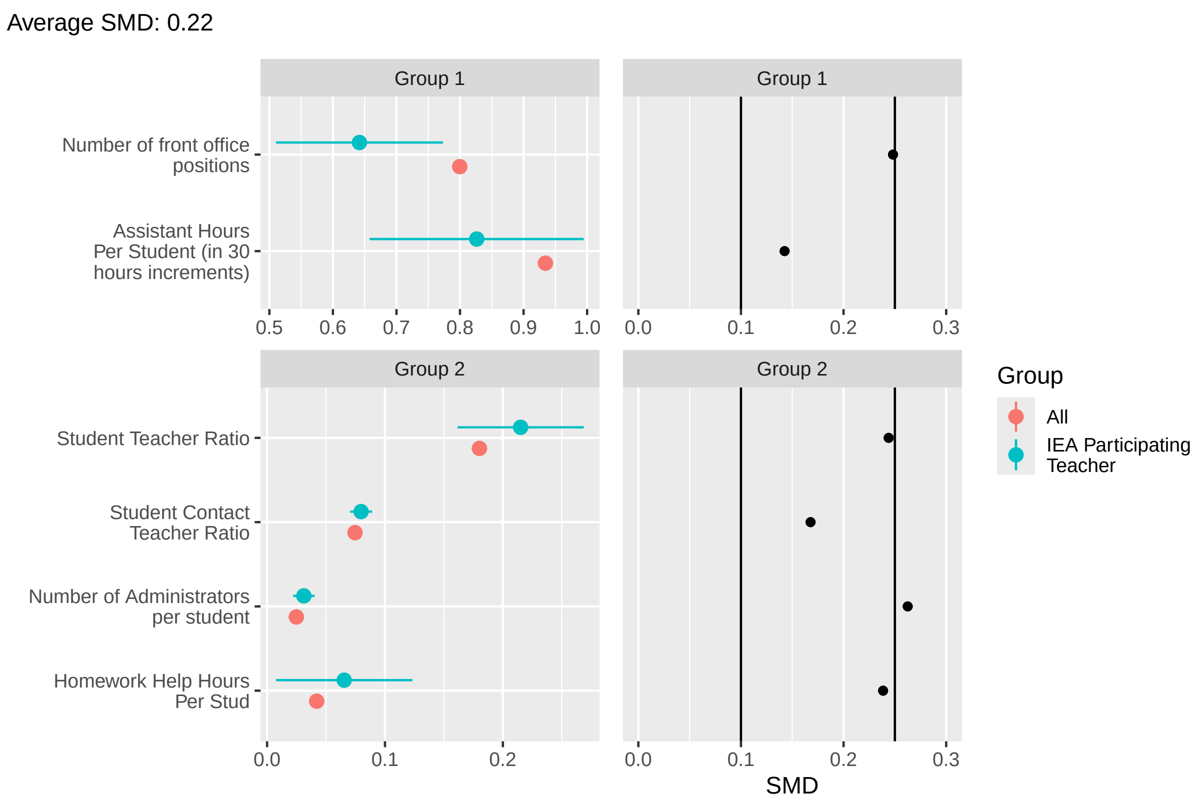


Figure 13: Comparison across school Size

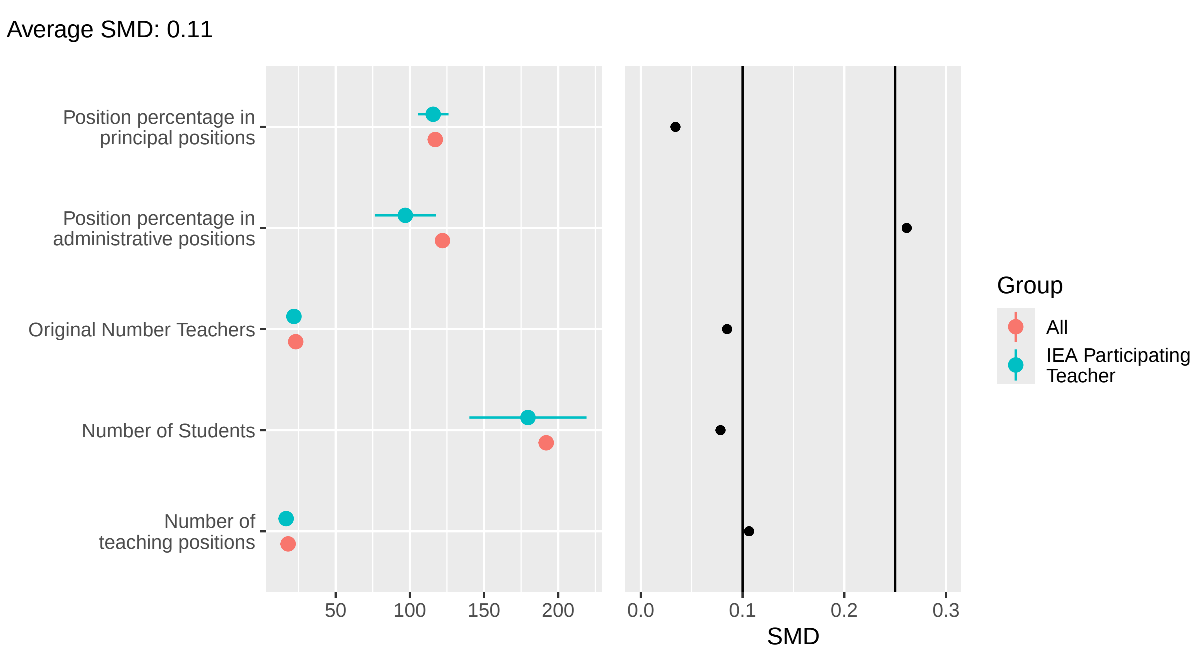
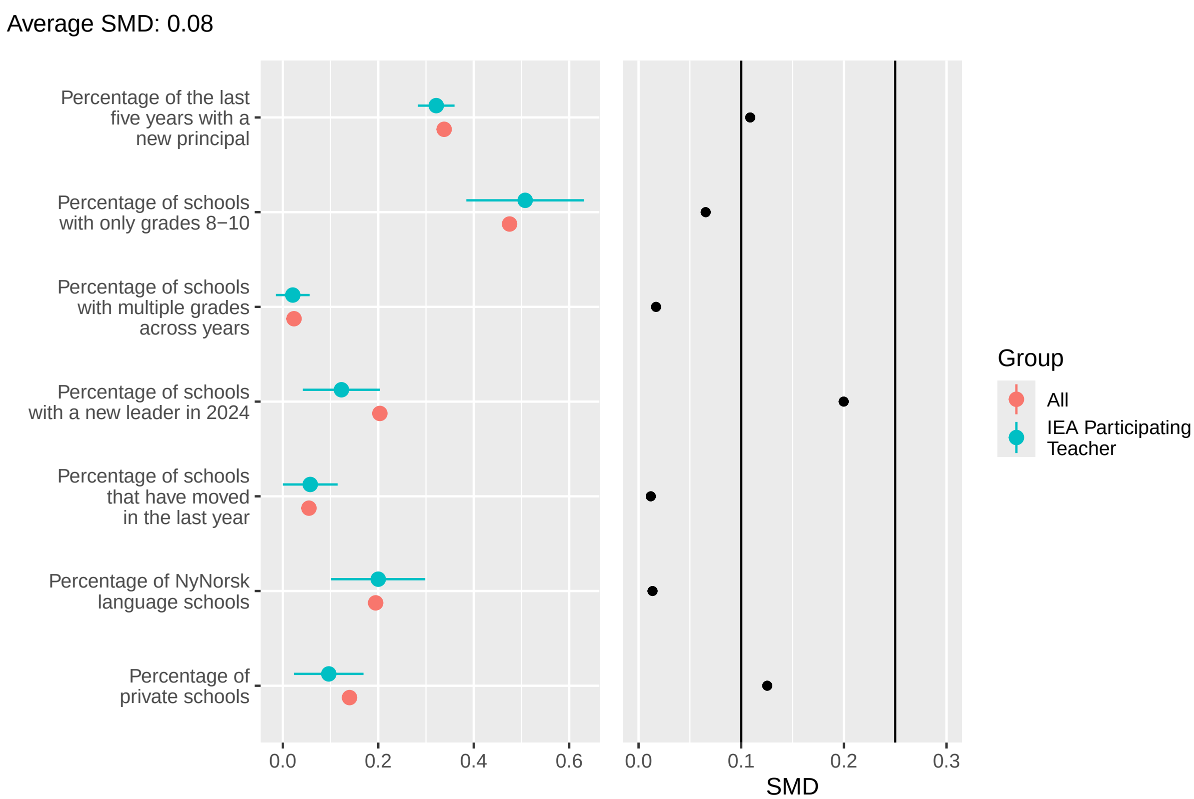


Figure 14: Comparison across characteristics of school



shows the same sample estimates for a set of variables related to instruction, bullying, and student characteristics.

Figure 15: Comparison across teacher and student Characteristics

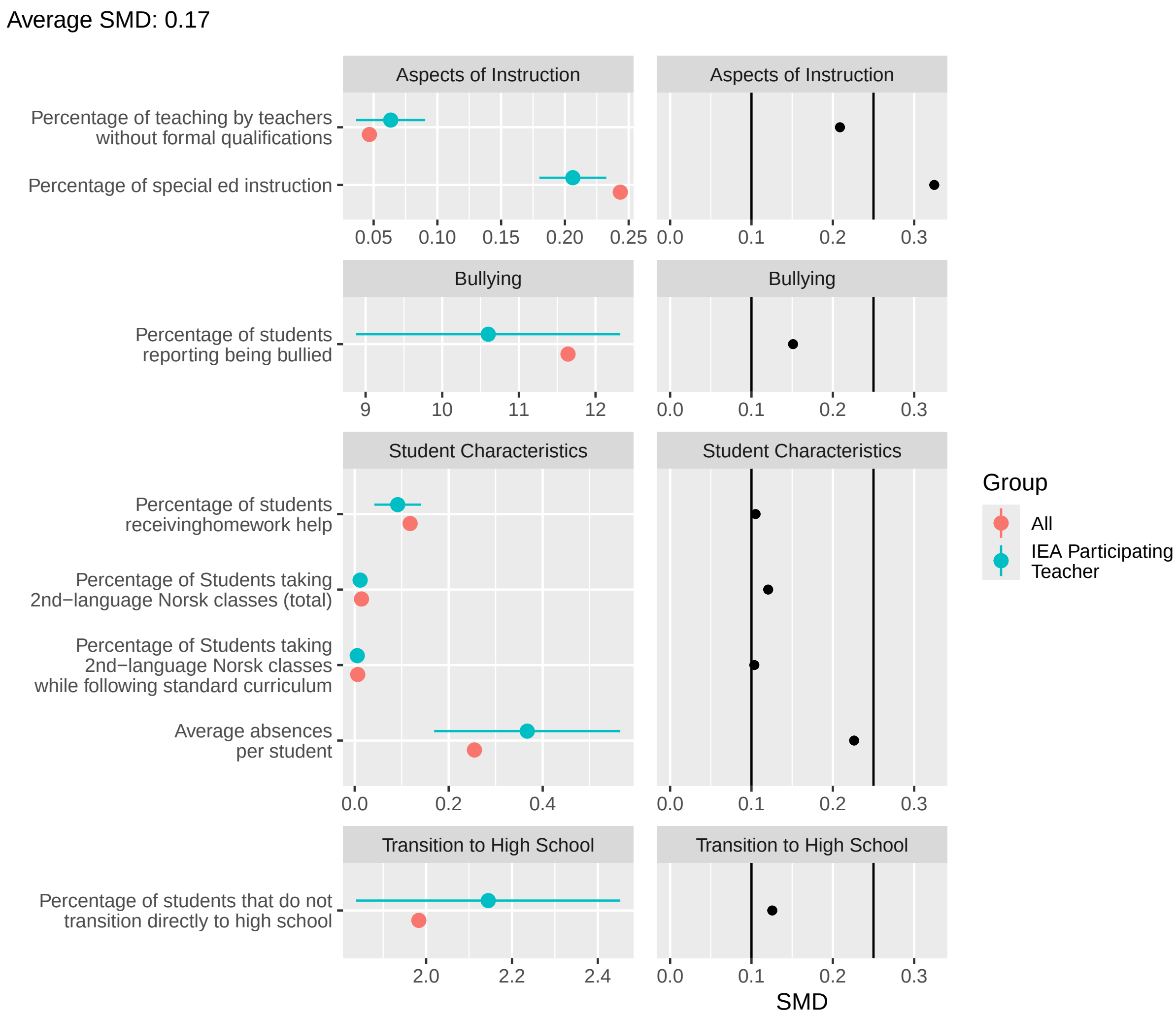
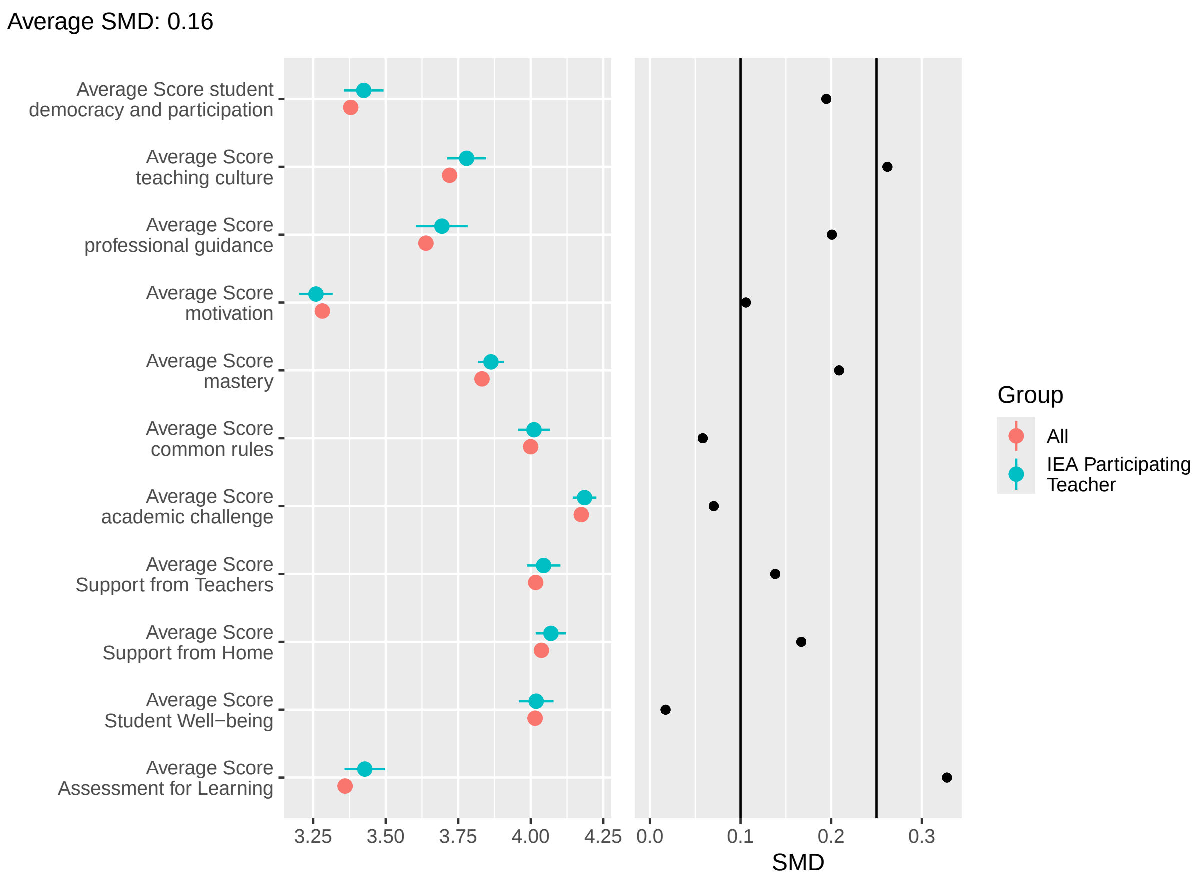


Figure 16: Comparison across Student Survey Scores



3.3 Comparing participating Schools from the original sample (without replacement) to the population.

This section compares schools that were *deemed as participating at the teacher level by IEA and in the original sample (without replacement)* to the set of all schools in the revised sampling frame (i.e., the population after removing small and non-Norwegian schools). Here, I look at whether sample-based estimates capture population mean values and the standardized mean differences (SMD) between the sample and the population. In general, the results are quite similar to the previous section for schools participating at the teacher level.

Figure 9 shows the results for the location of the school (i.e., which county or fylke the school belongs to). Here, the overall SMD remains approximately the same, but SMD for Telemark becomes much larger (though the population value remains in the confidence interval). It would be expected to have more widely varying SMDs when the sample is smaller due to random variation, so this would not necessarily indicate problems.

Figure 17: Comparison of estimates from participating schools to population value for school location (county/fylke)

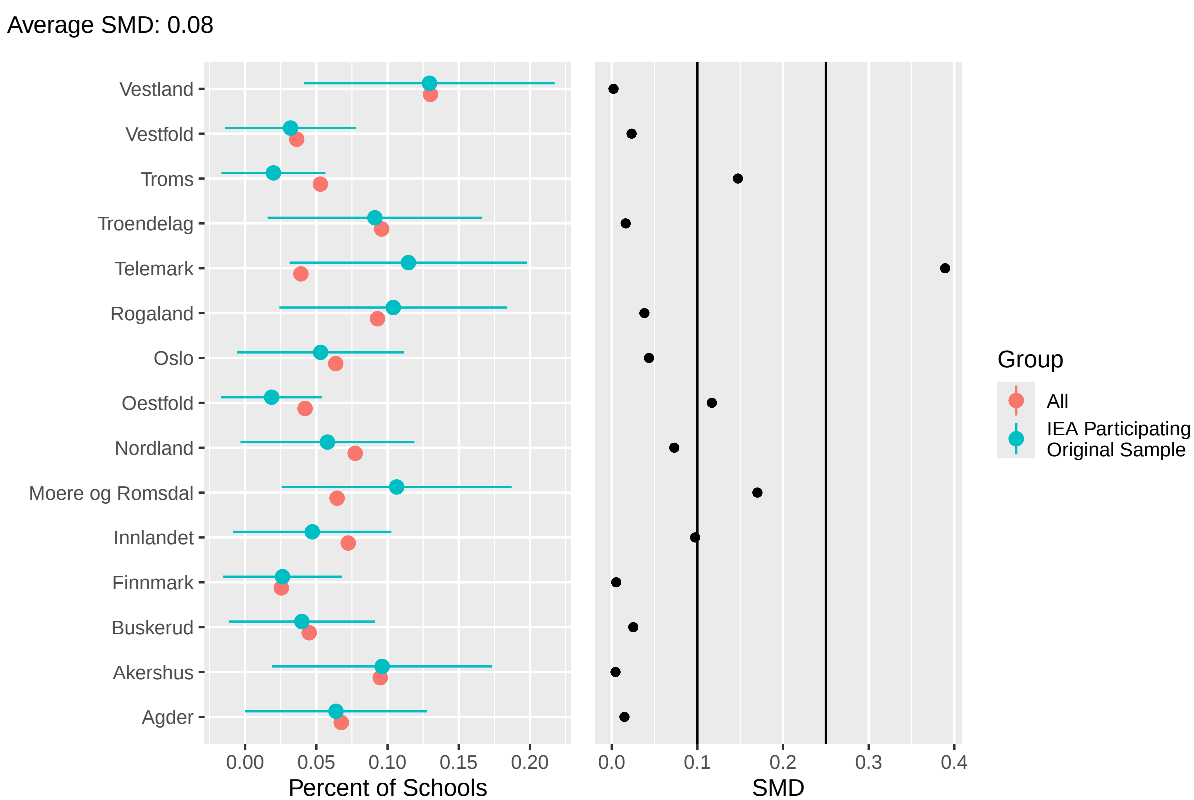


Figure 18 shows the sample-based estimates for school financing. The results here are approximately the same for the finding for all participating schools, but SMDs are actually a bit lower. Figure 19 shows the sample-based estimates of school-level national test scores without using replacement schools. Here, the population values fall within the estimated confidence intervals, but the SMDs are higher than for the full participating sample, especially for English and the slope of the reading test score across years. This would suggest that schools in the original sample (without replacement) are somewhat less representative of the population, compared to all participating schools. Figure 20 shows the same pattern when looking at school resources. The estimates without using replacement schools are accurate, but show higher SMDs and more error than estimates using all participating schools. There are quite large differences in the measures for teacher- and administrator- to student ratios and homework help, suggesting that sampled schools have more staff and more time spent in homework help than in the population. Figure 21 shows that estimates without replacement schools do not decrease in quality (compared to using all schools) when looking at school size metrics. This should not be surprising given the implicit stratification on school size.

Figure 18: Comparison across per pupil expenses

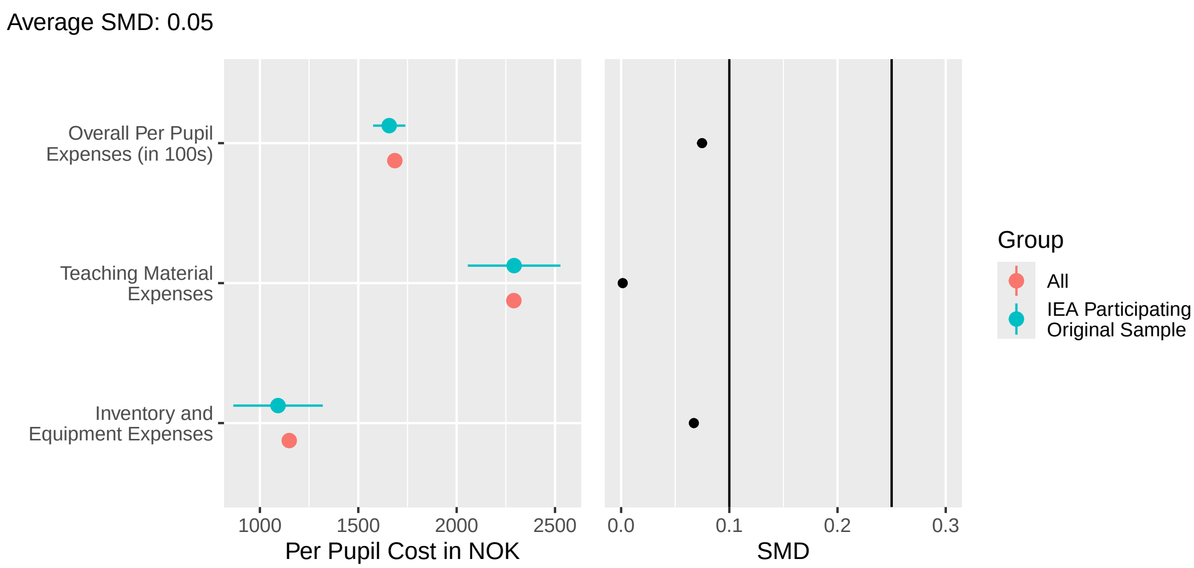


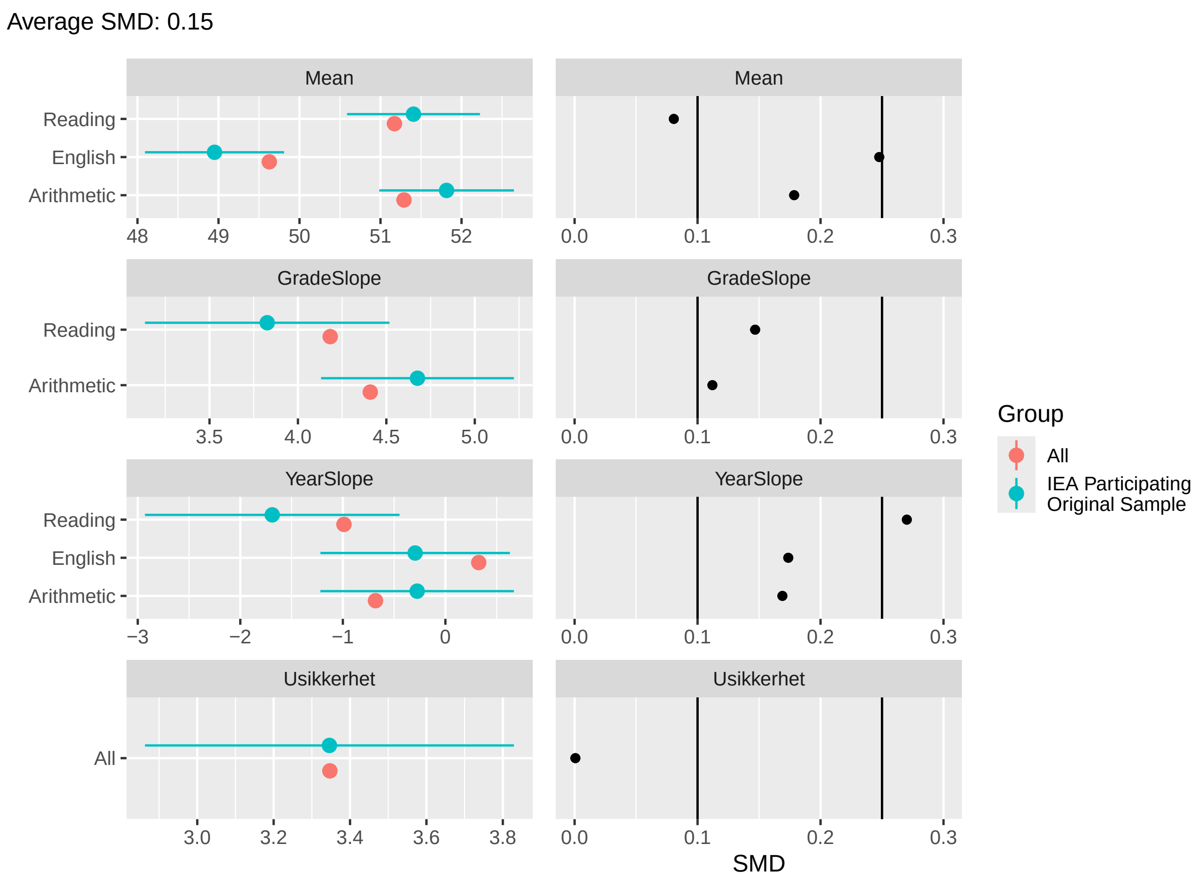
Figure 22 shows the same sample estimates for a set of variables related to school characteristics when using the original sample. Here again, the quality of estimates decreases and SMD slightly increases, but population values typically fall within the estimated confidence intervals. The one exception is the percentage of private schools, where the original sample significantly under-represents private schools. Adding in the replacement schools removes this bias (see Figure 14). Private schools are not seen as a particularly important or meaningful difference in Norway, as all schools follow the same curriculum and receive approximately similar resources (though private schools charge fees so could attract different student populations). As shown in Figure 23, the quality of estimates for teacher and student characteristics improves slightly when restricting to schools in the original sample (i.e., after dropping participating replacement schools), as the estimate for the percentage of instruction that is special education instructio now falls within the confidence interval of the estimate (though the SMD remains high).

Figure 24 shows the same sample estimates for school-level aggregates of student survey data from the national student survey after removing the replacement schools. Here, there is a noticeable decrease in the quality of population estimates with three population values falling outside of estimated confidence intervals (teaching culture, assessment for learning, and support from home). The average SMD is also over 0.26 suggesting an important risk of bias from differences in student survey scores between sample and population. Further, most student survey scores are higher in the sample than the population, which further suggests an important gap between sample and population.

3.3.1 Summary

To summarize, when the sample is restricted only to originally sampled participating schools (i.e., after dropping participating replacement schools), there is some evidence for bias in the sample (after weighting). Originally sampled participating schools have higher than expected student survey scores, fewer private schools, higher student-teacher ratios and hours of homework help. This would suggest that sampled schools may be, on average, slightly higher functioning than the population as a whole, but the difference becomes

Figure 19: Comparison across national test Scores



very small when including participating replacement schools. However, there remain only 4 (~6%) estimates that did not fall into the confidence interval, which not necessarily higher than would be expected by chance if a small sample were chosen at random.

Figure 20: Comparison across amount of School Resources

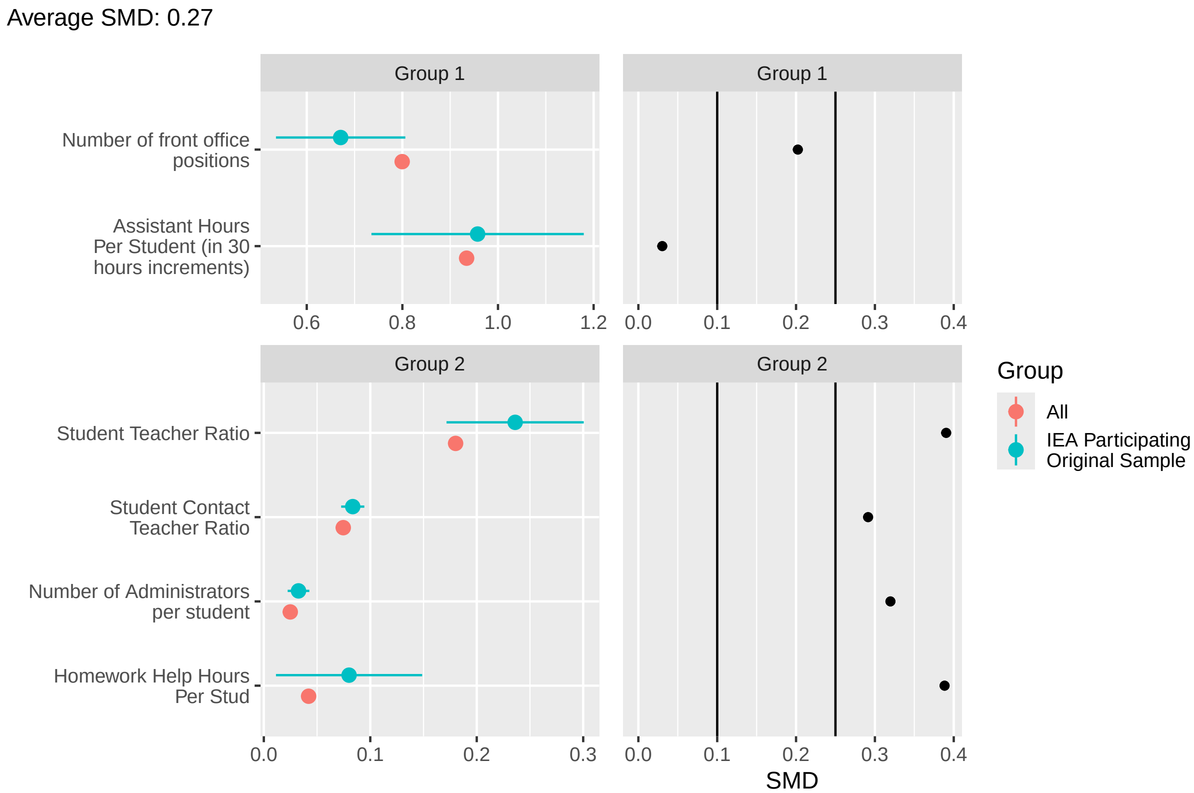


Figure 21: Comparison across school Size

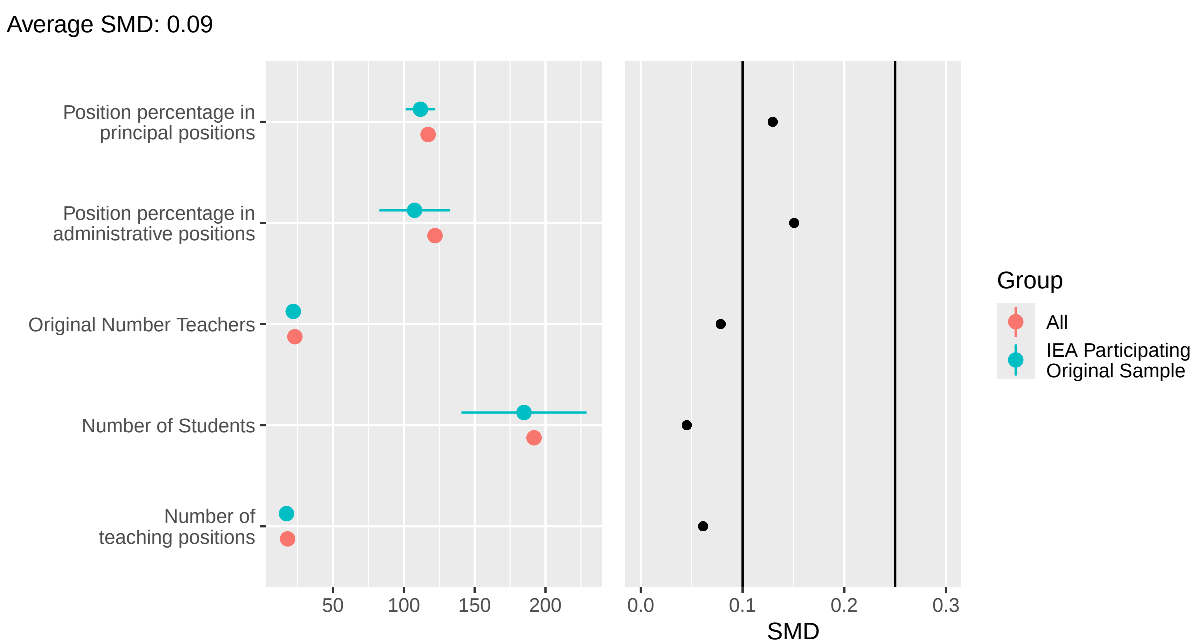
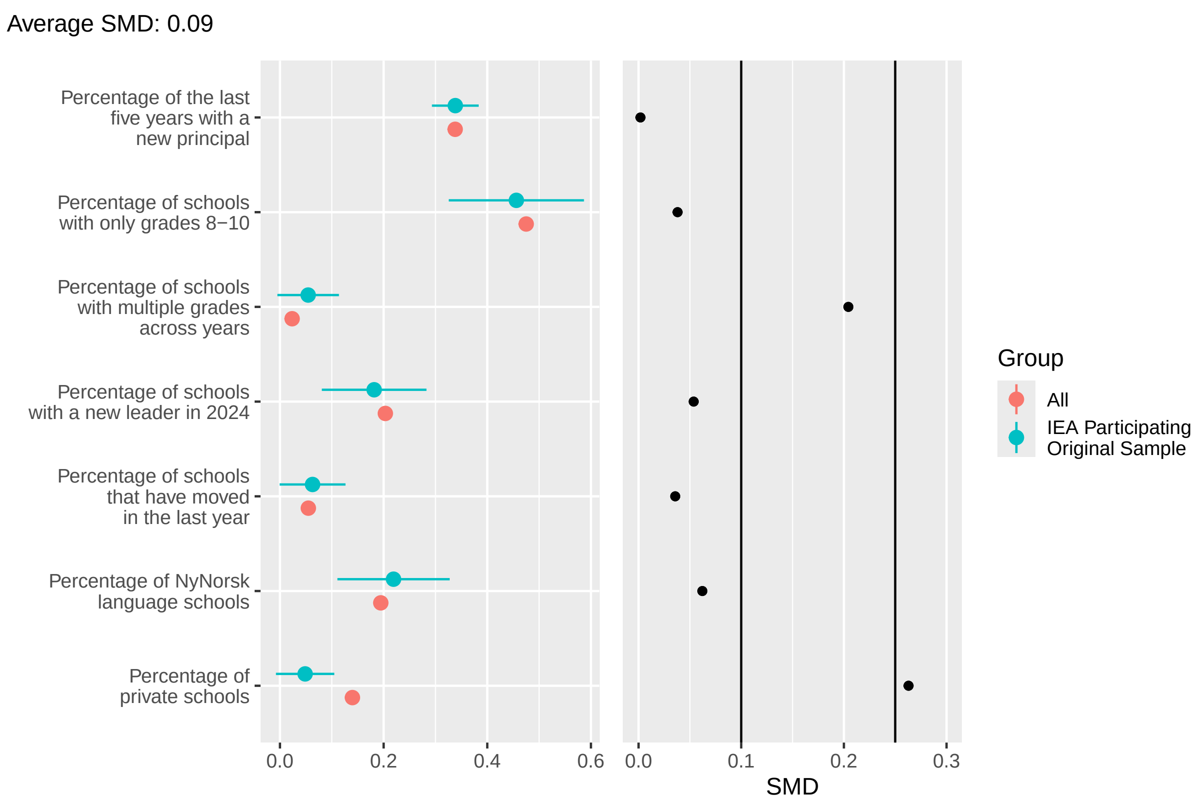


Figure 22: Comparison across characteristics of school



shows the same sample estimates for a set of variables related to instruction, bullying, and student characteristics.

Figure 23: Comparison across teacher and student Characteristics

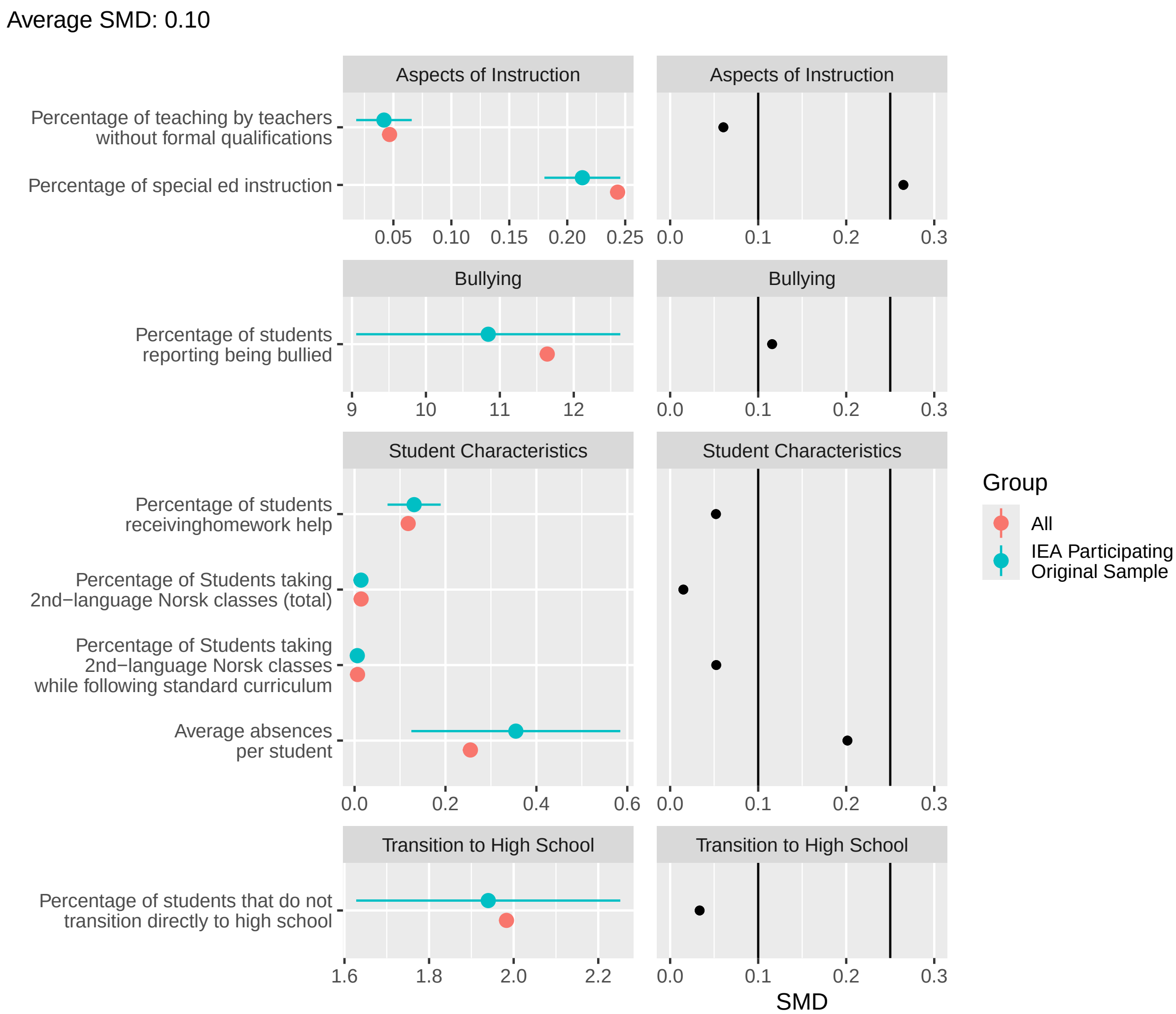
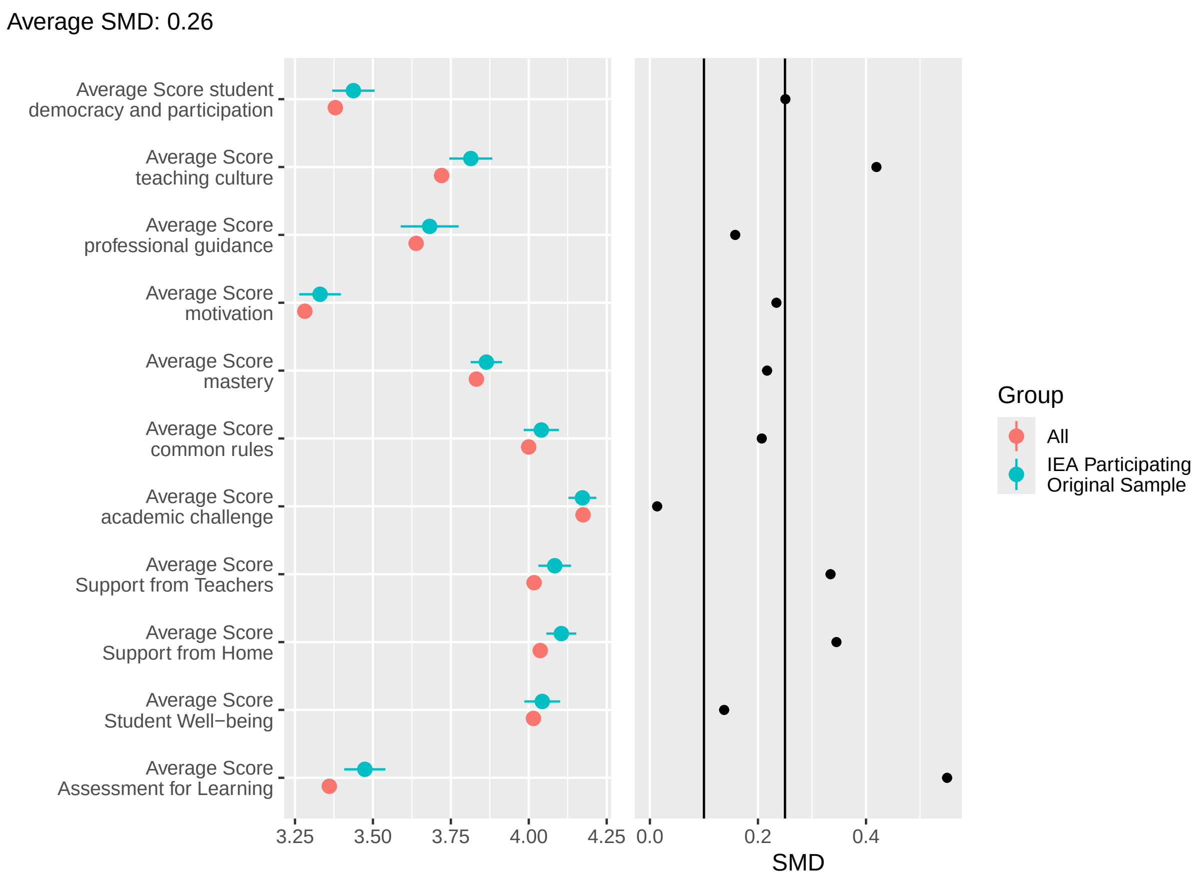


Figure 24: Comparison across Student Survey Scores



3.4 Comparing participating schools to schools that refused to participate

This section compares schools that were *deemed as participating at the teacher level by IEA* to *those schools that were contacted and refused to participate in TALIS*. Here, instead of comparing sample-based estimates, I compare the unweighted mean and standard deviation of the two groups to try and understand if schools that actively refused to participate had different characteristics than those that participated. Figures 25-32 show the results of this analysis. Across all of the variables considered, schools that refused participation were quite similar to those that participated with only one SMD above 0.25. Participating schools were slightly more likely ~10 percentile points to only contains grades 8-10 (i.e., ISCED level 2). About 20% (13/64) variables having SMDs higher than 0.1. This would suggest a strong similarity between participating and refusing schools, implying that refusing schools were not different than participating schools on any characteristics that we could identify.

Figure 25: Comparison of estimates from participating schools to population value for school location (county/fylke)

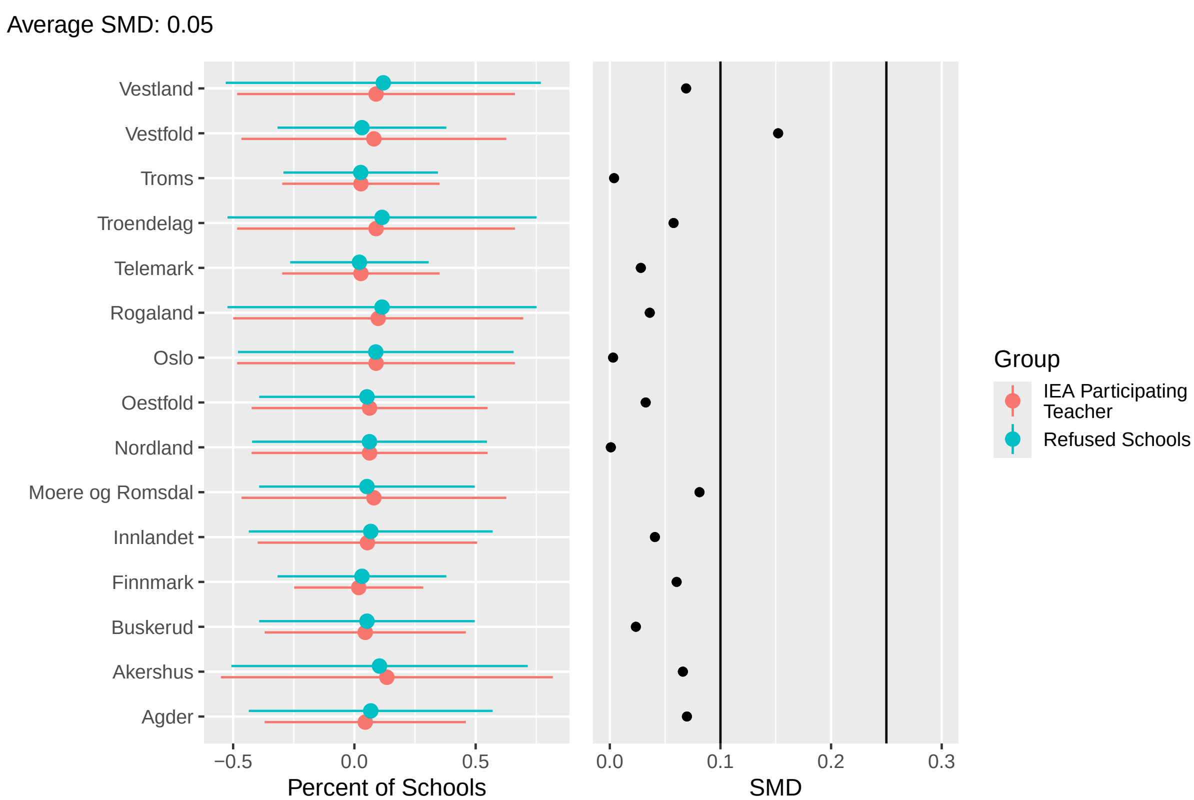


Figure 26: Comparison across per pupil expenses

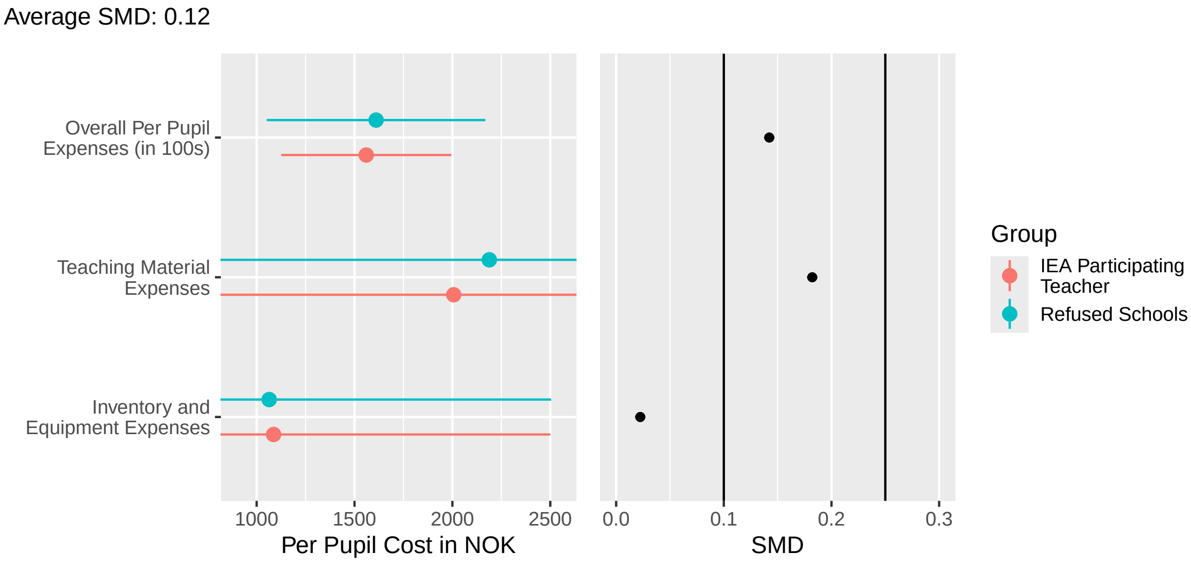


Figure 27: Comparison across national test Scores

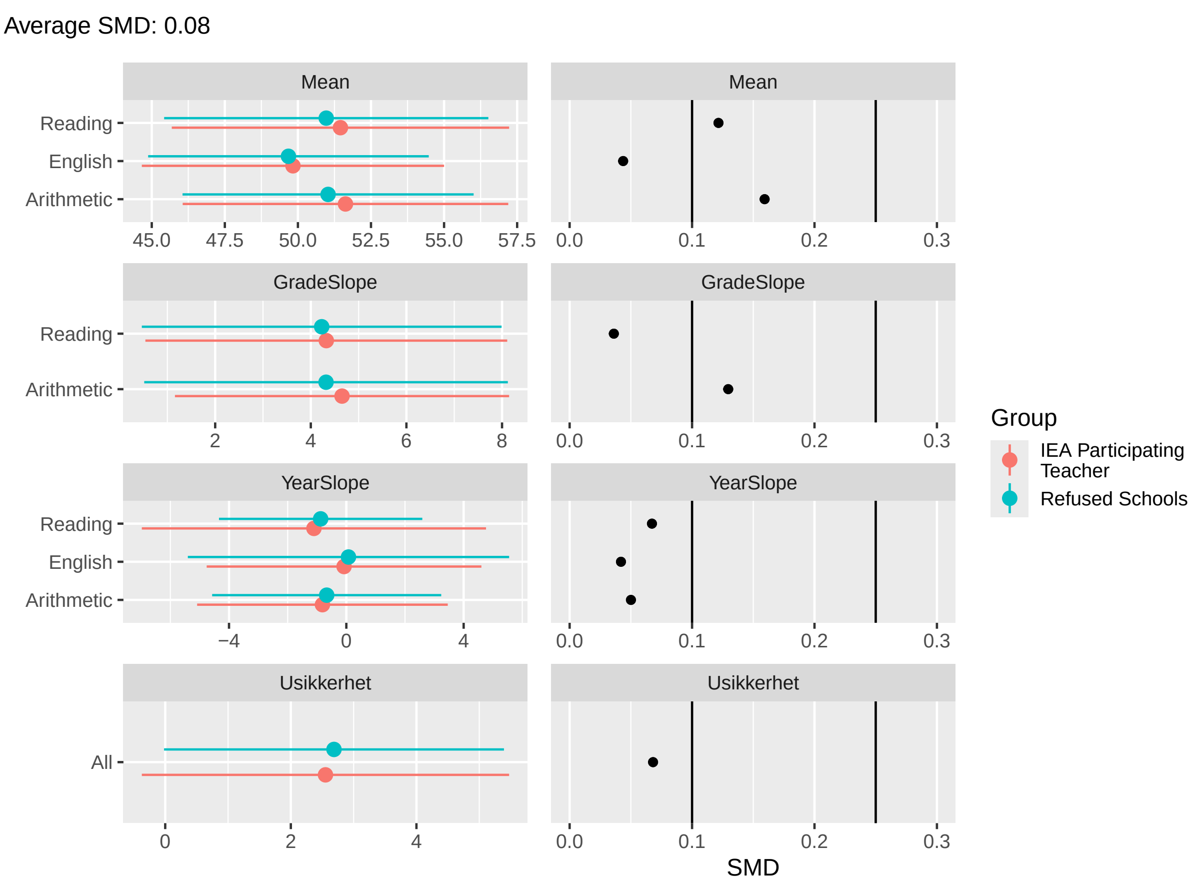


Figure 28: Comparison across amount of School Resources

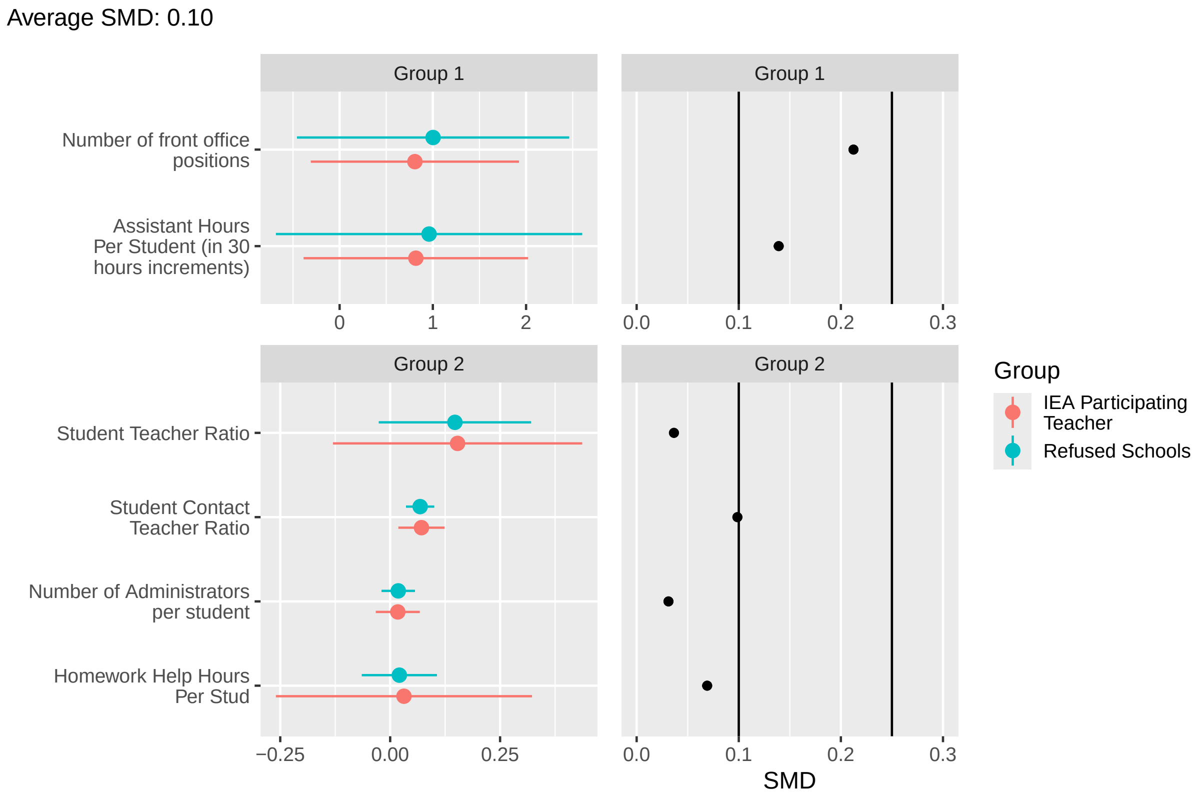


Figure 29: Comparison across school Size

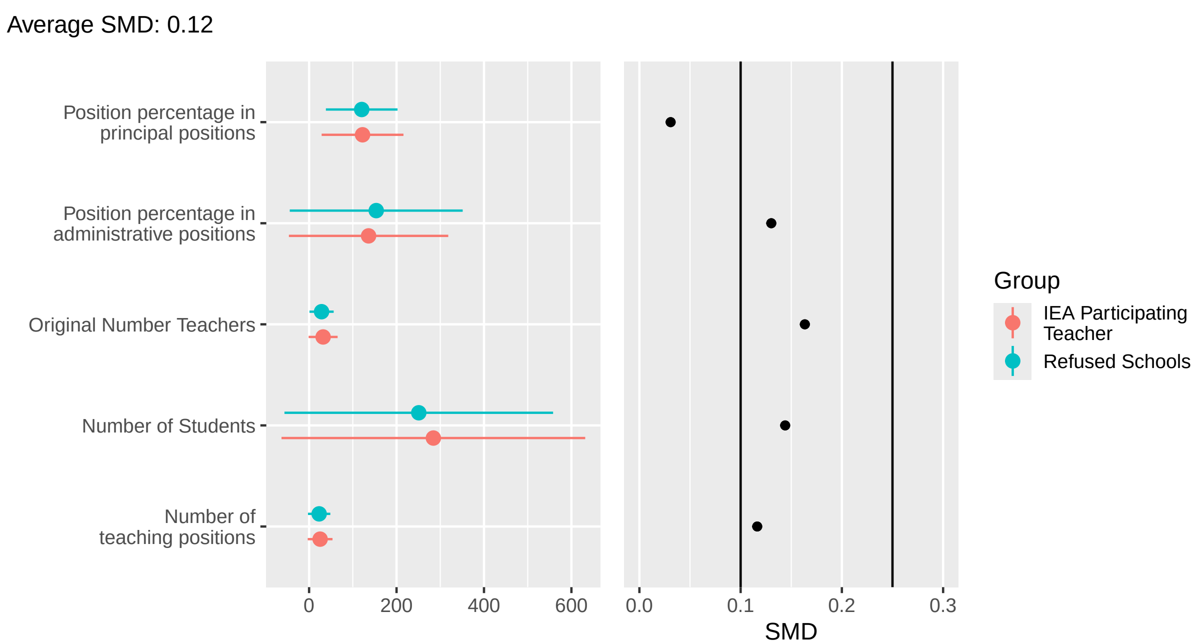
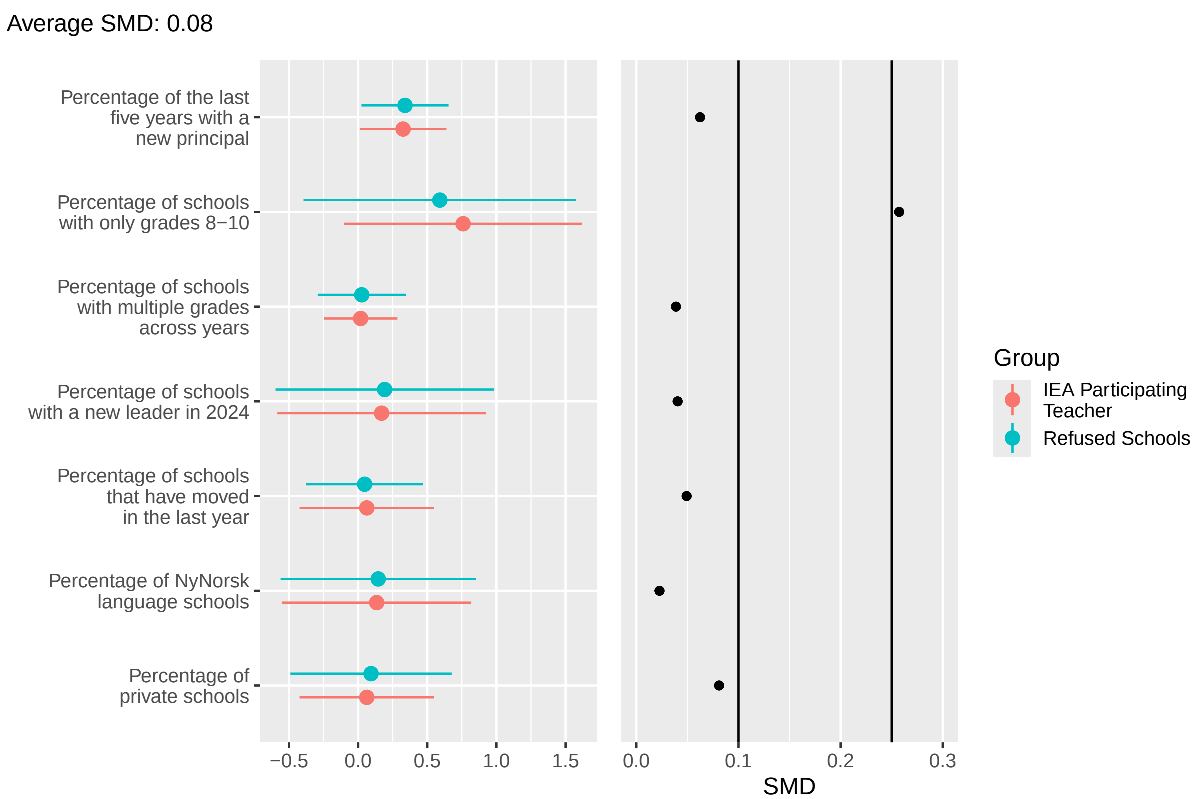


Figure 30: Comparison across characteristics of school



shows the same sample estimates for a set of variables related to instruction, bullying, and student characteristics.

Figure 31: Comparison across teacher and student Characteristics

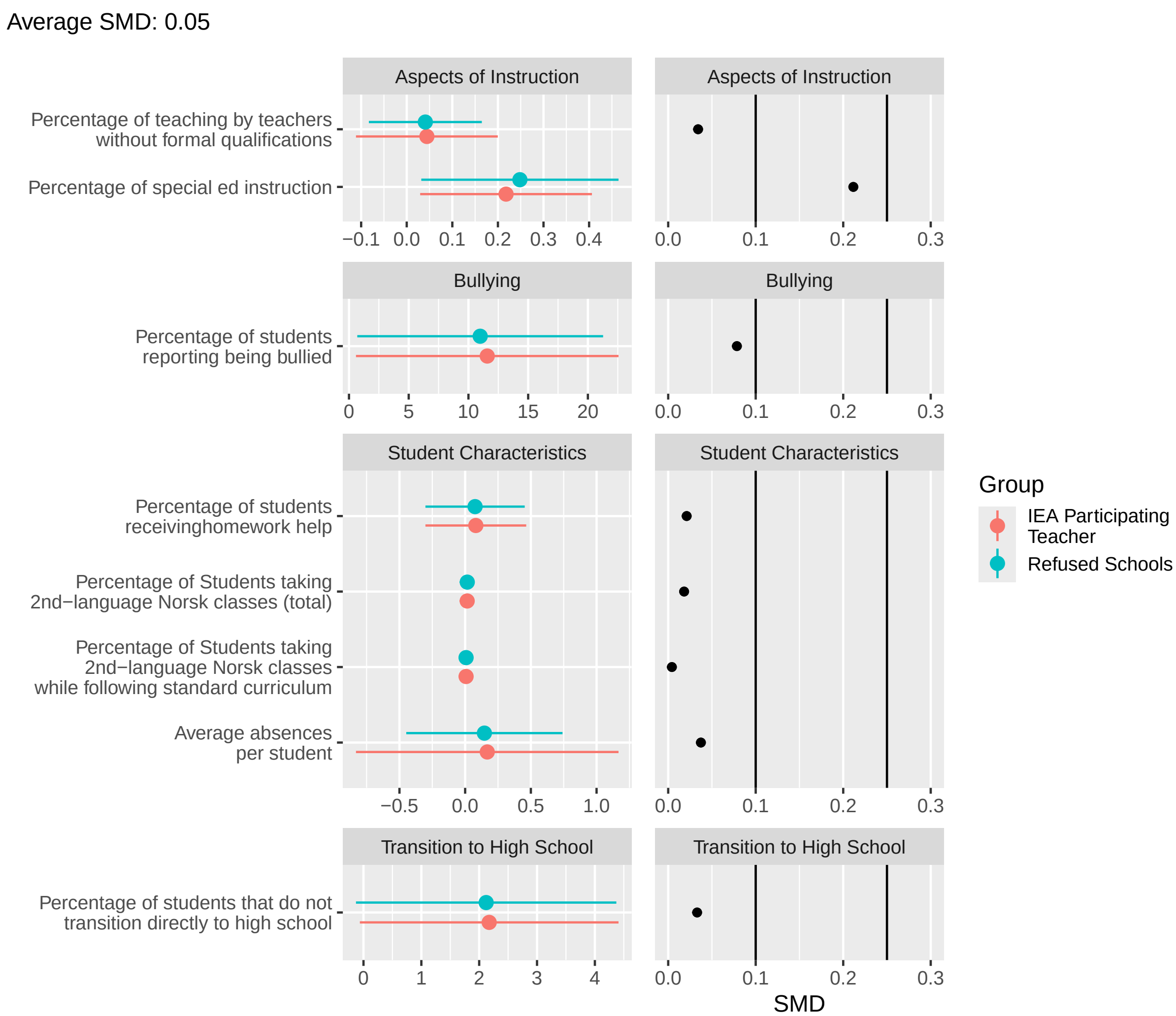
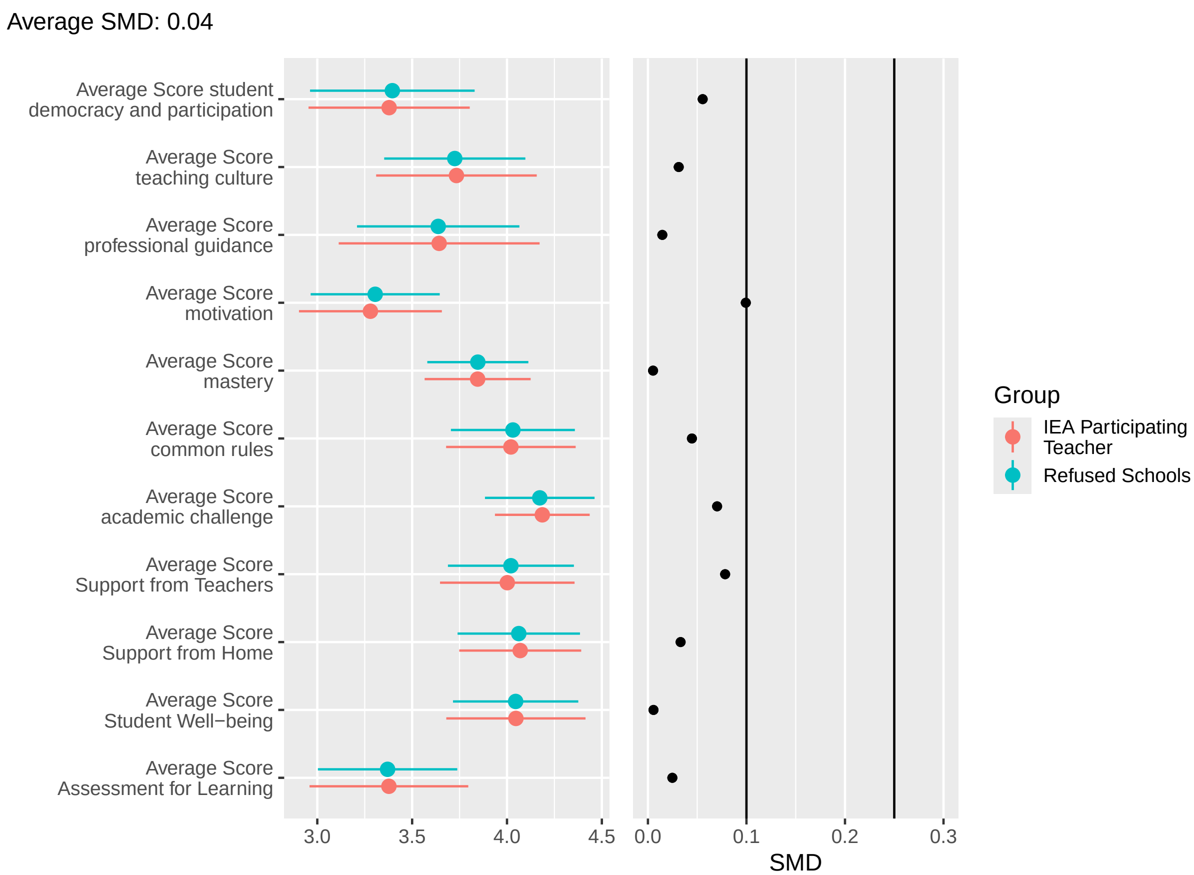


Figure 32: Comparison across Student Survey Scores



3.5 Comparing participating schools from the original sample (without replacement schools) to schools that refused to participate

This section replicates the previous section after restricting the sample to schools in the original sample (i.e., removing replacement schools). That is, it compares schools that were *deemed as participating at the teacher level by IEA and were in the original sample* to *those schools that were contacted and refused to participate in TALIS and were in the original sample*. Here, instead of comparing sample-based estimates, I compare the unweighted mean and standard deviation of the two groups to try and understand if schools that actively refused to participate had different characteristics than those that participated. Figures 33-40 show the results of this analysis. The same basic patterns as in the previous section apply, but there is a lot more variation in estimates of SMD, as would be expected when using a smaller sample. As a result, the standardized mean differences rarely move above the 0.25 threshold, but are often above the 0.1 threshold. Differences remain fairly small from a non-standardized perspective. For example, mean differences in the school-average arithmetic score from the national test are ~0.28 SMD, but are about 1 scale point (where the original [student-level] standard deviation is 10 scale points). SMD averages for the categories of variables range from 0.8-0.13, suggesting that overall differences are still quite modest.

Figure 33: Comparison of estimates from participating schools to population value for school location (county/fylke)

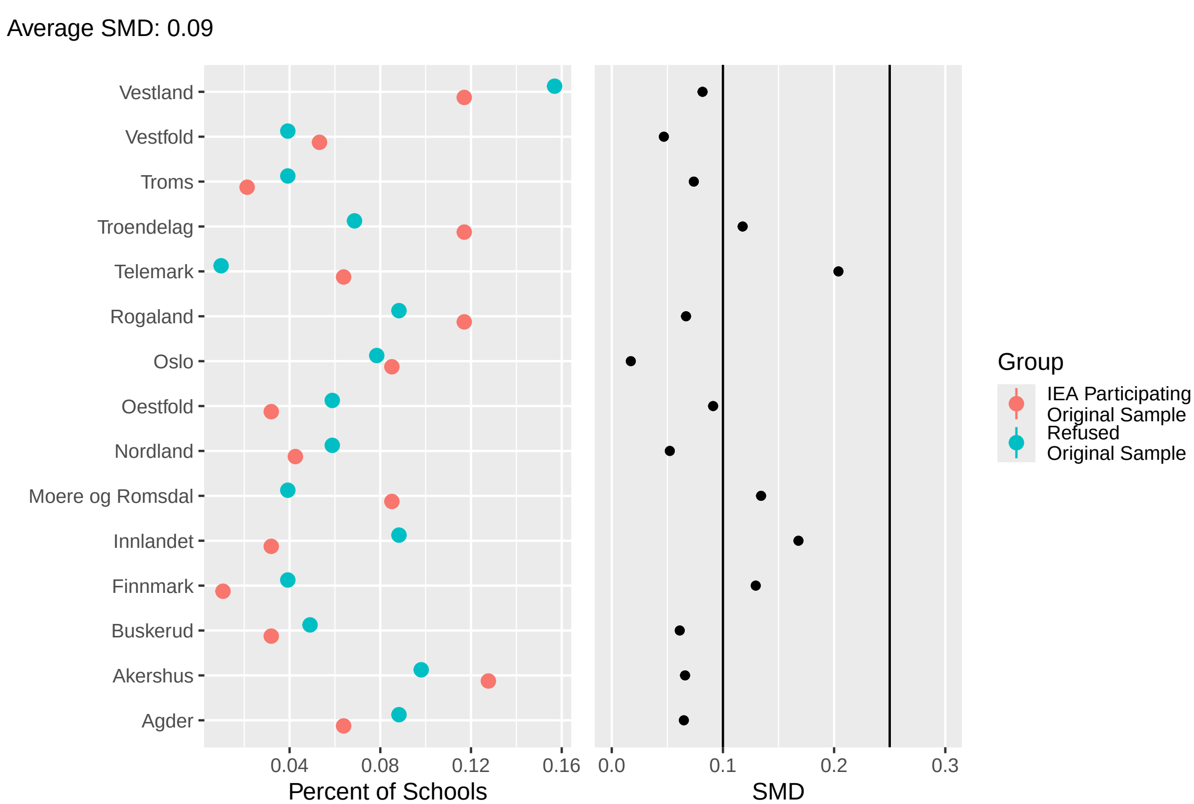


Figure 34: Comparison across per pupil expenses

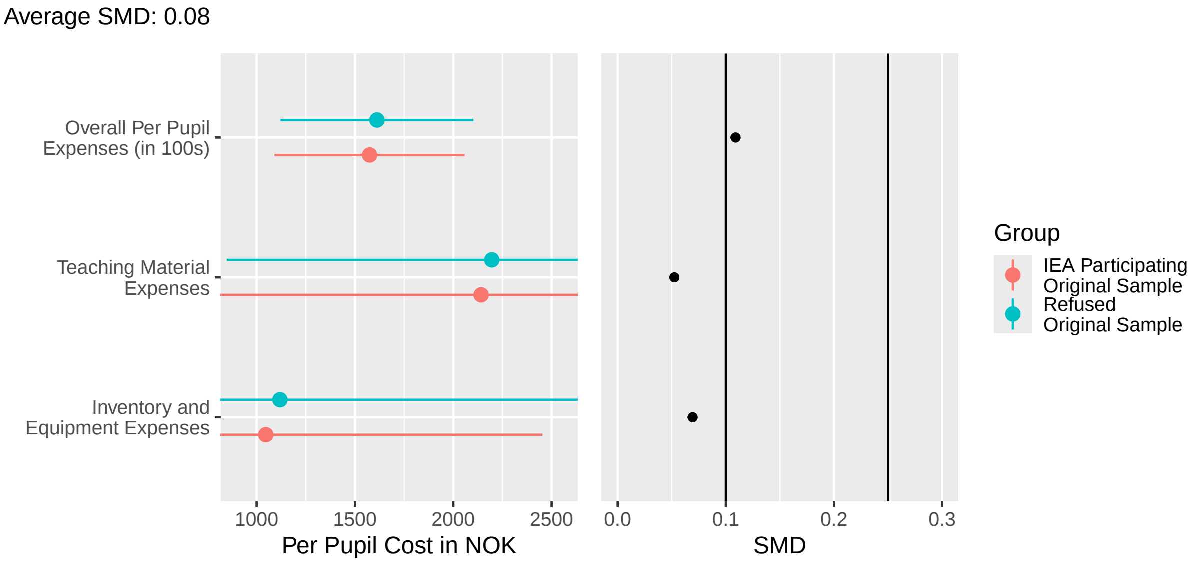


Figure 35: Comparison across national test Scores

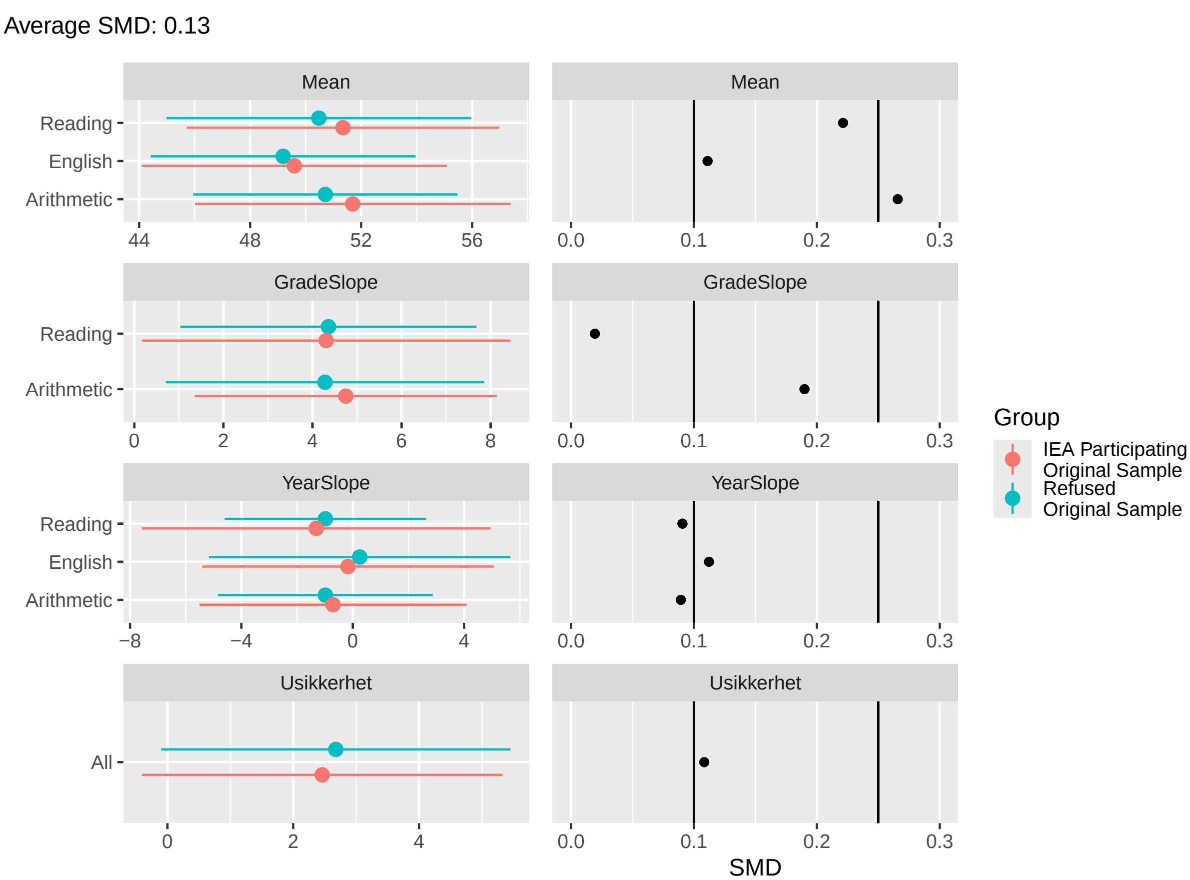


Figure 36: Comparison across amount of School Resources

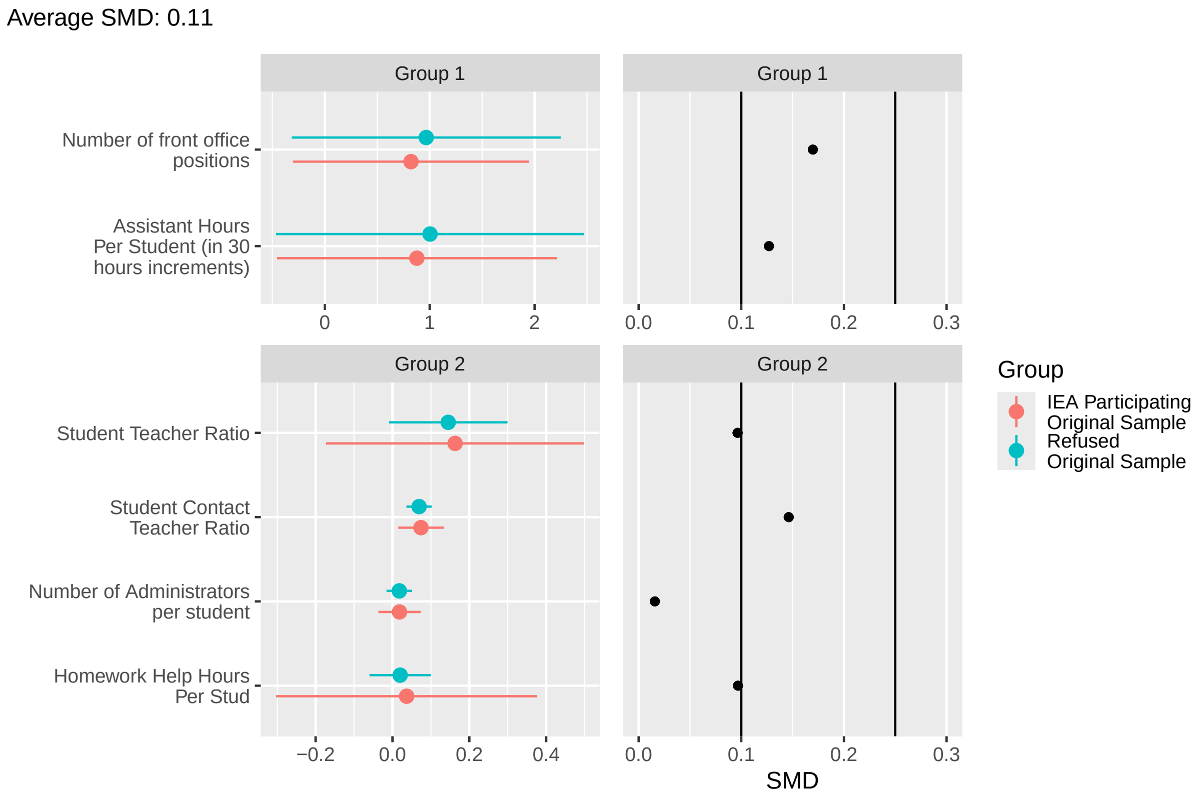


Figure 37: Comparison across school Size

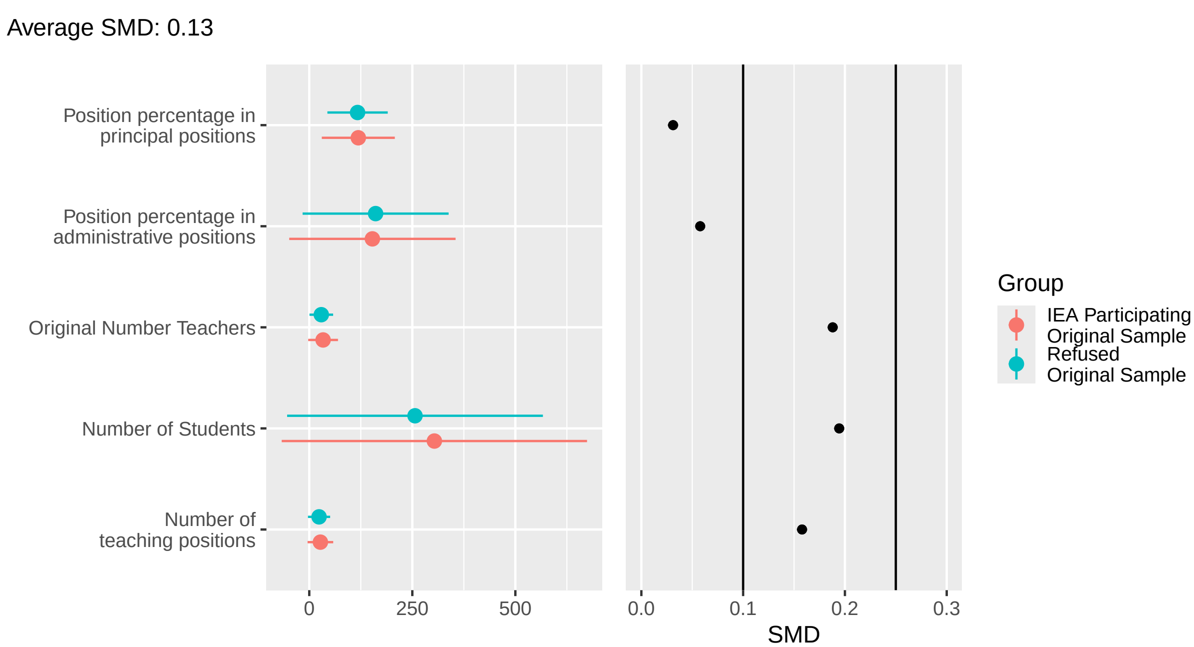
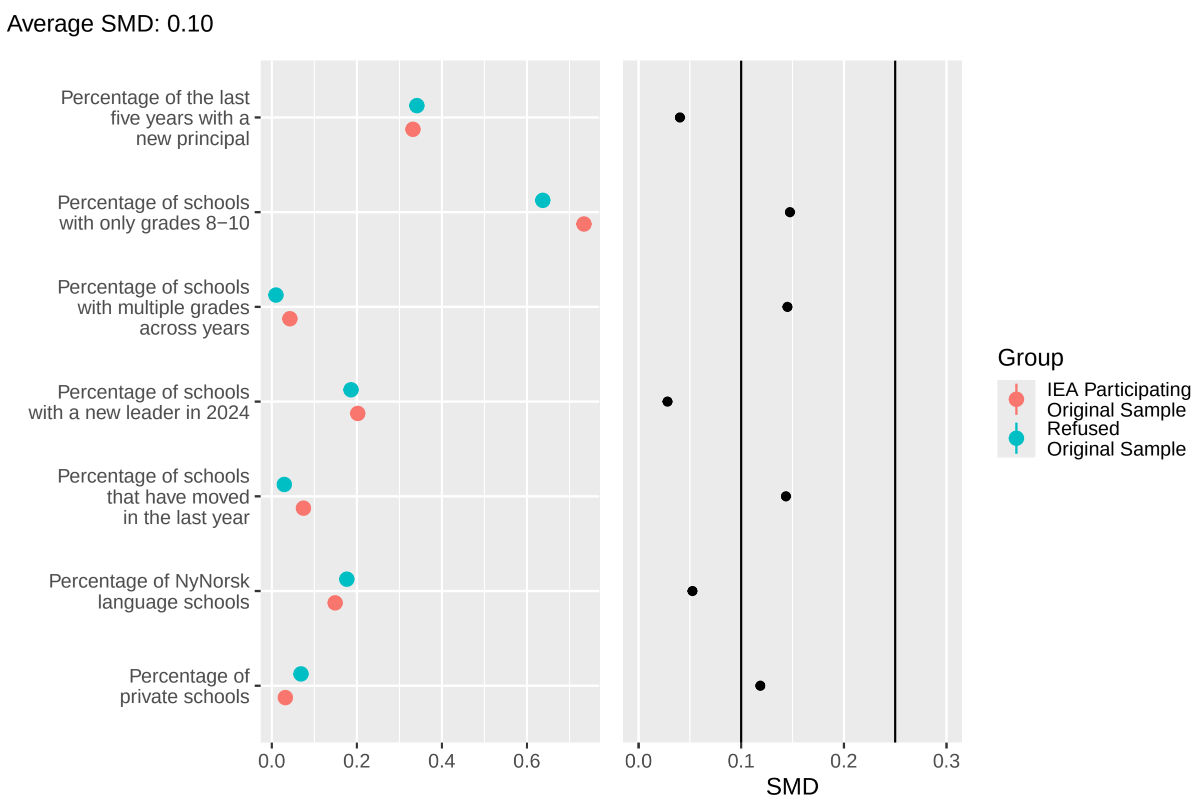


Figure 38: Comparison across characteristics of school



shows the same sample estimates for a set of variables related to instruction, bullying, and student characteristics.

Figure 39: Comparison across teacher and student Characteristics

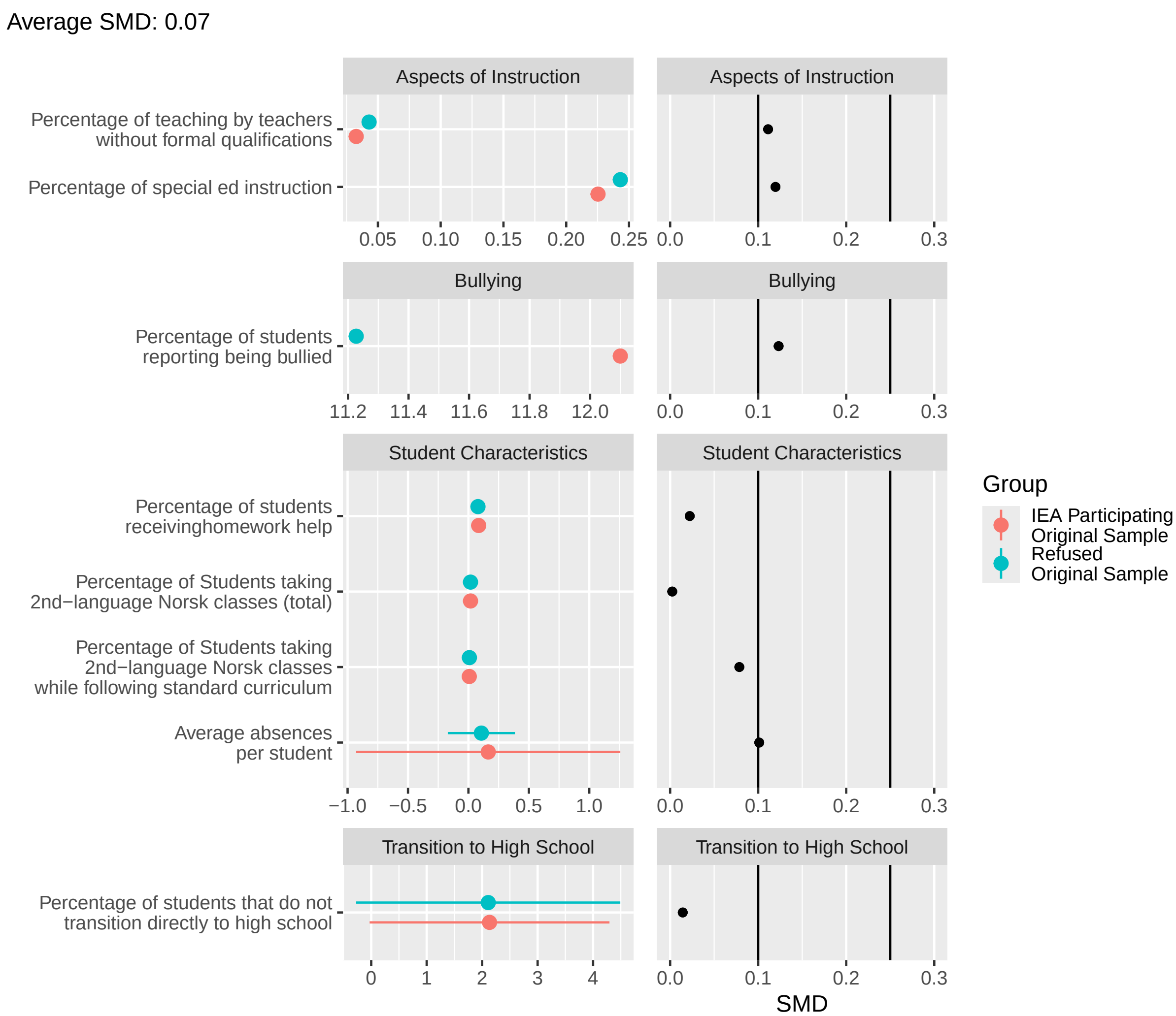
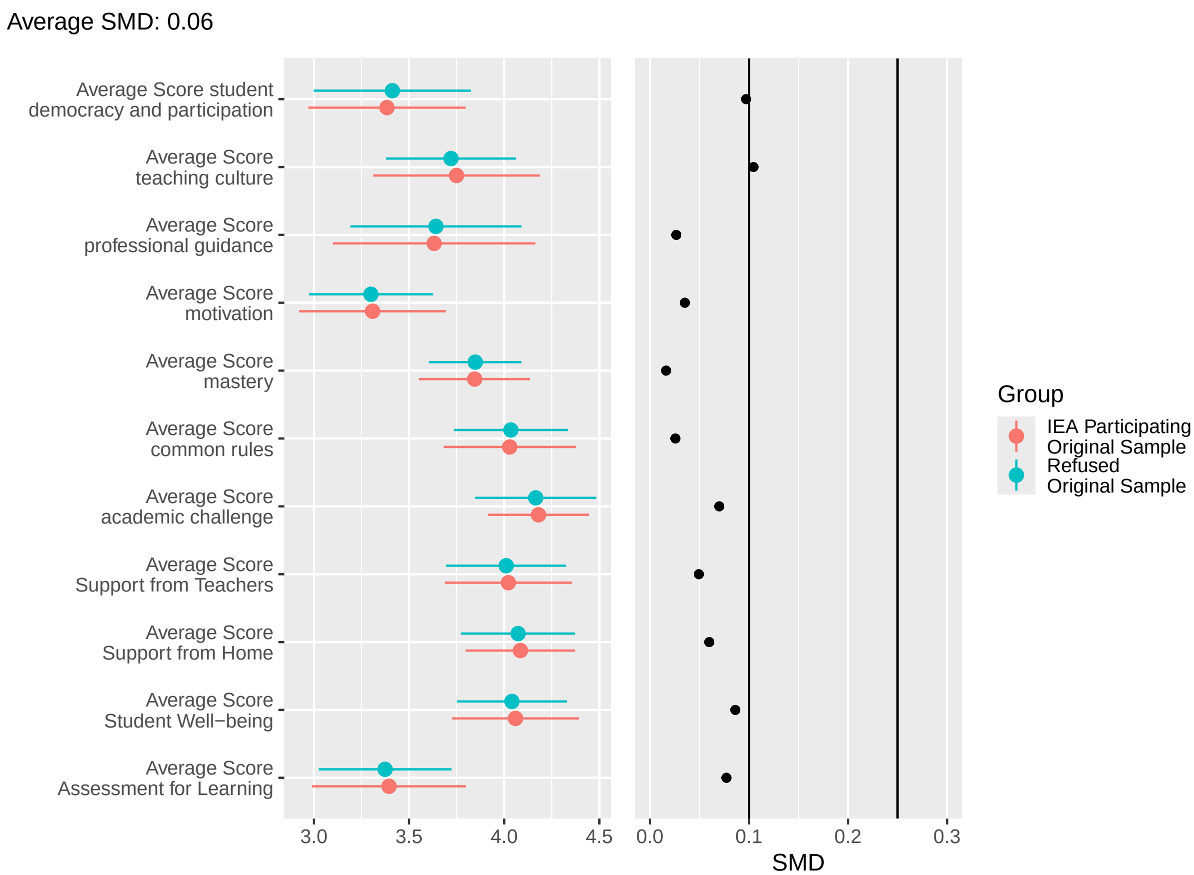


Figure 40: Comparison across Student Survey Scores



3.6 Comparing participating schools to schools with low participation

This section compares those schools that were deemed as participating by IEA to those schools with less than 50% participation raters among teachers. Figures 41-48 shows these results, focusing on mean and standard deviations for variables. Note that standard deviations are often hidden when showing percentages since they do not add information. Three SMDs are over the 0.25 threshold and 23 are between 0.1 and 0.25. Average SMDs however remain near or beow 0.1. The largest apparent differences come from school location, where differences in the percentage of schools in the two groups within a county are often about 5 percentile points (mean SMD 0.12). Further, low-participation schools have about 10 percentile points higher chance of having a new principal, 15 percentile points lower chance of having only grades 8-10, more administrative positions but are not larger overall, and a greater percentage of special educational instruction. Across many variables, though, the two sets of schools are quite similar. In summary, it would look like there are some differences, but also remarkable level of similarity across groups.

Figure 41: Comparison of estimates from participating schools to population value for school location (county/fylke)

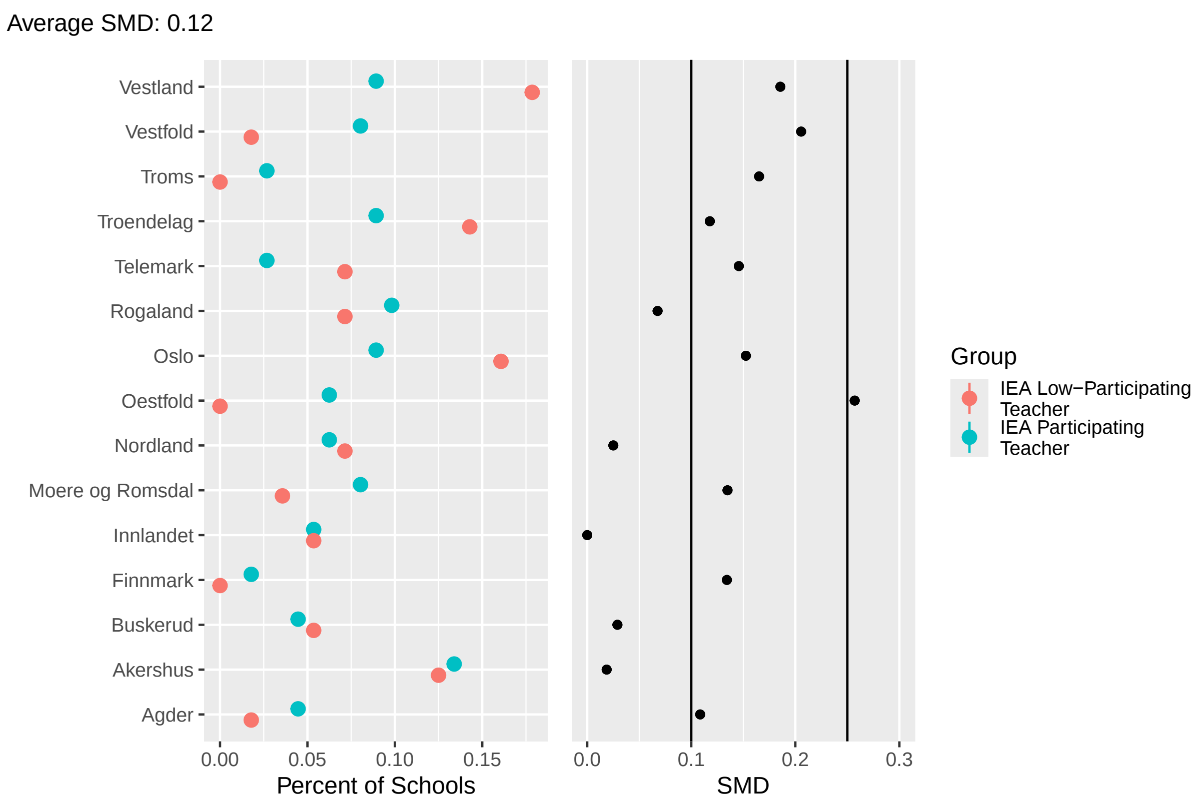


Figure 42: Comparison across per pupil expenses

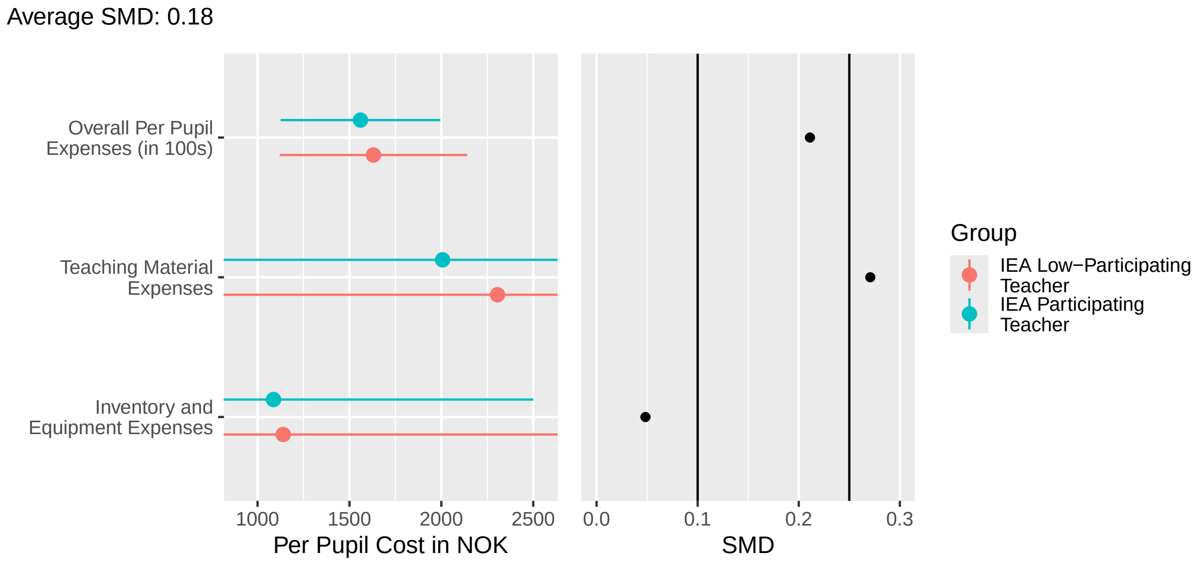


Figure 43: Comparison across national test Scores

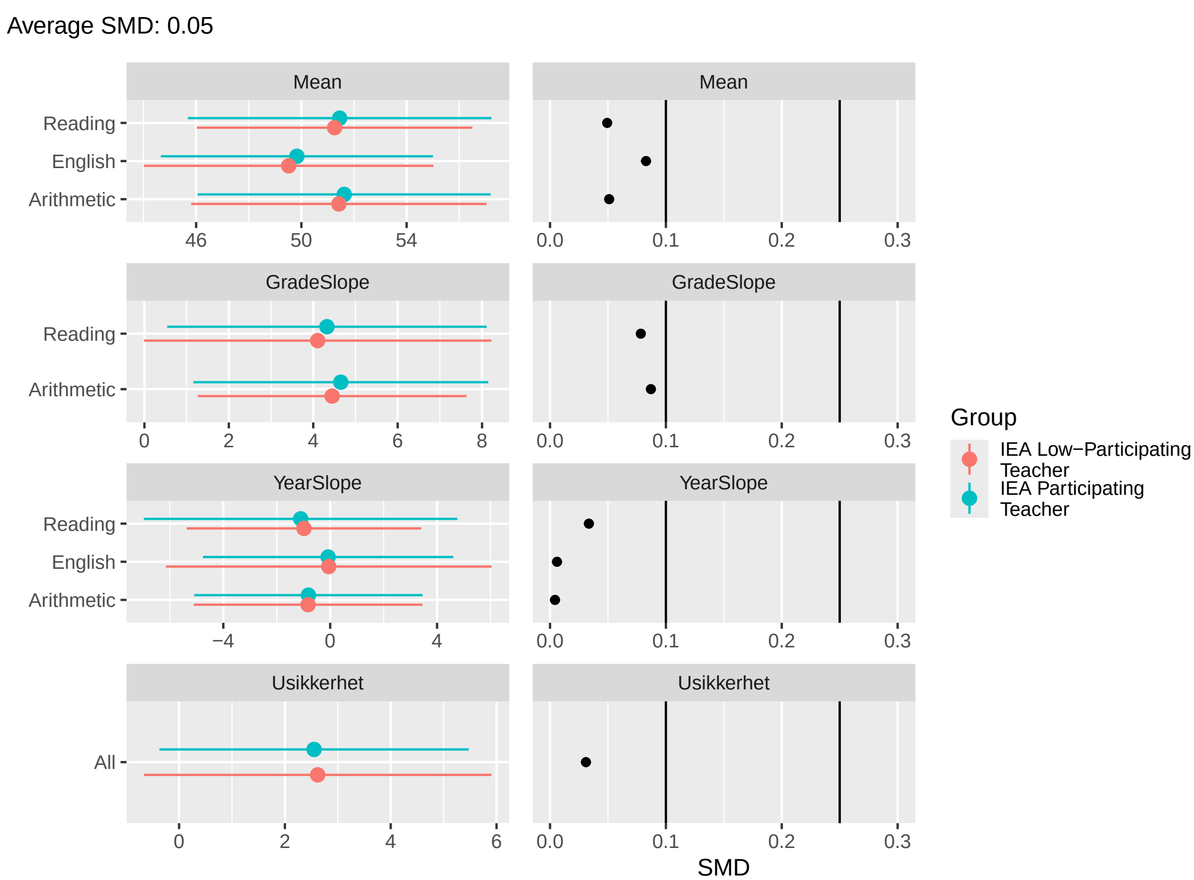


Figure 44: Comparison across amount of School Resources

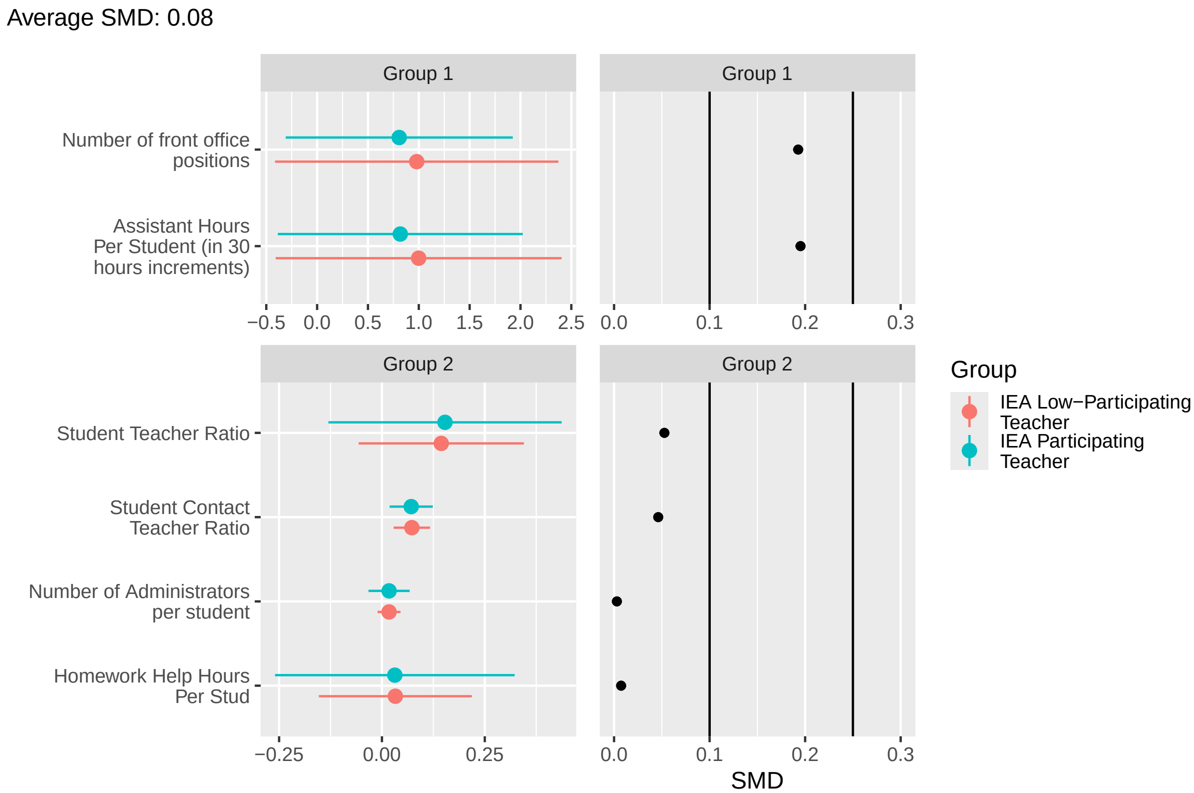


Figure 45: Comparison across school Size

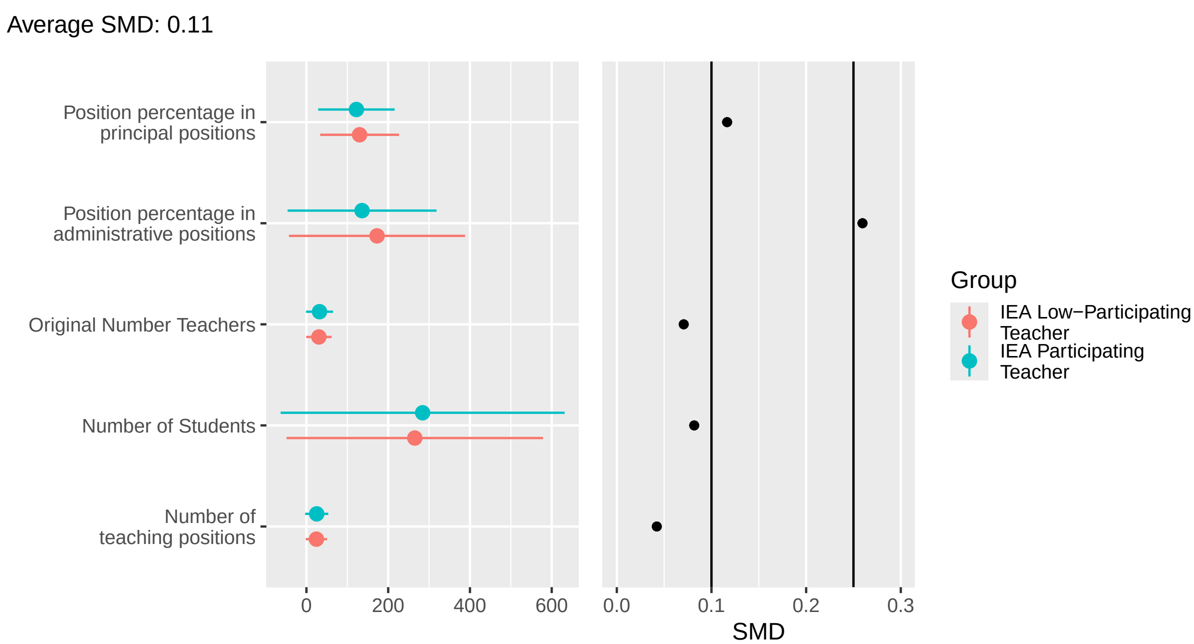
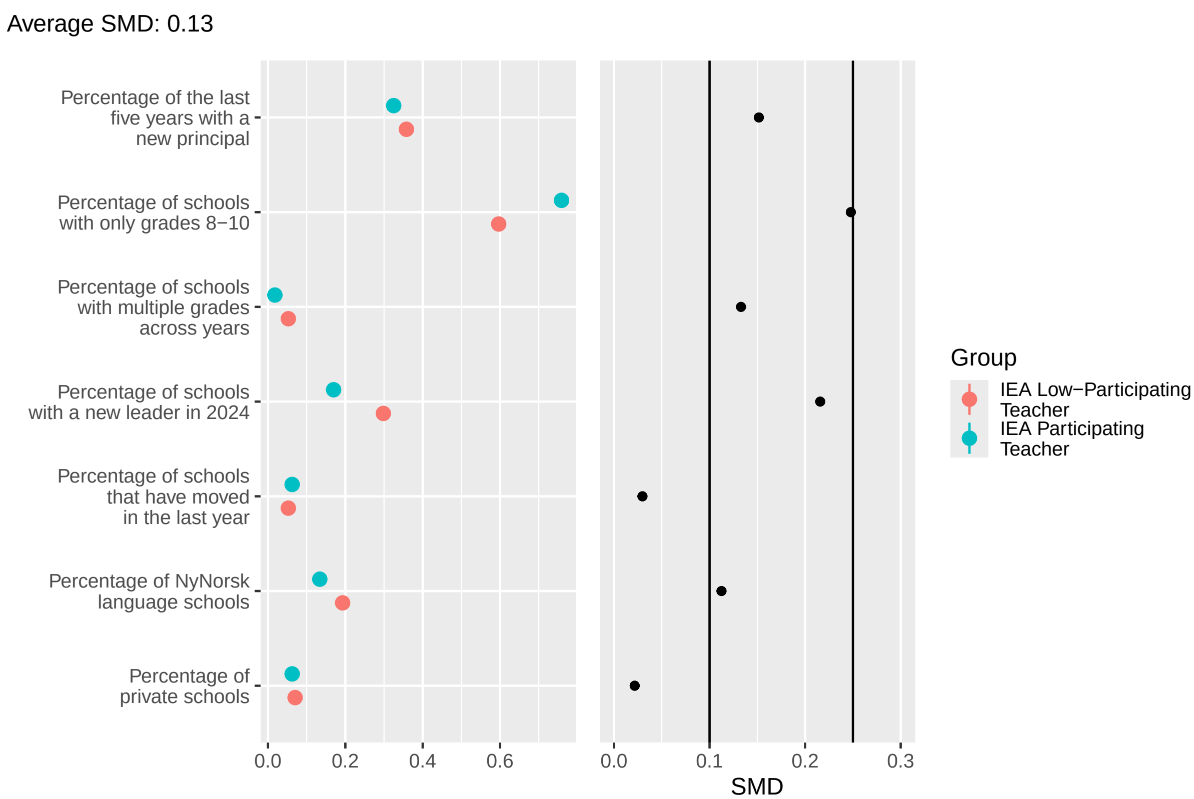


Figure 46: Comparison across characteristics of school



shows the same sample estimates for a set of variables related to instruction, bullying, and student characteristics.

Figure 47: Comparison across teacher and student Characteristics

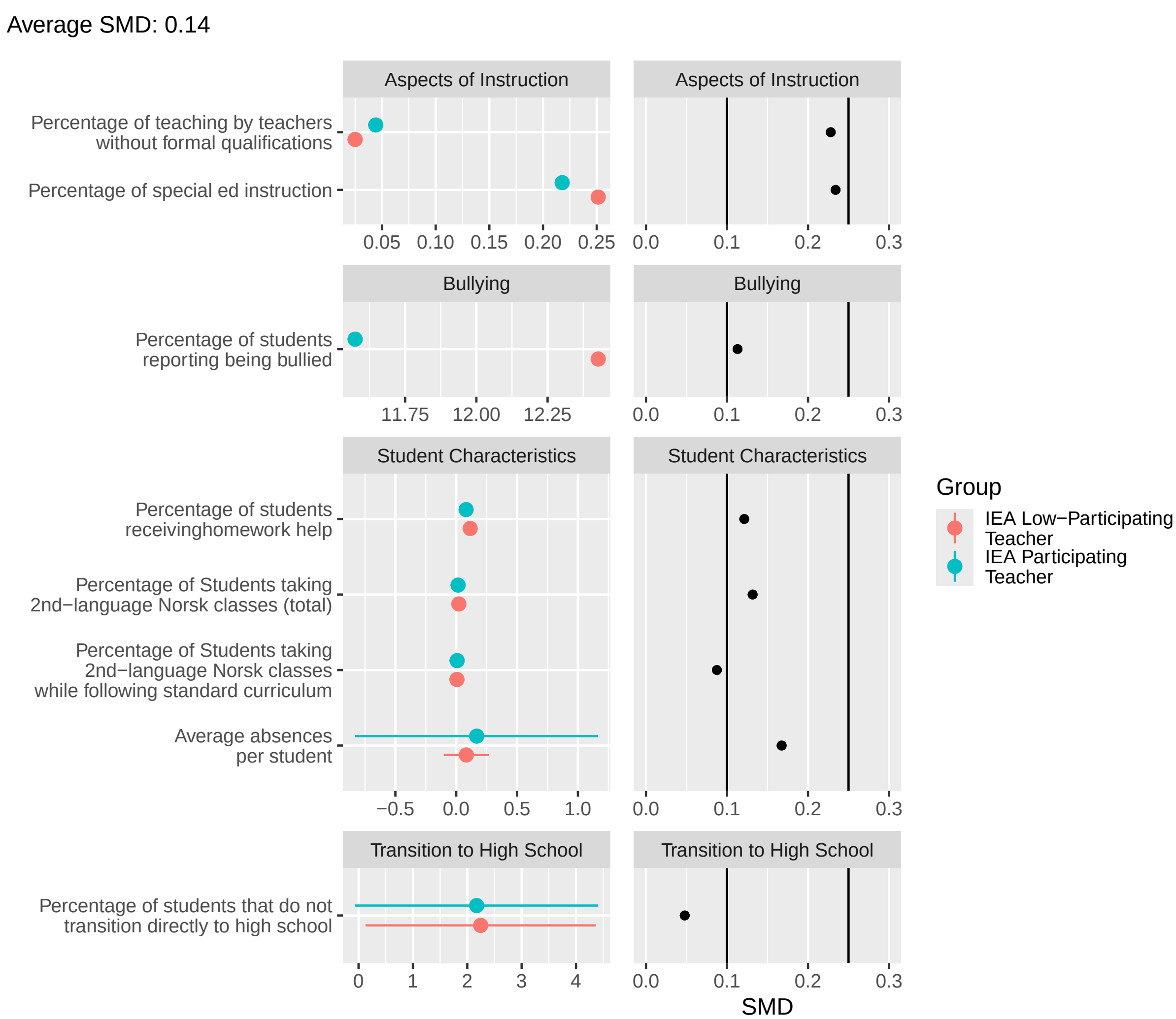
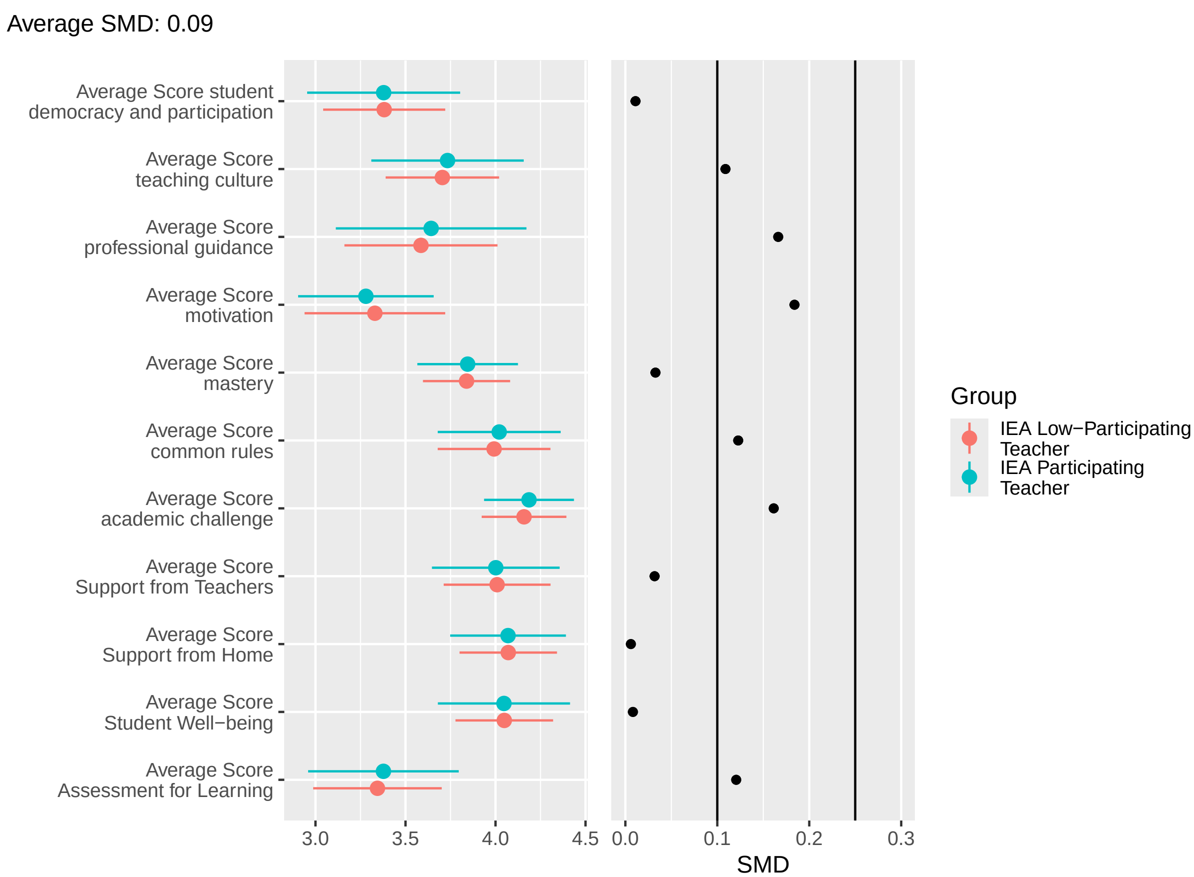


Figure 48: Comparison across Student Survey Scores



4 Differences between responding and non-responding teachers

In this section, I move towards examining differences between responding and non-responding teachers. Unfortunately, there is not nearly as much teacher-level data in Norway as school-level data. As a result, I was not able to get any estimates for population level means for important TALIS variables. Due to this, this section is restricted to comparing the variables that were collected on the teacher listing form, comparing responders and non-responders using unweighted means based on these variables. Figure 49 compares teachers that returned the survey to those that did not return the survey. The figure shows that differences in main subject and gender are less than 5 percentile points, which is a SMD <= 0.1. The differences in teacher age are close to zero. While there is little data at the teacher level, the data that does exist would suggest that there are relatively small differences between responding teachers and non-responding teachers.

Figure 49: Comparison of teacher characteristics across responders, non-responders, and non-sampled teachers

5 Relationship between non-response and survey outcomes

This section moves to trying to predict non-response more comprehensively, rather than looking at individual variables. Here, I use the `glmnet` package to run an elastic net regression to predict the probability of being a participating school. This regression approach allows for multicollinear predictors and shrinks regression coefficients towards zero based on their predictive power (i.e., less predictive coefficients are shrunken). I used cross-validation to select model parameters. The model started with all variables discussed previously as predictors. The result was a regression model with only 10 non-zero predictors (of 64 initial predictors): an indicator of whether the school was private, the number of students at the school, the percentage of teachers that fulfill competency rules based on amount of teacher education, total amount of teaching positions at the school, percentage of school instruction dedicated to special education, uncertainty in national test scores, per pupil expenditures, per pupil spending on teaching materials, an indicator for whether the school only included ISCED level 2 grades (or both ISCED levels 1 and 2), and an indicator capturing whether the school was in the Vestfold fylke. These variables, then, appear to predict participation in TALIS.

Figure 50 shows the distribution of the propensity score in participating schools and non-participating schools. The distribution for participating schools after weighting is also shown. Of note is that the propensity score distribution varies across groups, as would be expected. If the sample were purely random, the propensity score distribution would look similar across groups. Note also that there is not large regions where the propensity score distribution is zero in one group and non-zero in other groups. This indicates that the propensity score overlap, which is a positive sign, as it suggests that there exist some participating schools can represent each non-participating school. Last, note that the weighted distribution of the propensity score is much closer to the distribution of the propensity score in the non-participating schools, but the distributions are not identical. If the weights fully corrected for non-response, the distribution of the propensity score in the weighted and non-participating groups would be equivalent. Then, the IEA weights do not seem to fully adjust for non-response, which is expected since we include far more variables in the propensity score than IEA used to create the weights.

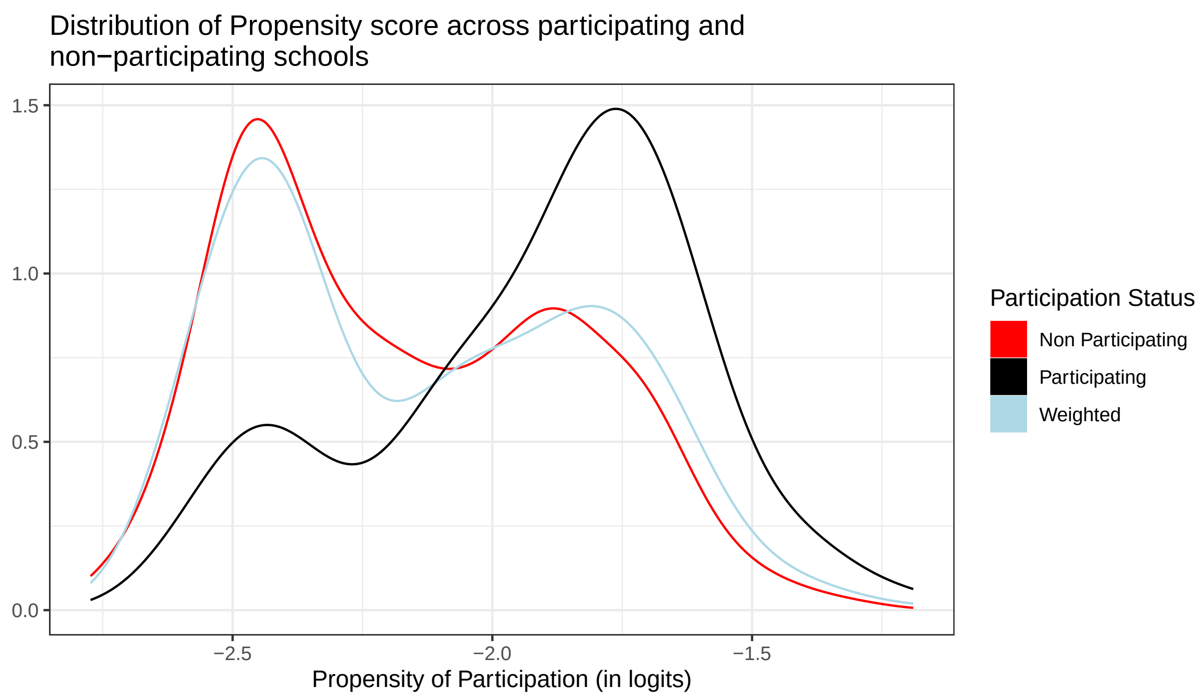


Figure 50: Distribution of propensity score across groups

Table 5: Correlations between propensity score and rektor variables

Scale	AveCorrW	AveCorrP	Diff
TCA04	0.0903	0.0825	0.0439
TCA08	0.0579	0.0900	0.0526
TCA09	0.1347	0.1111	0.0236
TCA10	0.0777	0.0823	0.0197
TCB02	0.2177	0.1964	0.0213
TCB03	0.3059	0.2438	0.0621
TCB04	0.2525	0.2143	0.0381
TCB05	0.2384	0.2484	0.0276
TCB06	0.2731	0.2326	0.0435
TCB07	0.6019	0.6024	0.0005
TCB08	0.1200	0.1195	0.0362

Table 5: Correlations between propensity score and rektor variables *(continued)*

Scale	AveCorrW	AveCorrP	Diff
TCB09	0.0715	0.0412	0.0442
TCB10	0.0555	0.0473	0.0131
TCB11	0.2309	0.1514	0.0795
TCC01	0.3763	0.2912	0.0851
TCC02	0.2056	0.1792	0.0657
TCC03	0.0997	0.0883	0.0363
TCC04	0.1105	0.1078	0.0465
TCC05	0.0643	0.0580	0.0300
TCC06	0.0747	0.0661	0.0447
TCC07	0.1104	0.1256	0.0296
TCC08	0.1168	0.0846	0.0417
TCD01	0.1254	0.1194	0.0070
TCD02	0.0917	0.0783	0.0520
TCD03	0.0525	0.0326	0.0851
TCD04	0.0857	0.0104	0.0752
TCD05	0.0522	0.0540	0.0397
TCD06	0.1001	0.0946	0.0436
TCD07	0.1007	0.0777	0.0543
TCE01	0.0595	0.0759	0.0847
TCE02	0.1321	0.1713	0.0457
TCE03	0.0964	0.1164	0.0372
TCE04	0.2431	0.2107	0.0433
TCF01	0.1090	0.0918	0.0172
TCF02	0.1077	0.0799	0.0451
TCF03	0.0974	0.0884	0.0296
TCG01	0.0510	0.0312	0.0352
TCG02	0.0712	0.0800	0.0386
TCG03	0.1762	0.1715	0.0047
TCH01	0.0405	0.0178	0.0582
TCH02	0.0649	0.0471	0.0203
TCH03	0.0647	0.0667	0.0297
TCH04	0.0838	0.0853	0.0250
TCH05	0.0985	0.0716	0.0383

Last, I look at the correlation between survey item responses and the propensity score and weights. The logic behind this analysis is as follows. The propensity score contains information from school's participation that is not contained in the weights. Thus, from the perspective of Hartman and Huang (2024b), the propensity score can be understood as an *estimate* of the "ideal weights". If the correlation between the propensity score and responses is different than the correlation between the weights and responses, this would imply that "error" in the weights is correlated with responses, leading to biases in the weighted estimates (see Hartman & Huang, 2024b). This should be understood as a sensitivity analysis, rather than a complete test of bias, since bias could result from variables not included in the propensity score. Table 5 shows the correlations between survey items (aggregated across sub-items and only for items where numeric answers are given) and the weights and propensity score. AveCorrW is the average correlation between weights and responses and AveCorrP is the average correlation between the propensity score and responses. Figure 51 show the distribution of the difference between the correlations. As can be seen in Figure 51, differences in

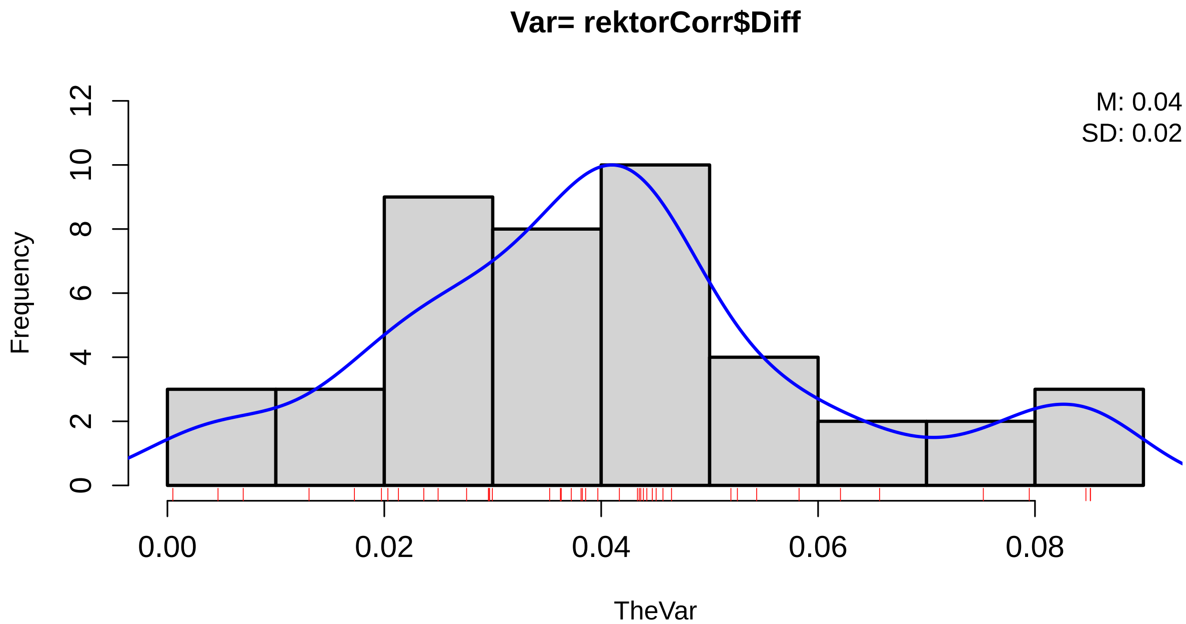


Figure 51: Distribution of correlations between Rektor items and propensity score

correlations are all below 0.1, which would suggest that all of the school level variables that we explored do not lead to changes in weights that would imply large biases in estimates from the principal survey. This is the strongest evidence we have yet that the sampling challenges in TALIS Core 2024 Norway do not have a strong biasing effect from information from the principal survey. That said, the differences in correlations are not zero, suggesting that future analyses will need to carefully assess if small biases exist.

Table 6: Correlations between propensity score and laerer variables

Scale	AveCorrW	AveCorrP	Diff
TTA01	0.0394	0.0004	0.0398
TTA02	0.0315	0.0278	0.0593
TTA05	0.0130	0.0237	0.0367
TTA06	0.0260	0.0259	0.0264
TTA07	0.0387	0.0201	0.0339
TTA08	0.0193	0.0240	0.0047
TTB02	0.1484	0.0209	0.1274
TTB03	0.3716	0.0003	0.3719
TTB04	0.0761	0.0290	0.0471
TTB05	0.0156	0.0298	0.0415
TTB06	0.0121	0.0134	0.0254
TTB07	0.0343	0.0121	0.0223
TTB08	0.0362	0.0421	0.0261
TTC01	0.0214	0.0210	0.0424
TTC02	0.0474	0.0313	0.0477
TTC03	0.0407	0.0212	0.0488
TTC05	0.0254	0.0244	0.0348

Table 6: Correlations between propensity score and laerer variables *(continued)*

Scale	AveCorrW	AveCorrP	Diff
TTC06	0.0473	0.0033	0.0440
TTC07	0.0180	0.0339	0.0302
TTC08	0.0221	0.0615	0.0473
TTC09	0.0176	0.0566	0.0628
TTD01	0.0489	0.0701	0.0372
TTD02	0.0189	0.0306	0.0319
TTD03	0.0159	0.0620	0.0673
TTD04	0.0146	0.0229	0.0356
TTD05	0.0347	0.0105	0.0452
TTD06	0.0224	0.0366	0.0232
TTD07	0.0106	0.0289	0.0205
TTD08	0.0239	0.0362	0.0601
TTD09	0.0313	0.0187	0.0354
TTD10	0.0244	0.0474	0.0652
TTD11	0.0273	0.0714	0.0442
TTD12	0.0722	0.0491	0.0649
TTD13	0.0597	0.0541	0.0547
TTE01	0.0134	0.0279	0.0200
TTE02	0.0213	0.0357	0.0421
TTE04	0.0353	0.0434	0.0490
TTE05	0.0189	0.0049	0.0140
TTE06	0.0474	0.0503	0.0029
TTF01	0.2736	0.2812	0.0077
TTF02	0.0457	0.0325	0.0344
TTF04	0.0125	0.0378	0.0253
TTF05	0.0517	0.0125	0.0637
TTF06	0.0341	0.0297	0.0510
TTF07	0.0232	0.0448	0.0355
TTF08	0.0220	0.0528	0.0502
TTF09	0.0359	0.0271	0.0088
TTF10	0.0239	0.0234	0.0473
TTF11	0.0238	0.0355	0.0245
TTF12	0.0425	0.0332	0.0167
TTF13	0.0243	0.0327	0.0198
TTF14	0.0483	0.0491	0.0119
TTF15	0.0381	0.0280	0.0301
TTF16	0.0152	0.0194	0.0286
TTG01	0.0379	0.0359	0.0479
TTG02	0.0391	0.0229	0.0458
TTG03	0.0383	0.0466	0.0810
TTG04	0.0590	0.0353	0.0631
TTG05	0.0274	0.0283	0.0287
TTG06	0.0345	0.0183	0.0256
TTG07	0.0868	0.0567	0.0663
TTG08	0.0727	0.1020	0.0368
TTG09	0.0363	0.0454	0.0817

Table 6: Correlations between propensity score and laerer variables
(continued)

Scale	AveCorrW	AveCorrP	Diff
TTH01	0.0277	0.0233	0.0421
TTH02	0.0078	0.0235	0.0312
TTH03	0.0264	0.0224	0.0486
TTH04	0.0286	0.0244	0.0501
TTH05	0.0380	0.0463	0.0647
TTH06	0.0210	0.0195	0.0265
TTH07	0.0241	0.0155	0.0208
TTH08	0.0289	0.0185	0.0473
TTI01	0.0650	0.0374	0.0310
TTI02	0.0797	0.0471	0.0596

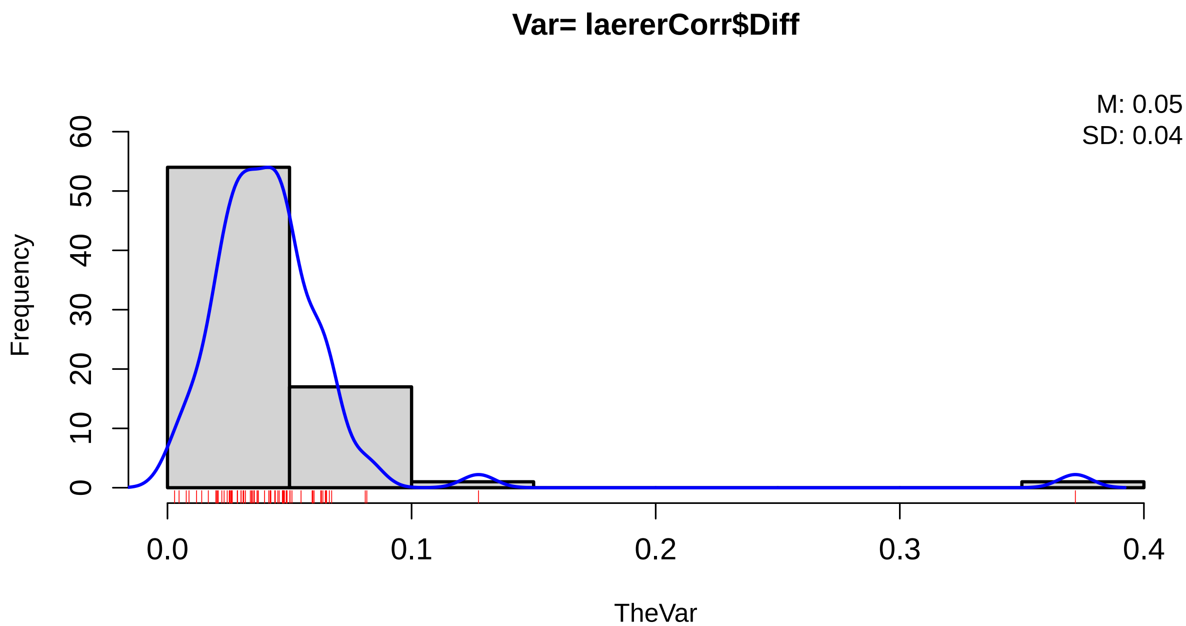


Figure 52: Distribution of correlations between laerer items and propensity score

Table 6 and Figure 52 show the corresponding table and figure for the teacher survey. Here, the distribution of correlations is generally below 0.1, but two differences in correlations are quite large. These are for the two items regarding whether teachers work at a different school and how many other schools they work at (TTB02 and TTB03). The propensity score is much more weakly related to these variables than the weights. This is because the propensity score only takes into account school level factors and the weights specifically adjust for these variables. Then, these high correlations reflect the design of the two measures and are not sources of bias. The other correlations are quite small. Here, we again see a lack of evidence of bias and the TALIS survey seems to be robust across to biases that might arise from differences between the sample and population in many school-level characteristics. This, though, is a much weaker sensitivity analysis than at the school level because there are no teacher level variables in the propensity score.

6 Appendix 1

6.1 Comparing Schools with any returned surveys to the population

This section compares those schools that returned any data (both participating schools and those with low participationg). This captures those schools that we could use data from for the national analyses. Figures 53-60 shows these results, which are not discussed.

Figure 53: Comparison of estimates from participating schools to population value for school location (county/fylke)

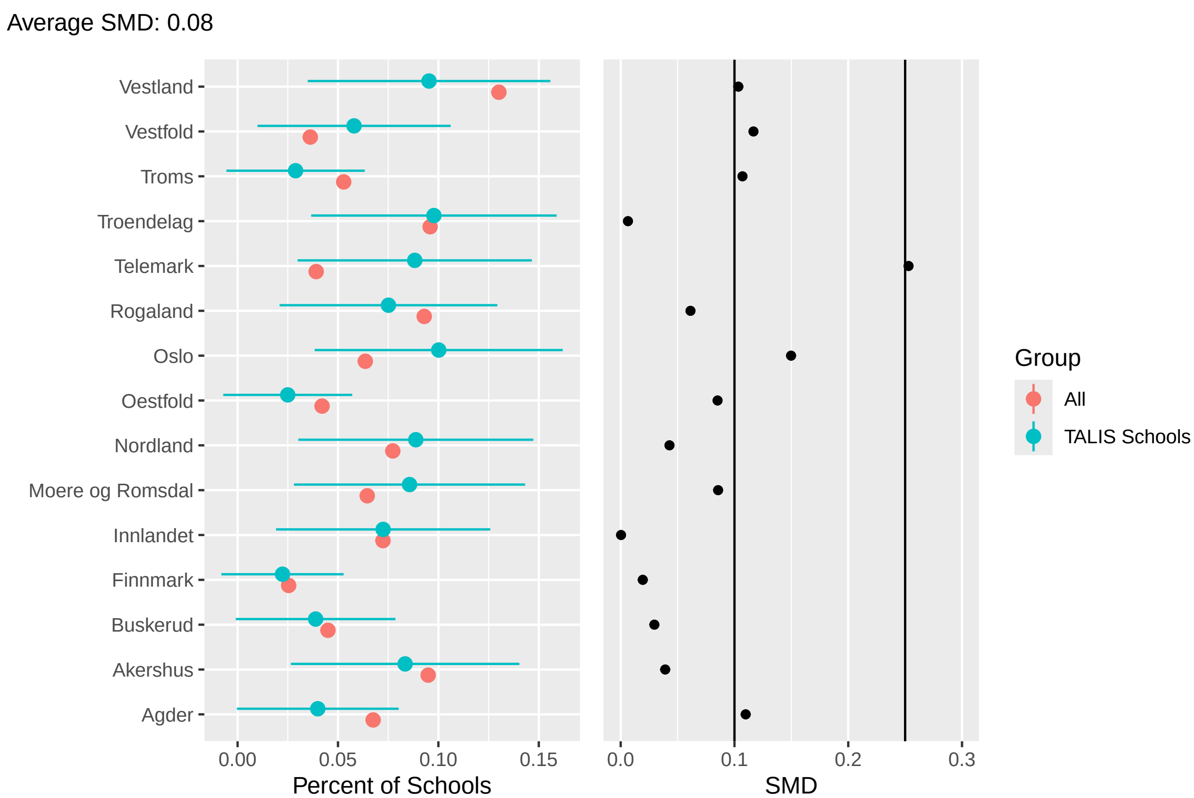


Figure 54: Comparison across per pupil expenses

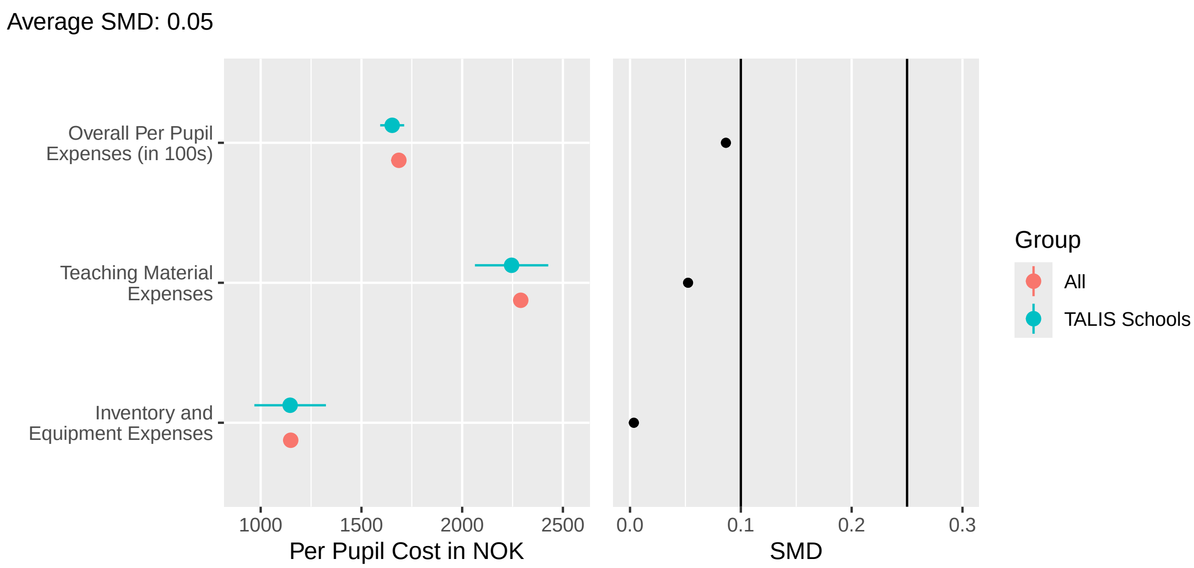


Figure 55: Comparison across national test Scores

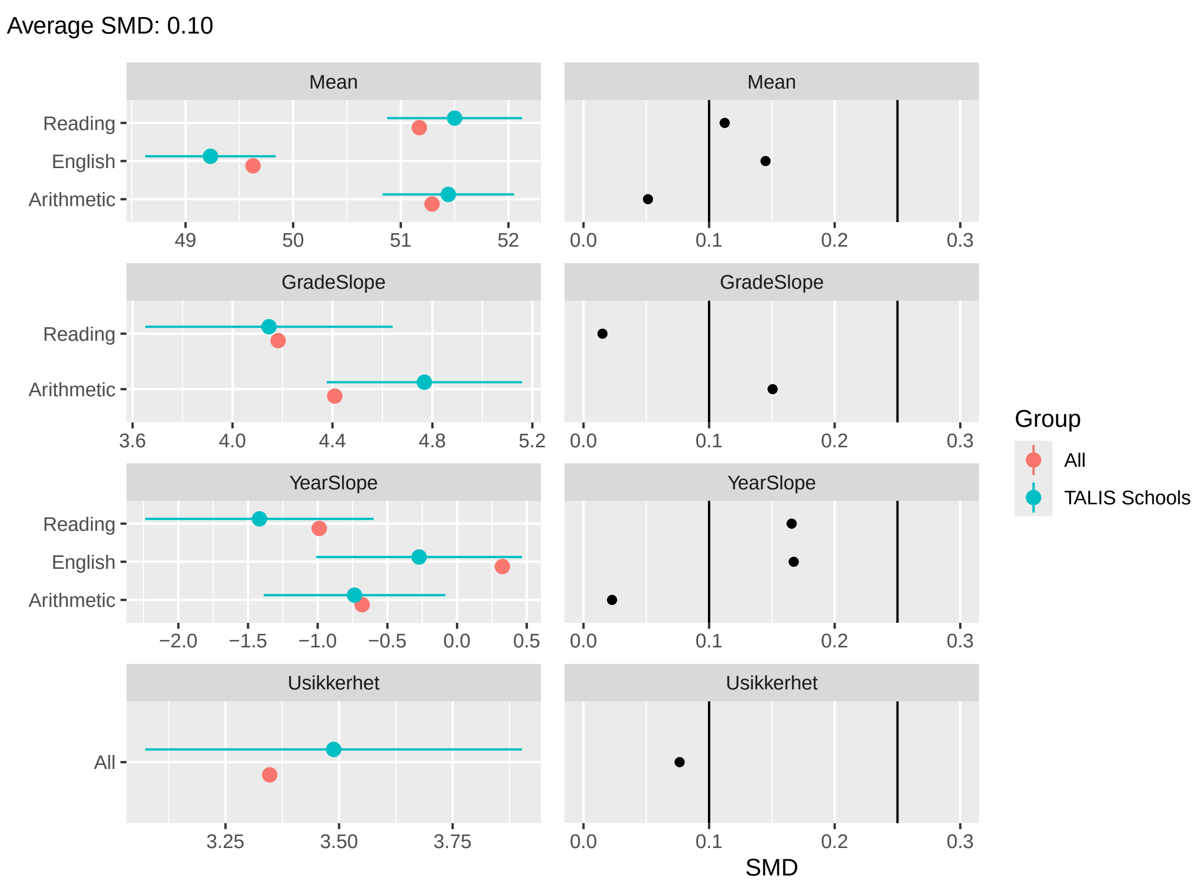


Figure 56: Comparison across amount of School Resources

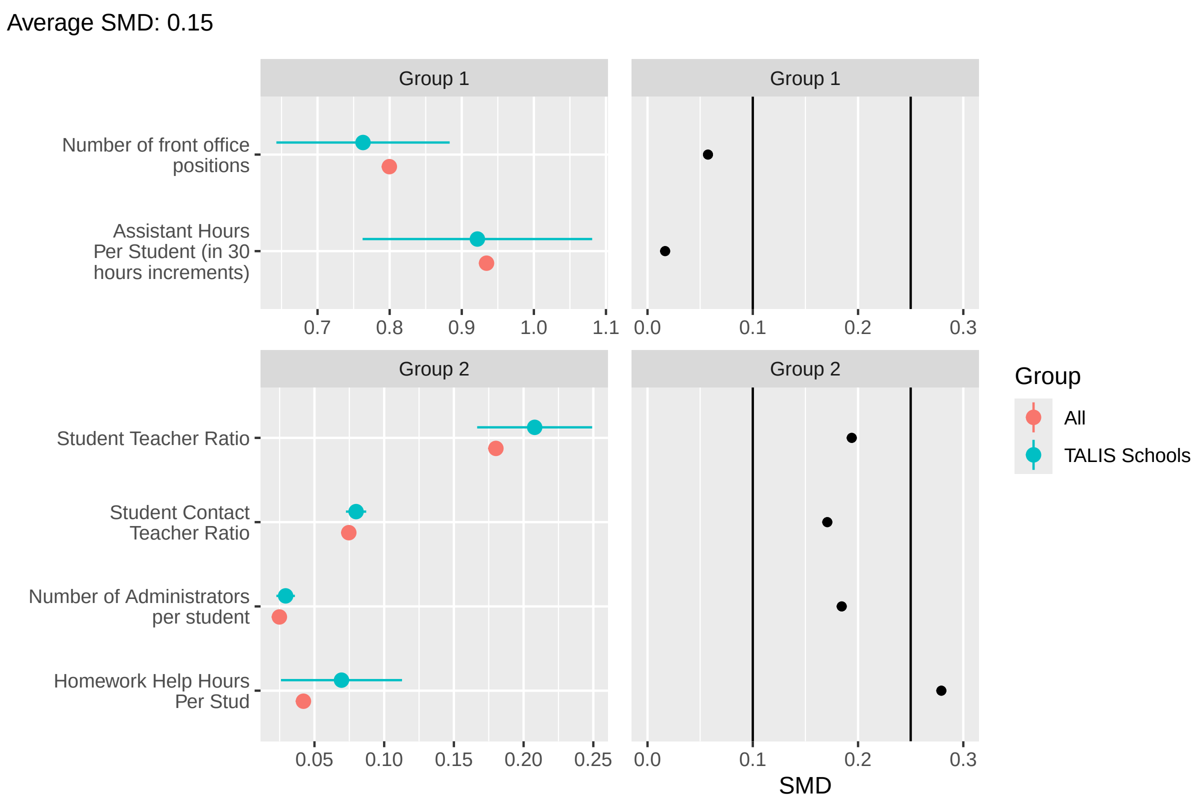


Figure 57: Comparison across school Size

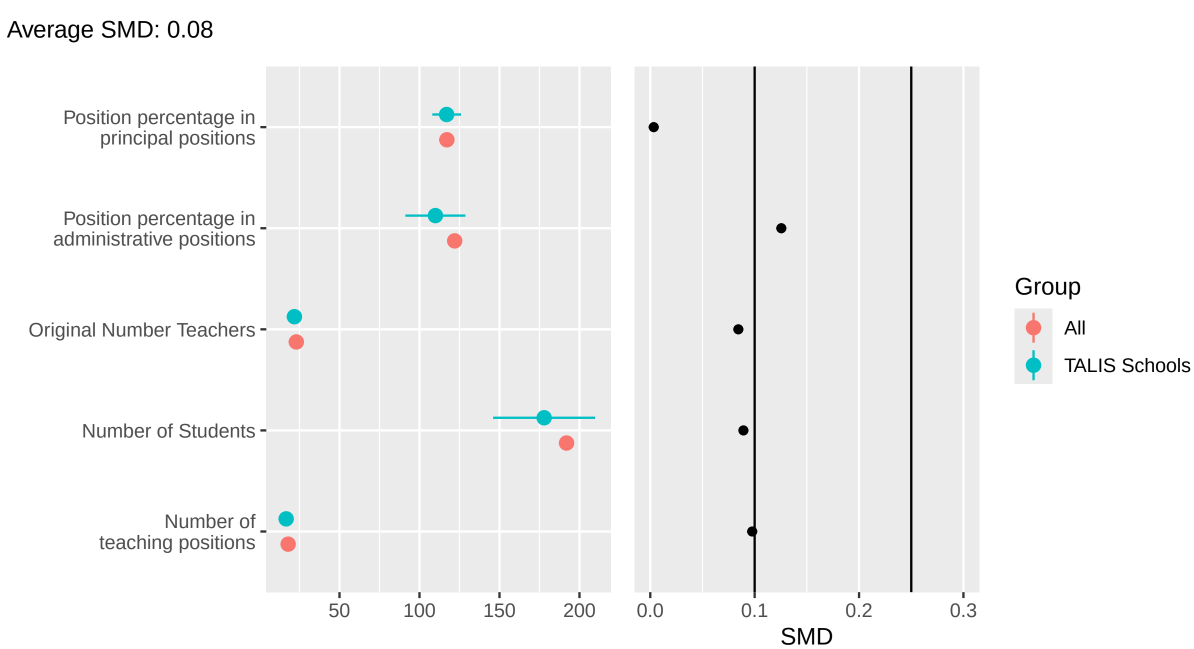
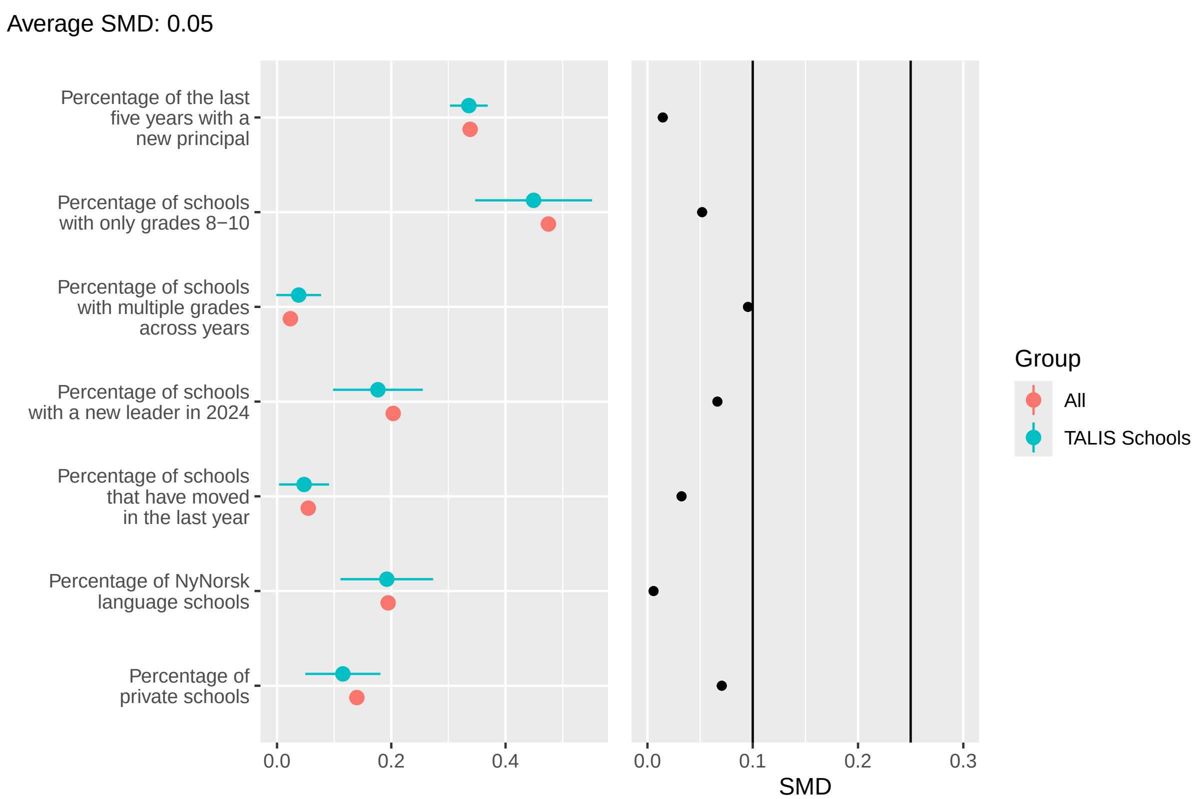


Figure 58: Comparison across characteristics of school



shows the same sample estimates for a set of variables related to instruction, bullying, and student characteristics.

Figure 59: Comparison across teacher and student Characteristics

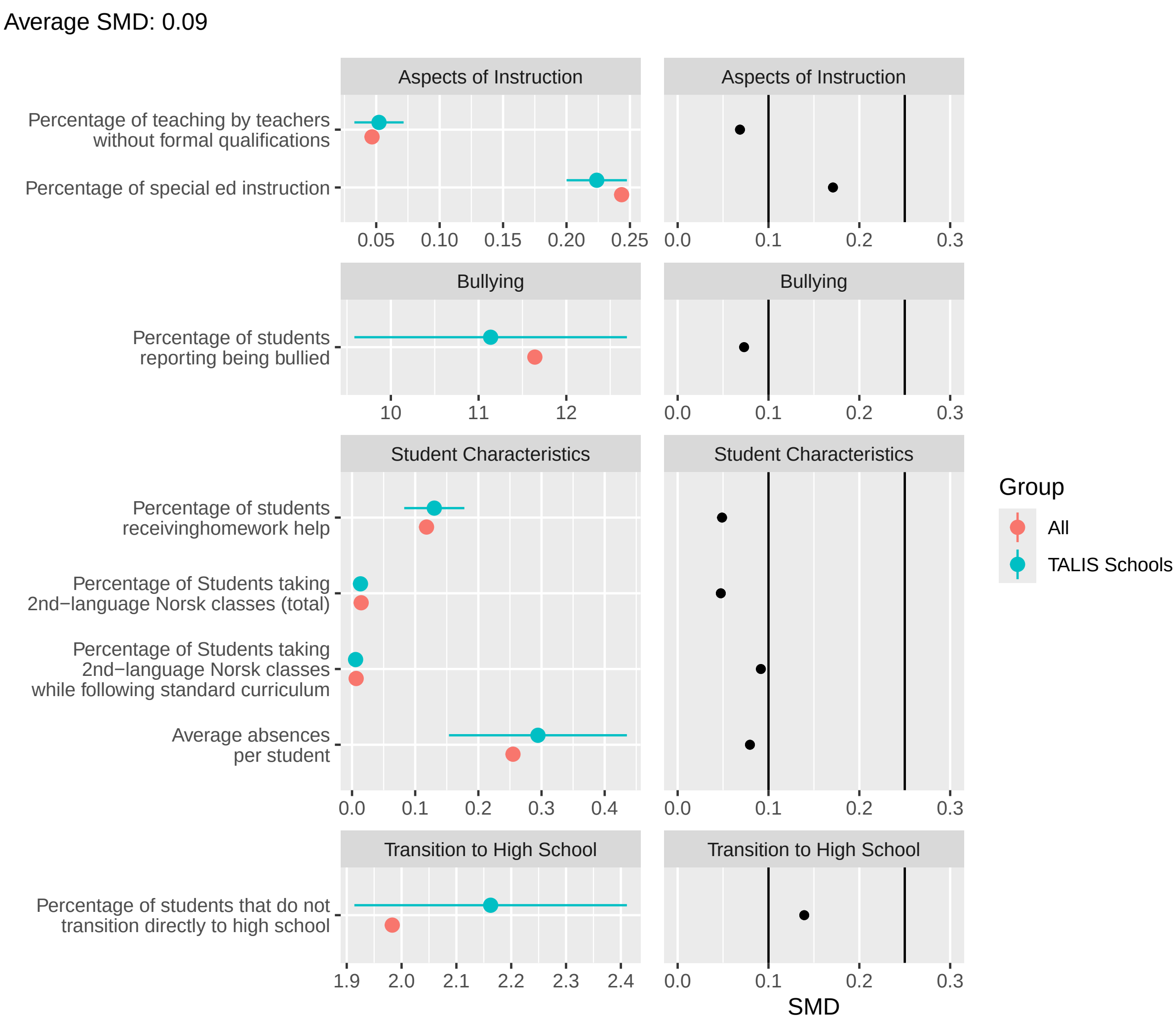
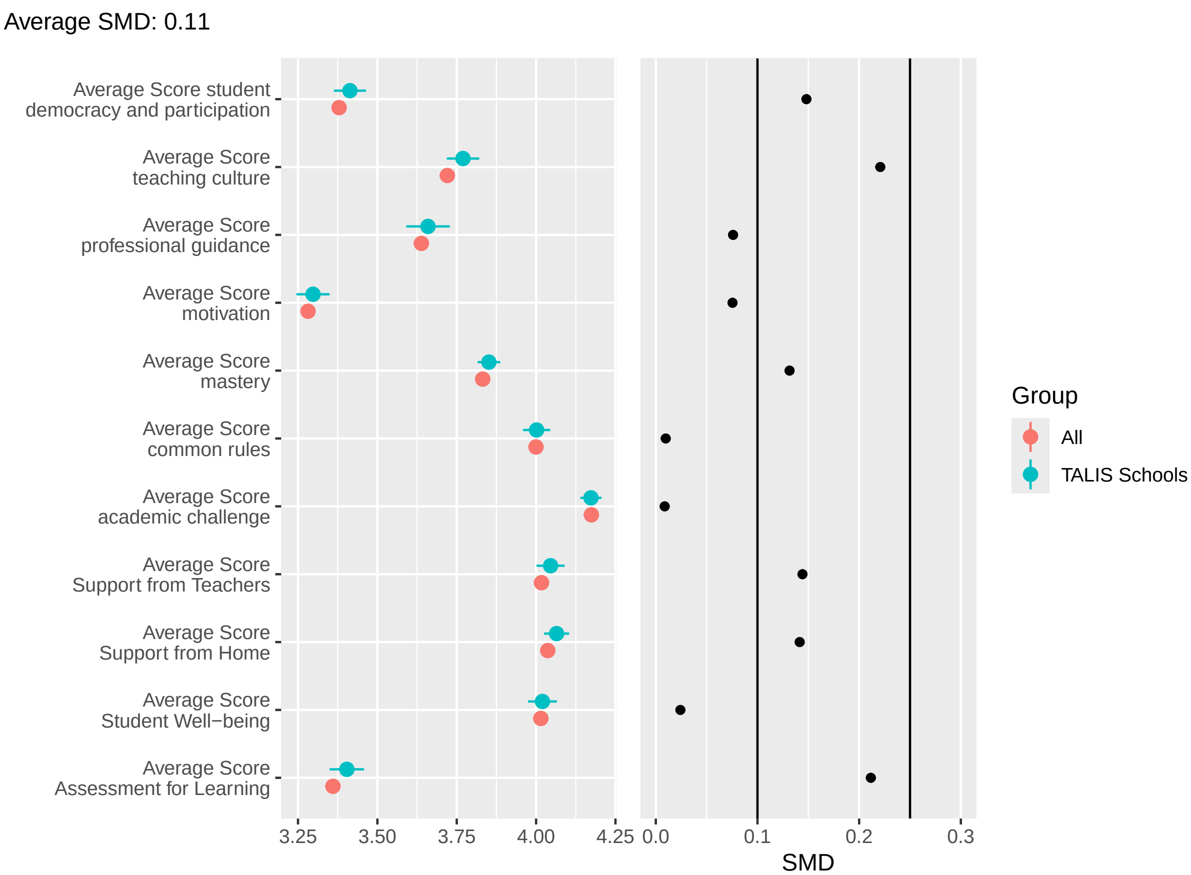


Figure 60: Comparison across Student Survey Scores



Bibliography

Assessing the representativeness of public opinion surveys. (2012). https://www.pewresearch.org/politics/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/

Curtin, R., Presser, S., & Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, *64*(4), 413–428. https://doi.org/10.1086/318638

Davern, M. (2013). Nonresponse Rates are a Problematic Indicator of Nonresponse Bias in Survey Research. *Health Services Research*, *48*(3), 905–912. https://doi.org/10.1111/1475-6773.12070

Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly*, *72*(2), 167–189. https://doi.org/10.1093/poq/nfn011

Hartman, E., & Huang, M. (2024a). Sensitivity Analysis for Survey Weights. *Political Analysis*, *32*(1), 1–16. https://doi.org/10.1017/pan.2023.12

Hartman, E., & Huang, M. (2024b). Sensitivity Analysis for Survey Weights. *Political Analysis*, *32*(1), 1–16. https://doi.org/10.1017/pan.2023.12

Hendra, R., & Hill, A. (2019). Rethinking Response Rates: New Evidence of Little Relationship Between Survey Response Rates and Nonresponse Bias. *Evaluation Review*, *43*(5), 307–330. https://doi.org/10.1177/0193841X18807719

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236. https://doi.org/10.1093/pan/mpl013

Little, R. J., & Vartivarian, S. (2004). Does weighting for nonresponse increase the variance of survey means? *The University of Michigan Department of Biostatistics Working Paper Series*, *35*, 20. http://www.bepress.com/umichbiostat/paper35