



Oxford Policy Management

ESSPIN Composite Survey 2

Overall report

Stuart Cameron

April 2015

Acknowledgements

The author is very grateful to the various people who helped in producing this report. Ravi Somani, Sourovi De, Peter-Sam Hill and Yo-Yo Chen worked on parts of the data analysis. Nicola Ruddle wrote the companion state reports. David Megill reviewed weights from the first Composite Survey and produced the sampling weights application for the second Composite Survey. Allan Findlay provided support with school identification and the annual school census. From Education Sector Support Programme in Nigeria (ESSPIN), Jake Ross, Kayode Sanni, Fatima Aboki, Sandra Graham, Pius Elumeze, Lilian Breakell, John Kay, and Simon Thomson reviewed preliminary notes and drafts and provided valuable comments. Gratitude is also due to everyone who conducted the second Composite Survey, including staff of Oxford Policy Management (OPM) and ESSPIN; state coordinators; staff of the State Universal Basic Education Boards who carried out the data collection; and not least to the large number of head teachers, teachers and students who took the time to participate in the study.

The Composite Survey was carried out by Oxford Policy Management for ESSPIN, which is funded by the UK Department for International Development (DFID) and managed by a consortium led by Cambridge Education. The survey project manager is Stuart Cameron. For further information contact stuart.cameron@opml.co.uk.

Oxford Policy Management Limited

6 St Aldates Courtyard
38 St Aldates
Oxford OX1 1BN
United Kingdom

Tel +44 (0) 1865 207 300
Fax +44 (0) 1865 207 301
Email admin@opml.co.uk
Website www.opml.co.uk

Registered in England: 3122495

Executive summary

This report presents findings from the first and second rounds of the ESSPIN Composite Survey (CS1 and CS2). These took place in 2012 and 2014, respectively. The survey covered a wide range of indicators at the teacher, head teacher, school-based management committee, and pupil levels. The aim is to understand change in schools over time, and whether schools which receive interventions are working better than those which do not. The main findings are as follows:

Teacher competence: There was no significant change in the proportion of teachers meeting ESSPIN standards for teacher competence in 2014 compared to 2012. But scores did improve, both for use of teaching aids and for use of praise and reprimands in the classroom. Teachers in ESSPIN schools do better on most criteria than those in non-ESSPIN schools. Those who individually received training do even better. There is also evidence that teachers benefiting from ESSPIN interventions improved faster (or, at least, worsened more slowly) between 2012 and 2014. Analysis of teacher English and mathematics tests shows that teachers continue to struggle with some primary-level material. They find higher grade material and English most difficult. But ESSPIN-trained teachers achieve higher scores than those who are not in ESSPIN schools.

Head teacher effectiveness is improving over time across the focus states. It is better in ESSPIN schools than non-ESSPIN ones. There is also evidence that effectiveness has improved faster among head teachers who received more ESSPIN support.

School development planning is improving in focus states. It is better in ESSPIN than non-ESSPIN schools. There is suggestive evidence that the pace of improvement may be faster in schools that had more ESSPIN intervention between 2012 and 2014.

Using our measure of **school inclusiveness**, fewer schools met the inclusiveness standard in 2014 than in 2012. In 2014, 25% of ESSPIN schools but only 8% of control schools met the standard. The drop in our inclusiveness score between 2012 and 2014 was also smaller in ESSPIN than non-ESSPIN schools.

School-Based Management Committees (SBMCs) functioned better in 2014 than in 2012. They are much better in ESSPIN schools than in non-ESSPIN schools. In CS2, 62% of ESSPIN schools met the SBMC functionality standard, compared to only 13% of non-ESSPIN schools. Both ESSPIN and non-ESSPIN schools were improving over time, but ESSPIN schools improved more quickly. SBMCs in ESSPIN schools were also more likely to be inclusive of women. 48% met this standard, compared to only 2% in non-ESSPIN schools. The same is true for inclusiveness of children, with 18% meeting the standard, compared to 2% in non-ESSPIN schools. SBMCs in ESSPIN schools had also improved in terms of their inclusiveness during 2012–2014. There was no significant change over time in non-ESSPIN schools.

School quality: The proportion of schools reaching the overall school quality standards increased significantly, from 3% to 10%. However, there was no significant change in the average level of quality, measured on a continuous scale. In 2014 there were significant and large differences in quality between ESSPIN and non-ESSPIN schools. Only around 1% of non-ESSPIN schools met the quality standard, compared to over 30% of ESSPIN schools. The pace of improvement between 2012 and 2014 was also faster in schools which received more ESSPIN intervention during that period than in schools which received less. The estimated number of children in good quality schools in the six states rose from under 200,000 in 2011/12 to over 650,000 in 2013/14. Of this increase of 450,000 in the number of children attending good quality schools, 90% were in ESSPIN schools.

Pupil learning: Test results in numeracy and English literacy for grade 2 and 4 pupils suggest that learning outcomes may be worsening over time in focus states. However, children in ESSPIN schools have significantly better test results than those in non-ESSPIN schools. This remains the case across all four types of test. There is some evidence that the change over time in pupil learning outcomes is more positive (improving faster, or worsening more slowly) in schools that received more ESSPIN intervention than in those that received less. The difference in results in 2014 between ESSPIN schools and non-ESSPIN ones remains even after controlling for possible confounding school characteristics, although the results when examining the rate of improvement over time are less positive. We considered whether rapid increases in enrolment and in pupil-teacher ratios might be responsible for slowing down, and even reversing, improvements in learning outcomes, but did not find evidence in support of this hypothesis.

Box 1. The good news and the bad news from the composite surveys

Positive results in this report include:

- Pupil test results are better in ESSPIN than in other schools (p. 44), even when we control for state and school characteristics (p. 54)
- School quality is better in ESSPIN than in other schools, and is improving greatly over time in ESSPIN, but not in control, schools (p. 41)
- Teachers trained by ESSPIN are more competent and are improving faster than those in other schools (p. 17)
- Head teachers are more effective in ESSPIN than in other schools, and have been continuing to improve in ESSPIN schools (p. 25)
- School development planning is improving over time and is much better in ESSPIN schools than in other schools (p. 29)
- ESSPIN schools are much more inclusive than other schools (p. 32)
- SBMCs are becoming more functional and inclusive over time, and 62% of ESSPIN schools meet the standard for functioning SBMCs (p. 35)

Some challenges identified in the report include:

- Pupil test results as a whole appear to be getting worse – although perhaps more slowly in ESSPIN than in other schools (p. 46)
- Most pupils score below 50%, and many score below 25%, in each test, with particular difficulties in writing at grades 2 and 4, reading with comprehension at grade 4, and multiplication and division at grade 4 (p. 47)
- Few students even in ESSPIN schools meet basic standards in literacy and numeracy (p. 45)
- Teachers in non-ESSPIN schools in the six states show no sign of improvement (p. 18)
- Teachers' own knowledge in English and mathematics is often weak. For instance, on average they score under 50% in English questions pitched at primary grade 2 and above, and they struggle with foundational concepts for teaching literacy and in writing (p. 21)
- Fewer than 20% of head teachers meet our effectiveness standard and there has been no improvement over time in non-ESSPIN schools (p. 24)
- Schools are becoming (by our measure) less inclusive over time – although the decline is less severe in ESSPIN schools (p. 31)

Table 1. Summary of findings

Indicator		Change over time across all schools in the six states			The ESSPIN effect in 2014	
		All, 2012	All, 2014	ESSPIN 2012	ESSPIN 2014	Non-ESSPIN 2014
Competent teachers	(p. 15)	70	66	83	73	62*
Competent teachers (new measure)			26		37	19*
Effective head teacher	(p. 23)	14	20	13	34	14*
School development planning	(p. 27)	4	7*	9	20	3*
Inclusive school	(p. 30)	19	13	24	25	8*
SBMC functions	(p. 33)	22	31*	30	67	17*
SBMC inclusive of women	(p. 36)	15	16	22	48	2*
SBMC inclusive of children	(p. 38)	6	6	10	18	2*
Good quality school	(p. 39)	4	10*	9	27	3*
Good quality school (new measure)			6		17	1*
Literacy grade 2	(p. 43)	30	30	33	41	26*
Literacy grade 4	(p. 43)	34	29*	38	40	25*
Numeracy grade 2	(p. 43)	48	38*	55	47	34*
Numeracy grade 4	(p. 43)	36	32*	40	39	30*

Note. ESSPIN here means schools that received output 3 intervention during 2009/10, 2010/11, 2011/12, and/or 2012/13. See chapter 2 for details. * against the results for non-ESSPIN schools in 2014 means that the results are significantly better in ESSPIN schools ($p < .05$). * against the results for all schools in 2014 means that the results for all schools have changed significantly over time. Results shown for teacher competence are specifically for teachers who have been trained by ESSPIN.

Table of contents

Acknowledgements	i
Executive summary	ii
List of tables, figures and boxes	vii
List of abbreviations	x
1 Introduction	1
1.1 Context: increasing enrolments in ESSPIN states	1
1.2 ESSPIN's School Improvement Programme	2
2 Methods	6
2.1 Evaluation strategy	6
2.1.1 Classifying the amount of ESSPIN intervention	6
2.1.2 Modes of analysis	7
2.2 Sample and weights	10
2.2.1 Sample design	10
2.2.2 Weights	12
2.2.3 Sample coverage	12
2.3 Training, pilots and fieldwork model	13
3 Findings	15
3.1 Teacher competence	15
3.1.1 Main analysis	15
3.1.2 Findings from the teacher content knowledge tests	19
3.1.3 Summary and discussion	23
3.2 Head teacher effectiveness	23
3.3 School development planning	27
3.4 School inclusiveness and SBMCs	30
3.4.1 School inclusiveness: meeting the needs of all pupils	30
3.4.2 How well do SBMCs function?	33
3.4.3 How inclusive are SBMCs of women?	36
3.4.4 How inclusive are SBMCs of children?	38
3.5 School quality	39
3.6 Pupil learning achievement in English literacy and numeracy	43
3.6.1 Main analysis	43
3.6.2 Distribution of test scores and sub-scale scores	47
3.6.3 Summary and discussion	51
4 Controlling for confounding school characteristics and changes in enrolment	52
4.1 Differences between ESSPIN and non-ESSPIN schools	52
4.2 Controlling for school and pupil characteristics	54
4.2.1 Timing of ESSPIN intervention and learning outcomes in 2014	55
4.2.2 Are learning outcomes better in ESSPIN schools in 2014, controlling for school and pupil characteristics?	56
4.2.3 Have learning outcomes improved faster in schools with ESSPIN intervention, controlling for school characteristics?	57
4.2.4 Summary	59
5 Conclusion and implications of survey findings for ESSPIN programme	60
References	62
Annex A Indicators	63

A.1	Pupil learning logframe indicators	63
A.2	Total test scores	66
Annex B	Note on changes to assessments for CS2	67
B.1	Introduction	67
B.2	Statistical analysis	67
B.3	Question removal and skip patterns	68
B.4	Pupil background	69
B.5	Disability	69
B.5.1	Ability to hear	69
B.5.2	Ability to speak	70
B.5.3	Ability to see	70
B.5.4	Ability to write	70
Annex C	ESSPIN output 3 interventions	72
Annex D	ESSPIN output 4 interventions	73
Annex E	Additional tables and figures: pupil test results in CS2	74
E.1	Disaggregated scores by sub-scale and state	74
E.2	Distribution of pupils by test score quartiles, by learning domain or grade level	75
E.3	Proportion of pupils scoring 0%–24%, by state and learning domain	77
E.4	Proportion of pupils scoring 75%–100%, by state and learning domain	79
E.5	Proportion of pupils in each band, by ESSPIN status and state	81
E.6	Proportion of grade 4 pupils in each band on grade 1/2 level items only, by ESSPIN status and state	85
E.7	Gap between ESSPIN and non-ESSPIN schools in the proportion of children scoring under 25%	87

List of tables, figures and boxes

Table 1. Summary of findings.....	iv
Table 2. Number of schools and enrolment in the 2009 and 2013 school censuses.....	2
Table 3. Proportion of schools receiving full package of ESSPIN output 3 interventions (%).....	3
Table 4. Definition of ESSPIN schools	6
Table 5. Sample in CS1 and CS2 and population of schools, by state and with intervention groups	11
Table 6. Sample coverage in CS2.....	13
Table 7. Instruments used in CS2	14
Table 8. Teacher competence in 2012 and 2014	16
Table 9. Teacher competence in 2014, by intervention group	17
Table 10. Teacher competence difference in differences (comparison of means)	18
Table 11. Teacher competence difference in differences (regression)	19
Table 12. Average scores in different types of item in the teacher tests (%).....	21
Table 13. Academic qualifications of teachers according to whether they sat English and mathematics tests in CS2	22
Table 14. Head teacher effectiveness in 2012 and 2014.....	24
Table 15. Head teacher effectiveness in CS2, by intervention group	25
Table 16. Teacher qualifications and absenteeism.....	25
Table 17. Head teacher effectiveness difference in differences (comparison of means)	26
Table 18. Head teacher effectiveness difference in differences (regression).....	27
Table 19. SDP effectiveness in 2012 and 2014.....	28
Table 20. SDP effectiveness in CS2, by intervention group	29
Table 21. SDP effectiveness difference in differences (comparison of means)	29
Table 22. SDP effectiveness difference in differences (regression).....	30
Table 23. School inclusiveness in 2012 and 2014.....	31
Table 24. School inclusiveness in CS2, by intervention group	32
Table 25. School inclusiveness difference in differences (comparison of means).....	33
Table 26. School inclusiveness difference in differences (regression).....	33
Table 27. SBMC functionality in 2012 and 2014.....	34
Table 28. SBMC functionality in CS2, by intervention group	35
Table 29. Difference in differences of SBMC functionality (comparison of means)	36
Table 30. Difference in differences of SBMC functionality (regression)	36
Table 31. SBMCs' women's inclusion in 2012 and 2014	37
Table 32. SBMC women's inclusion in 2014, by intervention group	37
Table 33. Difference in differences of SBMC women's inclusion (comparison of means).....	37
Table 34. Difference in differences of SBMC women's inclusion (regression)	37
Table 35. SBMC inclusion of children in 2012 and 2014	38
Table 36. SBMC children's inclusion in 2014, between intervention groups	39
Table 37. Difference in differences in SBMC children's inclusion (comparison of means)	39
Table 38. Difference in differences in SBMC children's inclusion (regression)	39
Table 39. School quality in 2012 and 2014	40
Table 40. School quality in CS2, by intervention group	40
Table 41. School quality difference in differences (comparison of means)	41
Table 42. School quality difference in differences (regression with continuous intervention variable)	42
Table 43. Test scores and proportion of children reaching logframe indicator in 2012 and 2014...	44
Table 44. Test scores and proportion of children reaching logframe indicator in 2014, by intervention group	44
Table 45. Pupil test score difference in differences (comparison of means)	46
Table 46. Pupil test score difference in differences (regression)	47
Table 48. Detailed scores in test sub-scales, 2012 vs. 2014 (%).....	48
Table 49. Detailed scores in test sub-scales, ESSPIN vs. non-ESSPIN.....	49
Table 50. Characteristics of ESSPIN and non-ESSPIN schools, by state	54
Table 51. Difference in test scores by timing of ESSPIN intervention.....	56

Table 52. Estimates of the effect of ESSPIN intervention on learning outcomes in CS2	57
Table 53. Identification of control and ESSPIN groups for difference in differences analysis.....	58
Table 54. Estimates of the effect of ESSPIN intervention on changes in learning outcomes between 2012 and 2014.....	58
Table 55. Grade 4 literacy test results in control and ESSPIN schools, in 2012 and 2014 (%)	59
Table 56. Detailed scores in test sub-scales, ESSPIN vs. non-ESSPIN (disaggregated by state). 74	
Figure 1. Number of schools receiving a full package of ESSPIN output 3 interventions.....	3
Figure 2. Number of teachers ESSPIN worked with during 2012, 2013 and 2014	4
Figure 3. Number of children according to how many years of output 3 intervention their school has received	4
Figure 4. Teacher competence score (strict version), by state and ESSPIN group.....	17
Figure 5. Teacher test scores by grade level and state	22
Figure 6. Number of head teacher effectiveness criteria met, by state and intervention group	26
Figure 7. Number of school development planning criteria met, by state and intervention group...	29
Figure 8. Inclusiveness score, by state and intervention group	32
Figure 9. Number of SBMC functionality criteria met, by state and intervention group	35
Figure 10. School quality score, by state and intervention group.....	41
Figure 11. School quality in 2012 and 2014, in control and ESSPIN schools	42
Figure 12. Number of children in a good quality school	43
Figure 13. Test scores by state and ESSPIN intervention	45
Figure 14. Pupil test scores in ESSPIN and control schools, in 2012 and 2014	46
Figure 15. Distribution of test scores by ESSPIN intervention in CS2.....	47
Figure 16. Test scores by grade level of item and state in CS2 (grade 4 tests)	50
Figure 18. Distribution of pupils by test score quartile for each learning domain	50
Figure 19. Distribution of pupils by test score quartile and grade level of questions	51
Figure 20. How matching techniques work (roughly)	55
Figure 21. Distribution of pupils by test score quartiles, learning domain and ESSPIN status	75
Figure 22. Distribution of pupils by test score quartiles, grade level of questions, and ESSPIN status.....	76
Figure 23. Proportion of pupils scoring 0%–24% in grade 2 literacy, by state and learning domain	77
Figure 24. Proportion of pupils scoring 0%–24% in grade 4 literacy, by state and learning domain	77
Figure 25. Proportion of pupils scoring 0%–24% in grade 2 numeracy, by state and learning domain.....	78
Figure 26. Proportion of pupils scoring 0%–24% in grade 4 numeracy, by state and learning domain.....	78
Figure 27. Proportion of pupils scoring 75%–100% in grade 2 literacy, by state and learning domain.....	79
Figure 28. Proportion of pupils scoring 75%–100% in grade 4 literacy, by state and learning domain.....	79
Figure 29. Proportion of pupils scoring 75%–100% in grade 2 numeracy, by state and learning domain.....	80
Figure 30. Proportion of pupils scoring 75%–100% in grade 4 numeracy, by state and learning domain.....	80
Figure 31. Grade 2 literacy.....	81
Figure 32. Grade 4 literacy.....	82
Figure 33. Grade 2 numeracy	83
Figure 34. Grade 4 numeracy	84
Figure 35. Proportion of grade 4 pupils in each band for grade 1/2 level literacy items, by ESSPIN status and state	85
Figure 36. Proportion of grade 4 pupils in each band for grade 1/2 level numeracy items, by ESSPIN status and state.....	86
Figure 37. Gap between ESSPIN and non-ESSPIN schools in the proportion of children scoring under 25%, by state and learning domain (grade 2 literacy).....	87

Figure 38. Gap between ESSPIN and non-ESSPIN schools in the proportion of children scoring under 25%, by state and learning domain (grade 2 numeracy) 87

Box 1. The good news and the bad news from the composite surveys	iii
Box 2. Difference in differences	9
Box 3. Logframe standard for teacher competence	15
Box 4. What was in the teacher content knowledge tests?	20
Box 5. Logframe standard for head teacher effectiveness.....	24
Box 6. Logframe standard for effective school development planning	28
Box 7. Standard for school inclusiveness (meeting needs of all pupils).....	31
Box 9. Logframe standard for SBMC functionality	34
Box 10. Logframe standard for SBMCs' inclusiveness of women.....	36
Box 11. Logframe standard for SBMC inclusiveness of children	38
Box 12. Logframe standard for school quality	40

List of abbreviations

BEd	Bachelor of Education
CAPI	Computer-assisted personal interviews
CBOs	Community-based organisations
CS1	Composite Survey 1
CS2	Composite Survey 2
DFID	Department for International Development (UK)
ESSPIN	Education Sector Support Programme in Nigeria
HND	Higher National Diploma
L2	Grade 2 literacy test
L4	Grade 4 literacy test
LGA	Local Government Authority
LGEA	Local Government Educational Authority
MLA	Measurement of Learning Achievement
N2	Grade 2 numeracy test
N4	Grade 4 numeracy test
NCE	National Certificate of Education
OND	Ordinary National Diploma
OPM	Oxford Policy Management
PGDE	Post-Graduate Diploma in Education
SBMC	School-Based Management Committee
SDP	School Development Plan
SE	Standard error
SIP	School Improvement Programme
SSCE	Senior Secondary Certificate of Education
SSIT	State School Improvement Team
SSO	School Support Officer
WASC	West African Senior School Certificate

1 Introduction

The aims of the ESSPIN Composite Surveys are to assess the effects of ESSPIN's integrated School Improvement Programme (SIP), and to report on the quality of education in the six ESSPIN-supported states. The surveys address five output indicators: teacher competence, head teacher effectiveness, school development planning, SBMC functionality, and inclusive practices in schools. They also address one outcome indicator, school quality, and one impact indicator, pupil learning achievement.

In particular, the second round of the Composite Survey (CS2), conducted in 2014, aims to provide post-intervention data that can be compared to data from the first round of the Survey (CS1) collected in 2012 (ESSPIN, 2013a), in order to evaluate the extent of improvements in key indicators and gauge programme success.

This report presents findings from CS2 and comparisons between CS1 and CS2, covering all of ESSPIN's output, outcome and impact indicators. A related set of reports present results for each of the six states, while a Gender and Inclusion Report (De and Cameron, 2015) sets out results on the extent to which schools and SBMCs are inclusive, as well as gender differences within the schools.

In this chapter, we present information on the context of increasing enrolments in ESSPIN states and on the scale of the SIP, drawing on data from the annual school census and ESSPIN offices. Chapter 2 describes the methods used to conduct the survey and to estimate ESSPIN's impact. Chapter 3 presents findings on teacher competence, head teacher effectiveness, school development planning, school inclusiveness, SBMC functionality and inclusiveness, overall school quality and pupils' learning outcomes. In each case we assess both the overall trends in ESSPIN states, and whether ESSPIN schools are doing better or worse than other schools. In chapter 4, we extend the analysis of pupil learning outcomes from chapter 3, by controlling for location and other school and pupil characteristics. Chapter 5 summarises the findings and concludes.

1.1 Context: increasing enrolments in ESSPIN states

Information from annual school censuses indicates large increases in enrolment during the past few years in ESSPIN states (Table 2). There is some uncertainty about the magnitude of these changes because the number of schools listed in the 2013 census is much larger than that in the 2009 census, and it is not clear whether these are genuinely new schools or just schools that were missing from the 2009 data.

Nevertheless, even within the schools found in both censuses, there are increases in Jigawa, Kaduna and Kano of 16%–19%, representing over 500,000 additional students in these three states alone. Across the six states total enrolment has increased by between 12% (comparing only schools listed in both censuses) and 28% (comparing the totals from all schools listed in each census).

These rapid changes in pupil enrolment have to be taken into account in understanding both learning outcomes across the six states, and the context in which ESSPIN's programmes operate. They are likely to have put a strain on existing schools, which are not likely to have had commensurate increases in teachers or resources. Pupil–teacher ratios appear to have increased somewhat during this period. Both the number and the composition of pupils are likely to have changed, as learners from disadvantaged backgrounds, who would previously have been excluded, are presumably among the new entrants to the system.

Table 2. Number of schools and enrolment in the 2009 and 2013 school censuses

State	2009		2013		Enrolment change (%)	Enrolment change (schools found in both censuses only, %)
	Number	Enrolment	Number	Enrolment		
Enugu	1188	237,548	2349	327,834	38.0	-23.7
Jigawa	1789	427,180	2157	584,037	36.7	18.5
Kaduna	3947	972,985	4223	1,151,876	18.4	16.2
Kano	4768	1,883,472	6467	2,591,175	37.6	18.1
Kwara	1448	199,604	1497	198,248	-0.7	-2.2
Lagos	986	388,577	1009	400,277	3.0	0.4
Total	14126	4,109,366	17,702	5,253,447	27.8	12.4

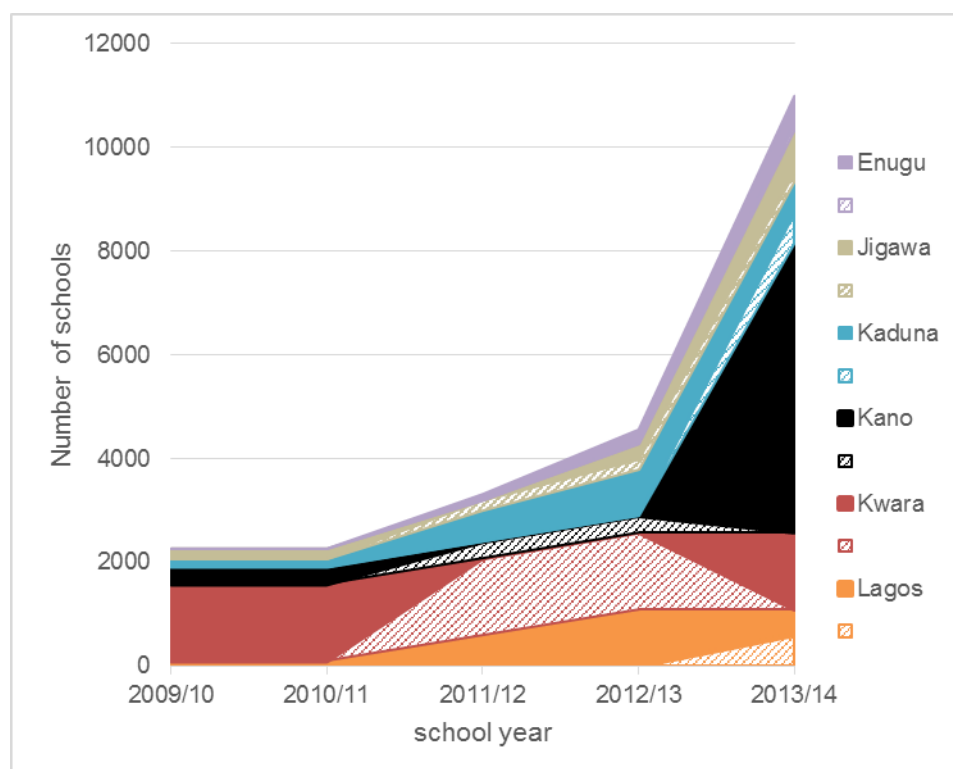
Note. Enrolment is for primary grades 1 to 6. The Enugu data for 2013 includes both public and private schools, as ESSPIN interventions have also covered some private (mission) schools; these schools were not captured in the 2009 census.

1.2 ESSPIN's School Improvement Programme

ESSPIN aims to bring about better learning outcomes for children of basic education school age in six states, with a range of activities at the state, national, local and school levels. It has four output streams, focusing on (i) strengthening federal government systems; (ii) increasing the capability of state and local governments as regards the governance and management of schools; (iii) strengthening the capability of primary schools to provide improved learning outcomes; (iv) and improving inclusion policies and practices in basic education (ESSPIN, 2013b).

Under the third of these outputs, ESSPIN's SIP aims to: provide and support the use of structured materials that ensure teachers can deliver quality instruction, and to strengthen teachers' own understanding of literacy and numeracy concepts; and to improve academic leadership and school improvement planning by head teachers (RTI International, 2014). The SIP typically works through a two-year modular programme of workshops and school visits, after which schools continue to receive school visits from government officers to help maintain and continue improving quality gains. Among the results of the programme, it is anticipated that 6,300 schools will meet a quality benchmark, representing 37% of all public primary and junior secondary schools in the six states, by 2016, and that 1.8 million students in primary grades 2 and 4 will have improved learning outcomes. At the same time, many of the same schools have been receiving interventions under the fourth output stream, facilitating community involvement and inclusion through SBMCs.

ESSPIN has massively scaled up its interventions during 2012/13 and 2013/14. In Lagos the SIP was rolled out to all schools in 2012/13 (although not all of those schools received additional intervention in 2013/14), and in Kano and Kwara it was rolled out to all schools in 2013/14. In total, it had reached 59% of schools in the six states in 2013/14, and an additional 9% had received its support at some point in the previous years (Table 3). The number of teachers that state officers reported having worked with showed a similar pattern of rapid scale-up during 2012 to 2014 (Figure 2), reaching over 80,000 in 2014, with the majority being in Kano.

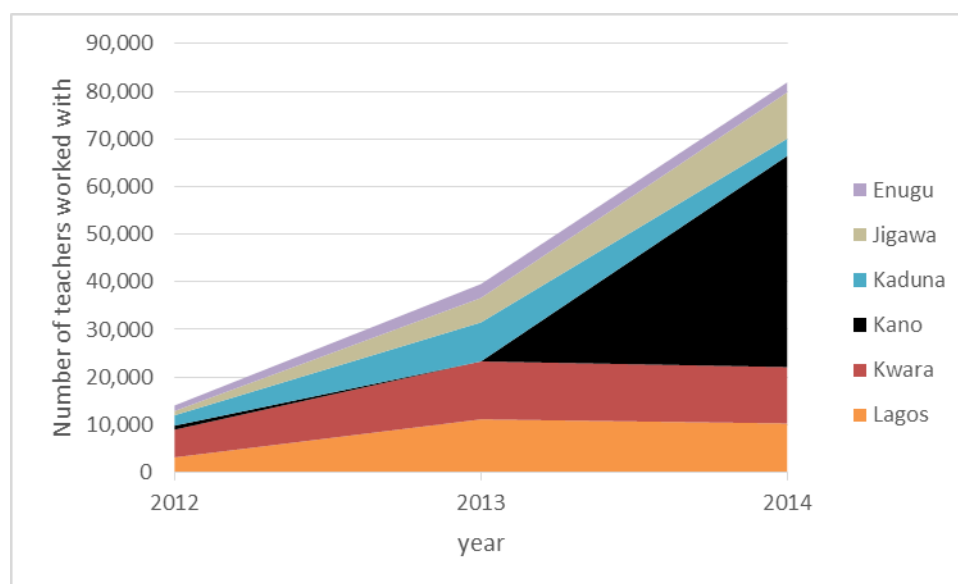
Figure 1. Number of schools receiving a full package of ESSPIN output 3 interventions

Source: author's calculations based on 2012/13 annual school census and intervention information provided by ESSPIN.
 Note. The areas with solid shading represent schools which received ESSPIN output 3 interventions in the year shown. The areas shaded with diagonal lines represent schools which did not receive any ESSPIN intervention in the given year, but that did in the previous years.

Table 3. Proportion of schools receiving full package of ESSPIN output 3 interventions (%)

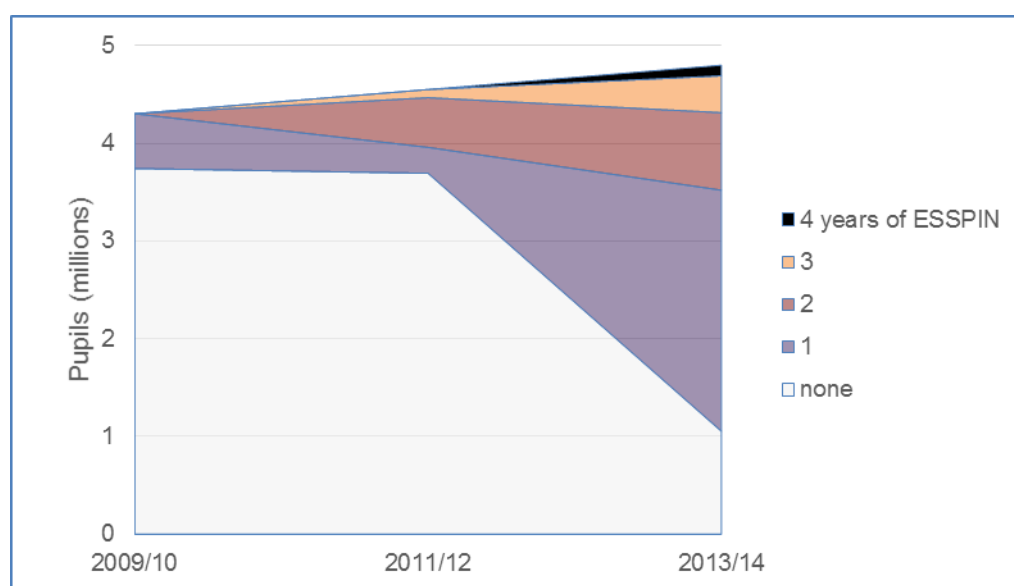
%	2009/10	2010/11	2011/12	2012/13	2013/14	Any year
Enugu	0	0	8	18	37	45
Jigawa	10	10	0	15	39	48
Kaduna	4	4	14	21	13	27
Kano	6	6	0	0	100	100
Kwara	100	100	0	0	100	100
Lagos	9	9	54	100	46	100
Total	14	14	9	16	59	68

Source: author's calculations based on 2012/13 annual school census and intervention information provided by ESSPIN.
 Note. Proportions are calculated relative to the total number of schools in the 2012/13 annual school census, and so the figures are not perfectly accurate for other years because the total number of schools changes slightly from year to year. Where census numbers are lower than ESSPIN's intervention tables, the information from ESSPIN is used, on the assumption that there are some missing data in the school census.

Figure 2. Number of teachers ESSPIN worked with during 2012, 2013 and 2014

Source: summary of state progress reports, provided by ESSPIN.

In 2014 over 3.7 million children were in schools that had received some output 3 ESSPIN intervention; this is 78% of the enrolled children in the six states. The number of children affected by an ESSPIN intervention increased by 2.9 million between 2011/12 and 2013/14 alone.

Figure 3. Number of children according to how many years of output 3 intervention their school has received

Source: ESSPIN intervention data and annual school censuses. The total is the estimated total number of children enrolled in school.

In the process of scaling up, there have been some changes to the model for delivering school support. During the pilot phase of ESSPIN (2009/10 and 2010/11), State School Improvement Teams (SSITs) trained directly by ESSPIN staff were responsible for supporting, and training head teachers and teachers directly. As the programme expanded, the School Support Officers (SSOs) – a second, larger group of employees working at the Local Government Educational Authority (LGEA) level – were trained by the SSITs and ESSPIN. Responsibility for working directly with

head teachers and teachers has been shifted progressively towards the SSOs, who are on average less qualified and have received less intensive training than the SSITs.

Thus, as the programme scales up, we may expect a narrowing in the difference between ESSPIN and non-ESSPIN schools, but because ESSPIN is reaching much larger numbers of schools, we can also expect the ESSPIN effect to start having an impact on school quality, teacher competence and learning outcomes for the states as a whole. For example, assume that 100% of ESSPIN schools but only 10% of non-ESSPIN schools meet a quality standard. In 2012/13, around 16% of schools across the six states receive ESSPIN output 3 intervention. That would mean that even with such a dramatic difference in quality between ESSPIN and non-ESSPIN schools, fewer than one in four schools overall would be meeting the quality standard. Even if the school quality gap between ESSPIN and non-ESSPIN schools then deteriorated somewhat, scaling up to cover over 50% of schools would make a dramatic difference to average school quality in the states as a whole.

For the present report, we are not able to assess whether the scale-up during the 2013/14 school year had had these expected effects. This is because CS2 was conducted in May/June 2014 – towards the end of the 2013/14 school year – which we judge to be too soon for the additional intervention to have taken effect in ways that can be measured with our indicators.

However, we are able to examine whether the relatively modest scale-up during 2012/13 has led to any dilution in the quality differential between ESSPIN and non-ESSPIN schools, and whether by this stage ESSPIN's SIP is starting to have an effect on state-level averages. Some 14% of schools across the five pilot states were reached during ESSPIN's pilot phase (Enugu did not participate in the pilot). An additional 14% have been reached during 2011/12 and 2012/13. It is in this period that we would expect changes to have taken place which would be reflected in the differences between the results of CS1 and those of CS2. There was extensive scale-up in Enugu, Kaduna and Lagos in particular. Despite this, the proportion of schools that had ever received ESSPIN intervention remained under 15% up to 2012/13 across Enugu, Jigawa, Kaduna and Kano, so our expectation is that even a large differential between the quality of ESSPIN and non-ESSPIN schools would have had a limited impact on trends at the state or overall level.

2 Methods

2.1 Evaluation strategy

2.1.1 Classifying the amount of ESSPIN intervention

ESSPIN was originally intended to be rolled out in a simple phased pattern across the six states, with schools falling into one of three groups: no intervention (control), phase 1 (roll-out prior to the 2012/13 school year) and phase 2 (roll-out in 2012/13 or 2013/14). In practice, as Figure 1 above shows, the roll-out has been more complicated, and in three of the states had reached all schools by 2013/14.

While this may be a sign of success for the ESSPIN programme, reflecting the enthusiasm with which state partners have taken on the programme and pushed for it to be rolled out to large numbers of schools, it presents a difficulty for evaluation. The original evaluation design for ESSPIN relied on maintaining a control group of schools with no intervention, which could be compared to those with a longer history of intervention (phase 1) and those where intervention started more recently (phase 2). As the schools no longer fall neatly into these phases, we instead grouped schools according to the amount of intervention they have received (see Annex C for full details). We focus on schools which had a ‘full package’ of output 3 interventions in a particular year – meaning leadership training for head teachers, teacher training, and school visits – and treat schools that had less than this full package as control schools. We also assume that there is a one-year lag between ESSPIN intervention and measurable impact.

We define ESSPIN schools slightly differently depending on whether we are looking at one point in time (cross-sectional analysis) or looking at change over time (see Table 4). In the tables and charts in chapter 3, we label ESSPIN schools either as schools which are *expected to be better* because of ESSPIN, or as schools which are *expected to have improved faster* because of ESSPIN. In each case the comparison point is a control group which did not have any intervention in the relevant period.

Table 4. Definition of ESSPIN schools

Labelled as	Defined as	Compared to (control group)	Type of analysis
Expected to be better	Schools that had at least one year of full package of ESSPIN output 3 ¹ intervention prior to 2013/14	Schools that did not receive the full package of output 3 intervention prior to 2013/14	One point in time (cross-sectional analysis of CS2)
Expected to have improved faster	Schools that had at least one year of full package of ESSPIN output 3 intervention during 2011/12 and 2012/13	Schools that did not receive the full package of output 3 intervention during 2011/12 and 2012/13	Change over time (difference in differences analysis)

In most cases these two definitions refer to the same set of schools. The schools that received a higher level of intervention during 2011/12 and 2012/13 are also those that have had more

¹ A companion report, ‘Composite Survey 2: Gender and Inclusion Report’ (De and Cameron, 2015), focuses on ESSPIN’s output 4 interventions, which run in parallel with output 3 and aim to improve inclusion and community participation in schools.

intervention overall (Table 5 in the following section shows the number of schools in each of these categories, and how they overlap). Consequently, the schools that we expect to be better in 2014 are mostly the same ones that we expect to have improved faster between 2012 and 2014. But there are some schools in Jigawa and Kano that were included in the pilot programme – and so are expected to be better overall – but that were not included in the roll-out of the programme during 2011/12 and 2012/13 – and so are not expected to have improved (relative to other schools) during the period between CS1 and CS2. The same applies to all schools in Kwara, which were included in the pilot programme during 2009/10 and 2010/11 but did not receive a full package of output 3 interventions during 2011/12 and 2012/13.² For these schools, we have to treat them differently depending on whether we are analysing the CS2 results alone (cross-sectional analysis) or analysing change over time (comparing CS2 to CS1).

For individual outcome indicators, we alter the classification scheme slightly according to the purpose of our analysis. For example, when examining teacher competence, we consider three different groups: teachers that have not been exposed to ESSPIN; teachers who are in schools that have received ESSPIN intervention but who have not themselves been trained by ESSPIN³; and teachers who have been trained through ESSPIN. We also use continuous versions of the intervention measures – for example, the number of years that a pupil has been exposed to expected improved school quality as a result of ESSPIN intervention. While categorical measures are easier to use for tables of descriptive statistics, a continuous measure makes sense in regression analysis, makes most use of the information, and helps us to avoid the risk that results might be altered by a slight change in the choice of categories.

2.1.2 Modes of analysis

The purpose of CS2 is both to provide insights into the changes over time in the six states where ESSPIN operates, and to evaluate whether ESSPIN is having an effect in the specific schools where its school improvement and community inclusion interventions have been applied. We are interested in a wide range of output indicators: teacher competence, head teacher effectiveness, school development planning, school inclusiveness, and the functionality and inclusiveness of SBMCs. Some of these same indicators are also combined to give an overall indicator of school quality. Finally, ESSPIN's impact is measured in terms of improved pupil learning outcomes, which we ascertain through test scores in numeracy and English literacy at grade 2 and 4. For each of these indicators we present in the following chapter three main types of analysis:

1. **Change over time** between CS1 and CS2, for ESSPIN states as a whole. These changes likely reflect changes that are beyond the control of ESSPIN. Although the recent large-scale roll-out of ESSPIN interventions has meant that the programme now has direct links with a very large number of schools in the six states, much of this roll-out happened in 2013/14, and so is unlikely to have started having a major impact by the time of our survey, near the end of the 2013/14 school year.
2. **Differences between ESSPIN and non-ESSPIN schools** within the CS2 results. In the group of schools that are 'expected to be better' we hypothesise that our output, outcome and impact measures will all be higher than in the control group. If this is the case, it provides good initial evidence that ESSPIN is effective, although it does not rule out the possibility that ESSPIN

² They did receive school visits but not any additional training. See Annex C.

³ Three to six selected teachers within each school attended workshops delivered by SSOs. In some states the same group of teachers continued to receive training year after year, while in other cases attempts were made to spread the training to teachers who had not yet received any. However, teachers in ESSPIN schools are also expected to receive more support through other channels, and particularly through professional development meetings organised by the head teacher (RTI International, 2014; and personal communications from ESSPIN). We distinguish the teachers who received direct training ('ESSPIN trained') from those who were not themselves directly trained, but are in ESSPIN schools and so are expected to have received support from their head teachers and colleagues ('ESSPIN schools').

schools' better results could come from differences in school and pupil background characteristics pre-dating the ESSPIN intervention.

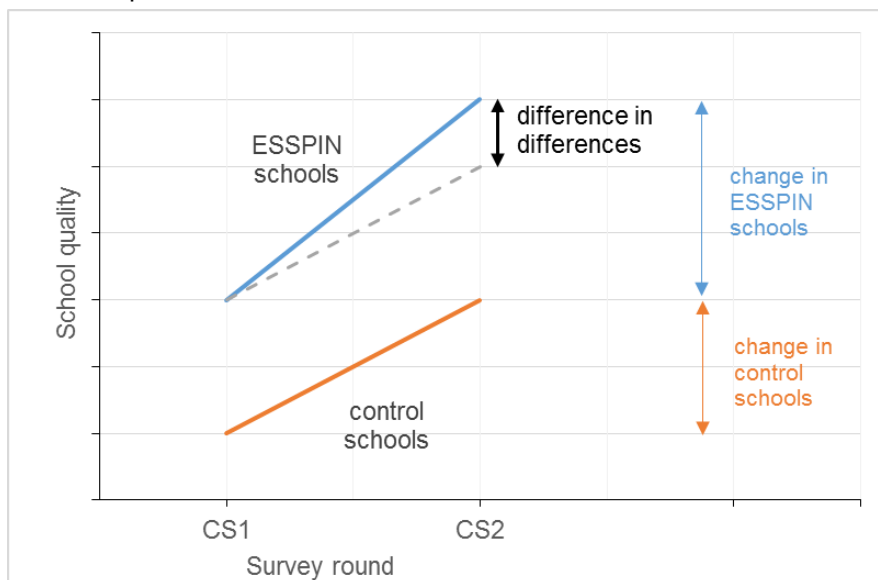
3. Difference in the differences between ESSPIN and non-ESSPIN schools and over CS1 and CS2. See Box 2 below.

In each case we use statistical significance tests (t-tests or z-tests) to give an indication of whether a difference in results (over time or between intervention groups) is significant (i.e. unlikely to have arisen by chance). This should not be taken as constituting rigorous hypothesis testing (given the large number of indicators tested) but it does provide a guide as to whether a difference between the weighted average results in two groups is large enough, relative to the variance of the results, to be able to provide us with a useful indication of likely differences in the population of schools in the six states.

As noted in Box 2, even using a difference in differences analysis does not entirely assure us that the issue of confounding has been taken care of. Chapter 4 establishes that there are significant differences in some of the states between the characteristics of ESSPIN and non-ESSPIN schools. It controls for these, using a number of techniques that match comparable schools in the ESSPIN and control groups prior to analysing pupil test results.

Box 2. Difference in differences

The Composite Surveys may reveal that ESSPIN schools are of better quality, or have better learning outcomes, than other schools, but how do we know whether this can be attributed to ESSPIN and is not just because ESSPIN schools were better in the first place? One way is to focus on change over time using difference in differences methods. The underlying idea of these methods is that schools that have had ESSPIN interventions between CS1 and CS2 – that is, between 2012 and 2014 – ought to have *improved faster* during that period than schools that did not have ESSPIN interventions. We can measure this by comparing averages of the indicator of interest – school quality, say – during CS1 and CS2, in control schools and ESSPIN schools. Is the change over time greater in the ESSPIN schools than in non-ESSPIN schools? If so – and if statistical tests confirm that this result is unlikely to have occurred by chance – then this is considered good evidence that ESSPIN itself had an effect and was not just lucky in choosing schools that were good in the first place.



Sometimes we want to use all of the available information and compare schools which have had *more* or *less* ESSPIN intervention – a continuous scale – rather than dividing them into *some* or *none*. In this case we can use regression analysis – a statistical process for estimating relationships among variables. We model the outcome indicator as depending on time (the round of the survey, CS2 versus CS1), the intensity of intervention, and a *treatment effect*, which is the interaction between time and intensity of intervention. The treatment effect tells us if an increase in the level of intervention increased the speed at which the outcome improved. Regression results are reported as a series of coefficients – numbers representing the strength of the relationship with the outcome of interest.

Coefficient	Meaning of coefficient if positive and significant
Time (CS2 v. CS1)	The outcome improved over time
Intervention	The higher the level of intervention the more effective (regardless of change over time)
Treatment	The higher the level of intervention, the more or faster the outcome improved over time – this is our key indicator of success

Does a significant difference in differences (or treatment effect) prove that the faster improvement in some schools can be attributed to ESSPIN? Not absolutely: it is still possible that there are other factors at play causing faster improvement in some schools than others. For this reason, in chapter 4 we use other statistical techniques to examine whether ESSPIN schools had different characteristics to start with, and to control for any such differences.

2.2 Sample and weights

2.2.1 Sample design

The aim of the sample design for CS2 was to allow follow-up on schools already sampled in CS1, and to allow inferences to be drawn about what is happening in the population of schools across the six states, and within each state, through the use of sample weights. In practice this meant tracking all of the schools sampled in CS1⁴, and adding some additional schools in some of the states. The sample design prioritised the ability to draw conclusions across the six states, conceding that it would not always be possible to obtain statistically significant estimates within each state, given a high degree of variability in the types of schools that are found in some of the states, which makes it difficult to construct a representative sample. The sampling design also incorporated the key aims of the study – to analyse change over time (between CS1 and CS2) and differences between ESSPIN and non-ESSPIN schools.

A stratified sampling design was used for CS1. The survey covered primary schools in the six ESSPIN states – Enugu, Jigawa, Kaduna, Kano and Lagos. The sampling frame was compiled using the annual school censuses, with stratification by ESSPIN phase and whether or not the schools had participated in an earlier attempt to measure learning outcomes, the 2010 Measurement of Learning Achievement (MLA) exercise. The sample design took account of MLA participation in order to allow comparisons between that 2010 round of data collection and the CS1.

Stratifying by MLA participation, however, resulted in high variability in sampling weights and contributed to problems in interpreting some of the results in CS1. There were, in any case, concerns about data quality in the 2010 MLA, so the attempt to compare with the 2010 data was dropped for CS2. The CS2 sample design (Megill, 2014b) also recommended increasing the sample size in two of the states (Kano and Kaduna) in order to reduce problems of high variability in weighted results, and in a third state (Enugu) for which fewer schools were sampled in CS1, to raise it to the same level as that of the other states.

The effective sample from each state is reported in Table 5. For explicit sampling stratification in CS2, schools were divided into four categories that represented the reality of how much ESSPIN output 3 intervention they had received: no intervention ('none'); less than two years of full and continuous intervention ('minimum'); two years of full and continuous intervention ('medium'); and more than two years of full and continuous intervention ('maximum'). Full intervention refers to the full package of leadership training, teacher training, and school visits. Continuous excludes cases where there was an interruption of one or more years in the delivery of interventions. Full information on the output 3 interventions is given in Annex C.

Medium or maximum schools are *expected to be better* in 2014 than none or minimum schools. The minimum schools only received intervention in 2013/2014, and so are not expected to be better than the schools that were receiving no intervention by the time of the survey in May/June 2014.

Maximum schools are also *expected to have improved faster* than schools that received no intervention between 2012 and 2014. For medium schools, some of them are expected to have improved faster, while others are not, depending on the timing of the intervention.

⁴ A few schools were not eligible for the survey because they were special schools or non-government Islamic schools, but which were erroneously sampled for CS1. These were excluded in CS2.

Table 5. Sample in CS1 and CS2 and population of schools, by state and with intervention groups

	Category for sampling purposes	CS1 sample (2012)	CS2 sample (2014)	Population	Categories for analysis	
					Expected to be better in 2014	Expected to have improved during 2012–2014
Enugu	None/minimum	35	70	1220	No	No
	Medium	35	35	272	Yes	Yes
Jigawa	None/minimum	32	32	1569	No	No
	Medium (1)	36	36	303	Yes	Yes
	Medium (2)	35	35	198	Yes	No
Kaduna	None/minimum	28	61	3413	No	No
	Medium	42	42	671	Yes	Yes
	Maximum	35	37	165	Yes	Yes
Kano	None/minimum	67	135	5238	No	No
	Medium	35	35	317	Yes	No
Kwara	Medium	102	105	1485	Yes	No
Lagos	Medium	69	69	1001	Yes	Yes
	Maximum	34	36	100	Yes	Yes

Note. The sample size shown is the actual sample for which data was collected. See section 2.2.3 on sample coverage.

Teachers were randomly sampled within selected schools in both CS1 and CS2. Following another recommendation from the sampling report, it was decided to reduce the number of teachers sampled per school, from 10 in CS1 to six in CS2, partly because sampling 10 teachers per school resulted in difficulties in detecting significant effects in some states⁵, and also because there are many schools which have fewer than 10 teachers. Although it would have been useful to have tracked individual teachers, this was not seen as feasible given the limited collection of identifying information during CS1.

Teachers within each school were sampled from the population present in the school on the day of the survey visit and who taught grades 1–6 in the present term, using the school's teacher attendance register. Fieldwork teams asked head teachers to complete such a register in cases where this had not already been done for the day of the visit. Team supervisors entered the number of eligible teachers into the computer-assisted personal interviews (CAPI) system, which then randomly selected six to be sampled.

Pupils were sampled from the pupil registers for grades 2 and 4 classes. The sample size of four pupils per test per school remained the same for both CS1 and CS2. In cases where the registers were not available, or were found to be inaccurate, data collectors went to the classrooms and counted pupils instead. Again, the CAPI system randomly assigned pupils to be sampled for each test (grade 2 literacy, grade 2 numeracy, grade 4 literacy and grade 4 numeracy) from the numbers present in grades 2 and 4. As with teachers, it would have been useful to trace the same pupils over time, but this was not seen as feasible because, for the children sampled in CS1, we only

⁵ In complex survey designs, clustering of results can make it difficult to obtain precise estimates. A high degree of similarity among teachers *within* schools, combined with a high degree of variation *between* schools, resulted in clustering effects in the teacher indicators. This, along with the type of sampling weights that were applied, resulted in high 'design effects' and a loss of precision in the CS1 analysis for some states. See ESSPIN (2013a) and Megill (2014a) for a full explanation.

have their names, which is not always sufficient information to identify the same children two years later. We therefore collected a random sample within each school in both CS1 and CS2.

In addition to the main sample of 16 pupils and six teachers, an additional four pupils and two teachers were selected in each school by the CAPI system as 'replacements'. Replacements were included in the survey in cases where teachers and pupils from the main sample turned out not to be available at the school, despite having been recorded as present in the register. Replacements could not be used in any other circumstances, however. In practice the option to replace was used very rarely.

A number of schools were found to operate double shifts, with some classes taught in the morning and others in the afternoon. Where necessary, the previous afternoon's register was added to the morning register for the day of the visit, in order to ensure that teachers and pupils were included whether they attended the morning or afternoon session.

2.2.2 Weights

Simple averages of the results from the Composite Survey data would not be representative of what is happening across the state, because (as Table 5 above shows) the profile of schools included in the survey is not identical to the profile of schools in the state as a whole. We overcome this by applying sample weights which give greater weight to the results in schools that are relatively under-represented in the survey. Sample weights were calculated for the CS1 and CS2 schools, teachers, and pupils. A smoothing technique was also applied to reduce the variability of the weights and avoid the design effects problem encountered in the CS1 analysis (see Megill, 2014b).

Most of the following analysis applies weights to sample statistics calculated within each round and intervention group, which can then be used as estimates of the whole population of schools in the six ESSPIN states. However, part of the analysis compares change within individual schools. For this we cannot use the additional CS2 schools and are limited to the original set of schools sampled for CS1. An additional set of weights was calculated for use with this 'panel' of schools present in both CS1 and CS2.

2.2.3 Sample coverage

The intended sample for CS2 consisted of 735 schools (Table 6). During fieldwork, six schools in the sample were found to be ineligible for the survey, either because they were purely private Quranic schools with no non-religious education, or (in one case) because it was a special school for deaf students using different methods from regular government schools. One school (in Jigawa) was not visited because of security concerns.

Within the schools, it was not always possible to administer all of the intended instruments. In some cases this was because the school was very small, and lacked a sufficient number of pupils and eligible teachers. It also sometimes happened that teachers and pupils were not present at 8am, when sampling was conducted; and occasionally pupils and teachers left the school after being sampled (for example, due to illness). Overall, complete test data was gathered in CS2 for 95% of the maximum possible number of students, a similar proportion to CS1. Interviews were gathered and lesson observations conducted for 80% and 78%, respectively, of the maximum possible number of teachers. This figure was relatively low due to there being fewer than six teachers in a large number of schools.

As explained in the following section, teacher content knowledge tests were administered separately from the main survey, in testing centres. This results in some attrition, with around 10% of the sampled teachers not found in the tests, and some possible selection bias in the test results. This is discussed in section 3.1.2 below.

When we present results from CS1, they differ somewhat from those reported in the original report on CS1 (ESSPIN, 2013a). This is the result of the adjustment to the CS1 weights, along with some schools in Kano and Jigawa being dropped from the data because they were found to have been ineligible (see footnote 4 above). Although some of our estimates are changed, the overall pattern of results remains the same.

Table 6. Sample coverage in CS2

	Schools		Teachers			Pupil tests			
	Intended sample	Actual	Interview	Less. obs	Tests	L2	L4	N2	N4
Enugu	105	105	532	519	494	388	395	384	382
Jigawa	105	103	430	425	415	393	399	396	398
Kaduna	140	140	638	625	586	539	521	538	517
Kano	175	170	773	764	645	667	635	669	632
Kwara	105	105	545	538	483	394	393	397	391
Lagos	105	105	569	545	526	415	410	413	409
Total	735	728	3487	3416	3150	2796	2753	2797	2729

Note. In this table and throughout this report, L2 refers to the grade 2 literacy test, L4 to the grade 4 literacy test, N2 to the grade 2 numeracy test, and N4 to the grade 4 numeracy test.

2.3 Training, pilots and fieldwork model

Fieldwork for CS2, including the pupil tests, was conducted using CAPI during May–July 2014. Children were given a printed pupil book to read and write in. The interviewers made use of a tablet computer, which prompted them on the questions to be asked orally to the children, gave instructions on the administration of the different test items, including timing, and allowed them to input whether each part of each question was answered correctly or incorrectly (or not attempted at all) by the pupil.

The instruments were piloted during March 2014 after state coordinators and monitoring officers had been trained. Pilots were conducted in Nassarawa, Kaduna and Abuja, with the coordinators and monitoring officers collecting the data, using both paper and CAPI. In total there were eight pilot days, with additional piloting of revised pupil tests in Abuja. Instruments were revised through consultation with ESSPIN and state coordinators.

Table 7 lists the instruments used in CS2, together with the indicators relevant to outcomes, outputs or impact that were gathered from each instrument. The instruments were also used to gather intervention information, such as whether individual teachers had received ESSPIN training or not, and pupil-level information on socio-economic status, age, language spoken at home, and gender. The data gathered in general allows more detailed analysis than that presented in this report, some of which is presented in the six state-level reports and the Gender and Inclusion Report that will accompany this report. The data will also be published in anonymised form for use by ESSPIN and other researchers.

The process of revising instruments for CS2 does leave some possibility of measurement error in comparisons between CS1 and CS2. Given that training and fieldwork were extremely challenging

for CS1, the priority for CS2 was to ensure consistent and manageable data collection within CS2, by setting clearer guidance for data collectors through detailed data collection manuals, greater oversight, and through a single intensive training session for all data collectors across the six states. Although we avoided large changes in instruments that would compromise comparability with CS1, any change in questionnaire format or wording, training, and data collection procedures can potentially affect the results, and this should be kept in mind. However, the change in instruments should not have affected our difference in differences results, provided that changes in measurement are consistent across ESSPIN and non-ESSPIN schools.

Table 7. Instruments used in CS2

Instrument	Outcome / output / impact indicators
Structured interview with head teacher	Number of lesson observations during past two weeks; number of professional development meetings this school year; teacher attendance book; actions by head teacher to promote teacher attendance and improve pupil attendance; written evidence of school self-evaluation process for school year; School Development Plan (SDP) for school year available; activities relating to strengthening teaching and learning in the SDP; activities relating to improving access in the SDP; evidence of activities in the SDP being carried out; up-to-date cashbook.
Structured interview with SBMC chairperson and members	Number of SBMC meetings this school year; SBMC awareness-raising activities; steps taken by SBMC to address exclusion; SBMC networking with community-based organisations (CBOs), traditional or religious institutions, other SBMCs, and LGEAs; SBMC has a women's committee and a children's committee, and how often these committees meet; SBMC has contributed resources to the school; visits by the SBMC to the school this school year; number of SBMC meetings attended by at least one woman and by at least one child; issues raised by female and child members; action taken on issues raised by female and child members; whether children's committee had a trained facilitator; action for commonly excluded groups; SBMC raised issue of children's exclusion.
Structured interview with teacher	Knowledge of English and maths curriculum benchmarks; school opening time.
Lesson observation	Number of forms of classroom organisation used; number of teaching aids used; number of times teacher praised or reprimanded children; participation of children from different zones of the classroom; participation of boys and girls in the lesson.
Teacher tests conducted at the end of the survey in testing centres	Teacher test scores in English literacy and numeracy.
Pupil tests	Pupil test scores in English literacy and numeracy at grades 2 and 4.
General observation	Length of morning break; number of classes where pupils and teachers are in class within half an hour of starting time.

3 Findings

3.1 Teacher competence

The ESSPIN logframe sets four criteria for judging the competence of teachers (Box 3). A teacher who teaches English or maths is defined as competent if he or she meets at least three of these, while teachers of other subjects are exempted from one of the four criteria (knowledge of the English or maths curriculum) and defined as competent if they meet two of the remaining three criteria.

For CS2, a fifth criterion was added, based on teacher content knowledge test results. Teachers are defined as competent if they are competent according to the original criteria, and can also score at least 50% in primary school level literacy and numeracy tests.

Box 3. Logframe standard for teacher competence

A teacher must meet three out of four of the following criteria to meet the competence standard if he/she teaches English and/or maths. Teachers of other subjects must meet two out of three criteria (excluding 1 below).

- 1) Knowledge of English or mathematics curriculum (based on interview)
- 2) Use of at least one teaching aid during lesson observation
- 3) Greater use of praise than reprimands during lesson observation
- 4) Class organisation: assigning individual or group tasks at least twice during lesson observation (or for two contiguous five-minute blocks)

For CS2, a new stricter indicator of teacher competence has been introduced. This excludes reading from or writing on, or having pupils copy from, the blackboard as use of a teaching aid, and adds a fifth criterion:

- 5) English and mathematics content knowledge: scores at least 50% in both an English literacy and a numeracy test

3.1.1 Main analysis

Table 8 compares the results across the six ESSPIN states in 2012 and 2014. (The fifth criterion is listed here as teacher tests were not conducted in 2012.) Use of teaching aids and use of praise more than reprimands during classes have both improved. English and mathematics teachers' knowledge of curriculum benchmarks appears to have declined dramatically, from 57% to 35%, although there is some possibility of measurement error in this indicator.⁶ Assignment of individual and group tasks did not change significantly. The proportion of teachers meeting the overall competence standard, and the average value of our overall competence score – a score from 0 to 100 based on the extent to which teachers met each of the four criteria (with equal weight given to each) – also did not change significantly.

⁶ CS2 introduced clearer guidance about which grade of the curriculum teachers should be quizzed on, in order to improve consistency within the CS2 data. In addition, CS1 fieldwork in each school was spread over several days, giving teachers an opportunity to revise their knowledge of curricula guidelines. In CS2, fieldwork in each school was conducted on a single day.

Table 8. Teacher competence in 2012 and 2014

	2012	2014	
(1) Knowledge of English/maths curriculum	56.9	34.6	-
(2) Use of one or more teaching aid	88.1	96.3	+
(3) Praise more than reprimand	70.4	80.2	+
(4) Assigns two or more individual or group tasks	56.1	53	
Competence score (old version)	69.8	68.4	
Teacher competence standard (old version)	69.7	66.1	

Note. + = significant improvement between 2012 and 2014; - = significant worsening between 2012 and 2014 (using a t-test; $p < .05$)

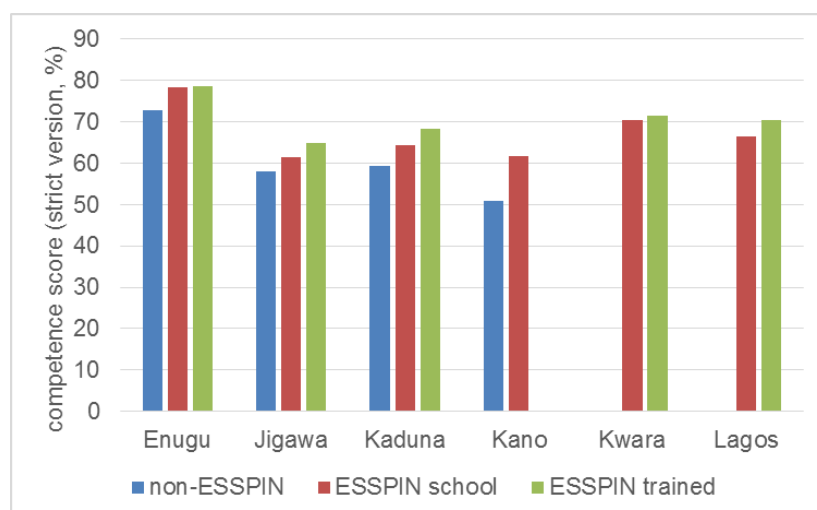
In regard to the findings in 2014, there is a clear pattern of teachers performing better in ESSPIN schools than in non-ESSPIN schools (Table 9). We distinguish between three groups of teachers: (i) those who are in schools that received no ESSPIN intervention; (ii) those who are in schools that received ESSPIN intervention but who did not individually receive ESSPIN teacher training; and (iii) those who are in ESSPIN schools and individually received ESSPIN teacher training. For four of the five criteria, teachers in ESSPIN schools are significantly more likely to meet them than those in non-ESSPIN schools, and the difference is larger still for teachers who individually received ESSPIN training. Teachers in ESSPIN schools, and who received ESSPIN training, also did better in English and mathematics tests, although the overall scores are still rather low. Fewer than one-third of teachers in non-ESSPIN schools passed the English and mathematics tests, compared to 48% in ESSPIN schools and 62% of those who were individually trained by ESSPIN. ESSPIN-trained teachers were more likely to meet the competence standards and their competence scores were significantly higher.

However, some of the differences shown in Table 9 are the result of two states (Kwara and Lagos) where there are no non-ESSPIN schools and where relatively high proportions of teachers meet the competence standards. Breaking down the results by state (Figure 4), the differences within each state between the three intervention groups are small in magnitude but uniform in direction: ESSPIN-trained teachers are more competent than teachers who are in ESSPIN schools but who do not directly receive training; and, in turn, both are more competent than teachers in non-ESSPIN schools.

Table 9. Teacher competence in 2014, by intervention group

	(i) Non-ESSPIN	(ii) ESSPIN school		(iii) ESSPIN-trained	
(1) Knowledge of English / maths curriculum	29.0	44.5	+	39.6	
(2a) Use of one or more teaching aid	95.2	97.9		98.3	
(2b) Use of one or more teaching aid excluding reading/writing on/copying from blackboard	69.8	79.8	+	89.9	+
(3) Praise more than reprimand	75.0	84.4	+	95.2	+
(4) Assigns two or more individual / group tasks	49.9	59.4		56.6	
English score (%)	41.8	49.7	+	55.1	+
Mathematics score (%)	56.5	66.5	+	71.5	+
(5) Passes English and mathematics test	32.7	47.8	+	61.6	+
Competence score (old version)	65.4	73.9	+	72.8	+
Teacher competence standard (old version)	61.8	74.9	+	71.5	+
Competence score (new version)	57.2	66.7	+	69.5	+
Teacher competence standard (new version)	19.5	32.8	+	42.6	+

Note. The CS2 version of the competence score adds the teacher's performance in the literacy and numeracy tests to the number of other criteria met by the teacher. For example, a teacher who met all four original criteria and also scored 100% in the literacy and numeracy tests would receive a competency score of 100%. + indicates a significant difference from the results in non-ESSPIN schools ($p < .05$)

Figure 4. Teacher competence score (strict version), by state and ESSPIN group

Note. Very few teachers in Kano said they had directly received ESSPIN training and so we merge this group with the ESSPIN school group. There were no non-ESSPIN schools in Kwara or Lagos.

Did teachers who benefited from ESSPIN interventions between 2012 and 2014 improve faster than those who did not? We address this question in two ways: firstly, by comparing the difference in mean competency scores between CS1 and CS2, within each intervention group; and secondly by using regression analysis (see Box 2 above). The comparison of means (Table 10) suggests that teachers who are in ESSPIN schools, but were not individually trained by ESSPIN, improved their competency scores significantly more than those in non-ESSPIN schools. In fact, those in non-ESSPIN schools became, on average, less competent between CS1 and CS2, while those in ESSPIN schools became slightly more competent. However, no such significant difference was

found in the case of teachers who individually received ESSPIN training. Among this category of teachers, competence scores appear to have worsened between 2012 and 2014, although the change is not statistically significant. (Note that the teachers interviewed in 2014 were not necessarily the same as those interviewed in 2012, and training was rolled out to a larger number of teachers during 2012–2014.)

Table 10. Teacher competence difference in differences (comparison of means)

Teacher competence scores (CS1 version)	(i) Non-ESSPIN	(ii) ESSPIN school		(iii) ESSPIN-trained
CS1	67.3	67.8		79.5
CS2	65.4	73.9		72.8
Difference	-1.9	6.0	*	-6.7

* indicates a significantly different difference compared to that in non-ESSPIN schools ($p < .05$)

An alternative method involves using regression analysis to examine the same question (Table 11). We model the outcome indicator (competence score) as depending upon time (the round of the survey) and the intensity of intervention. Intensity of intervention is measured either as the number of years of full ESSPIN output 3 intervention package (columns labelled ‘school improvement’), or more specifically as the amount of teacher training delivered to the school (columns labelled ‘training’). We also use an alternative intervention measure that adjusts for the length of time a teacher has been in his or her present school. A teacher who only joined the school in 2012, for example, cannot be expected to have benefited from ESSPIN training delivered in 2010 or 2011, and the intervention variable can be adjusted to reflect this.

The interaction effect between intervention and time, labelled ‘treatment’, if significant, would provide evidence that schools with more ESSPIN intervention improved more rapidly between 2012 and 2014. Time effects are mixed and not significant, consistent with the finding in Table 8 above that there is no significant change over time in teacher competence. Intervention effects are mostly positive and significant, confirming the finding that teachers who benefit from more ESSPIN intervention appear to be more competent than those receiving less ESSPIN intervention. Treatment effects for school improvement are not significant, suggesting that teachers whose schools benefited from more ESSPIN intervention did not generally improve faster than those whose schools had less ESSPIN intervention. The treatment effect for teachers who individually received ESSPIN training is not significant when no adjustment is made for the time when the teacher started teaching in the school, but is significant when this adjustment is made.

Table 11. Teacher competence difference in differences (regression)

Regression on competence scores (CS1 version)		Non-adjusted				Adjusted			
Intervention variable		School improvement		Training		School improvement		Training	
Time (CS2 v. CS1)	Coefficient	-0.8		-0.2		0.6		-1.8	
	Standard error (SE)	2.2		2.3		2.1		1.9	
Intervention	Coefficient	2.2	*	1.4	*	2.8	*	-1.2	
	SE	0.7		0.6		0.8		0	
Treatment	Coefficient	-0.1		-1.4		-0.9		1.3	*
	SE	0.8		0.9		0.8		0.6	
	sample	1263		1263		1102		1102	

* indicates a significant coefficient ($p < .05$). School improvement refers to the expected change in school quality resulting from ESSPIN output 3 intervention. Training refers to the number of days of teacher training provided during the relevant period. The 'adjusted' results take into account the length of time that the teacher has been working in his or her current school in calculating the amount of school improvement or training that they have been exposed to, while the 'non-adjusted' results ignore this.

3.1.2 Findings from the teacher content knowledge tests

The teacher tests included items pitched at different primary school grades and focusing on different areas: foundational skills for teaching literacy; writing; reading; grammar; number concepts; calculation; and other numeracy skills (Box 4). A breakdown of the items in the teacher tests helps us to understand the areas of strength and weakness for teachers in ESSPIN and non-ESSPIN schools (Table 12). Many of the English literacy test items appeared to present particular difficulties for the teachers, while relatively high proportions were able to answer the mathematics questions.

Within English, teachers appeared to struggle most with writing, and with letter–sound correspondence and word building, skills considered by ESSPIN to be foundational for teaching English literacy. Low writing scores may reflect fairly stringent marking guidelines. For example in the grade 5 writing question shown in Box 4 ('Think about a journey...'), each sentence needed a capital letter, full stop, and no more than one spelling mistake, for the answer as a whole to be marked correctly. Only 8% of teachers were able to reach this question, which appeared towards the end of the test, and to answer it correctly. Low scores in foundational literacy items may reflect a lack of familiarity with these items, especially among teachers not trained by ESSPIN. In non-ESSPIN schools the average score in these items was only 35%, while ESSPIN-trained teachers scored 39% (across Enugu, Jigawa, Kaduna and Kano) and 53% (across Lagos and Kwara).

Teachers' scores were higher in reading, number concepts and calculation, suggesting that they are familiar with these types of questions and comfortable with the required skills, although there remains a minority who are unable to answer these questions. Over 80% of teachers could write number lines and add three-digit numbers, for example.

Results across all types of item were stronger for ESSPIN-trained teachers compared to teachers in non-ESSPIN schools, although the difference was only large enough to be statistically significant for foundational literacy skills, grammar, and number concepts.

Box 4. What was in the teacher content knowledge tests?

Teachers were given written tests in English and mathematics, lasting one hour each. Some example items are listed below. The tests each included 40 items, with eight items pitched at each primary school grade level from 1 to 5. In English, within each grade level there were two items focusing on each of four areas: foundational literacy skills, writing, reading, and grammar. In mathematics, within each grade level there were three items on number concepts, three items on calculation, and two items on miscellaneous numeracy skills, such as measurement, shapes and statistics. Although some questions have multiple parts, the marking scheme allocates either one point if all parts are correct, or zero points otherwise, for each question; no partial credit is allowed. Some spelling mistakes are allowed in the English test but most other types of mistake result in a mark of zero.




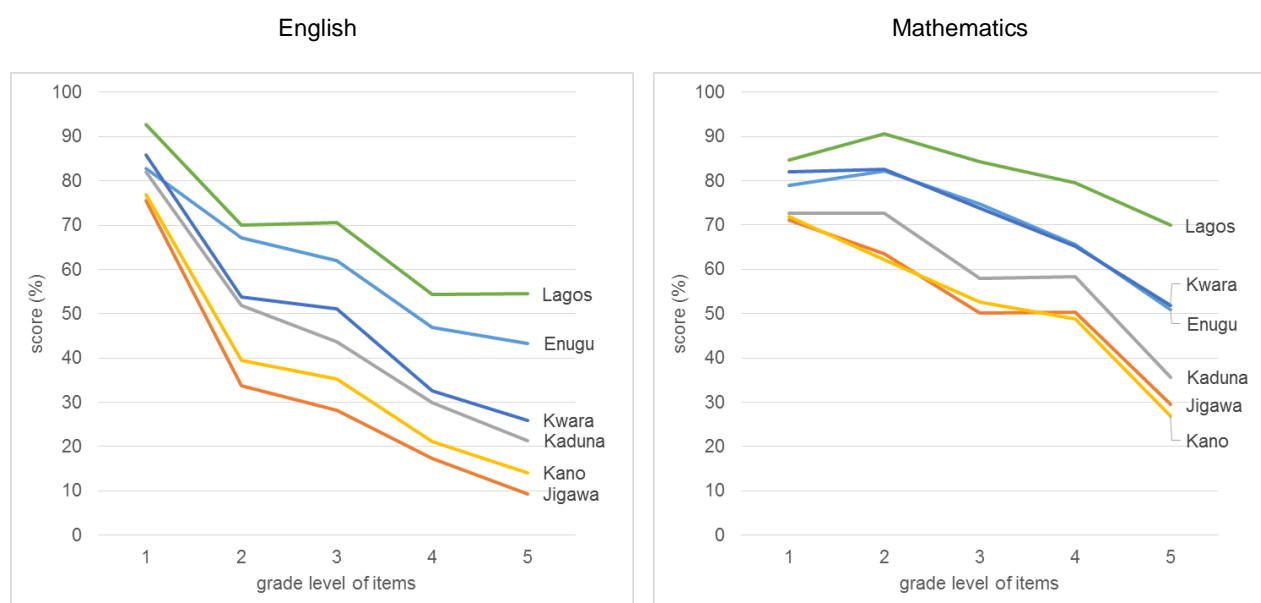
Item type	Question	% of teachers marked correct
Grade 3 literacy foundational skills	Write the number of sounds in each word. The first one has been done for you: steep = s-t-ee-p = 4 cloud = head = wish = August =	16%
Grade 5 writing	Think about a journey you have made. Write three sentences about it. Explain when you travelled, where you travelled to and what happened on the way there.	8%
Grade 2 reading	Read this story, then answer the questions: Ali is playing football. The teacher is blowing the whistle. Ali has no shoes on but he wears shorts when he plays football. The teacher wears a white shirt and a pair of shorts too. Last week Ali's team scored four goals. What is the teacher doing? How many goals did Ali's team score last week?	83%
Grade 1 grammar	Here is one bird. Complete the sentence: Here are two ____  	84%
Grade 1 number concepts	Write in the missing numbers on this number line: 	93%
Grade 3 calculation	Calculate 236 + 480 = ____ You may use this space to work out the answer.	87%
Grade 5 other numeracy	Match the correct unit to each label. The first one has been done for you. centimetre cm volume millilitre ml weight gramme g short lengths metre m longer lengths	45%

Table 12. Average scores in different types of item in the teacher tests (%)

		Enugu, Jigawa, Kaduna, Kano					Lagos, Kwara		
		Non-ESSPIN (i)	ESSPIN school (ii)		ESSPIN-trained (iii)		ESSPIN school (ii)	ESSPIN-trained (iii)	
English	Grade 1	78.1	82.5		82.8	+	86.8	90.3	+
	Grade 2	45.5	49.0		47.9		60.5	60.5	
	Grade 3	39.8	42.5		46.1	+	56.6	61.6	+
	Grade 4	26.2	28.1		30.0		39.7	43.4	
	Grade 5	19.3	20.3		22.5		36.9	38.7	
	Foundational	35.0	37.3		39.4	+	49.6	53.0	
	Reading	55.4	60.4		59.9		72.5	75.4	
	Writing	28.6	28.9		30.2		35.4	38.6	
	Grammar	48.3	51.3		53.8	+	66.9	68.6	
Mathematics	Grade 1	72.4	74.8		76.7	+	82.7	83.5	
	Grade 2	67.6	70.7		71.6		84.4	87.2	
	Grade 3	56.8	56.8		59.8		77.9	78.5	
	Grade 4	53.3	56.8		58.6	+	69.4	72.7	
	Grade 5	32.6	36.4		33.6		59.5	59.4	
	Number concepts	62.6	67.4	+	67.8	+	81.1	82.2	
	Calculation	53.0	54.1		56.6		71.7	73.0	
	Other	52.8	54.1		53.6		70.0	72.2	

Note. There is no control (non-ESSPIN) group in Lagos and Kwara, so the results are shown separately. + indicates a statistically significant difference ($p < .05$). The comparison group is non-ESSPIN schools for ESSPIN schools and ESSPIN-trained teachers in Enugu, Jigawa, Kaduna and Kano. The comparison group is teachers who are in ESSPIN schools but who were not individually trained by ESSPIN in Lagos and Kwara.

There is a steep drop in test scores as the test progresses from lower grade to higher grade items. Around 80% of teachers can answer grade 1 level English questions correctly, but only 20% can answer grade 5 level questions correctly. In mathematics, the declining gradient of scores with the grade level of test items is less steep, but it is still evident: teachers score 72% in grade 1 but only 33% in grade 5 level items. Moreover, these patterns vary dramatically by state (Figure 5). Only in Lagos and Enugu is there a relative consistency across grade levels in English, with teachers scoring over 40% even in grade 5 level items. In Jigawa teachers scored on average 34% in grade 2 English items, and only 9% in grade 5 English. In mathematics, most teachers across the six states can answer questions up to grade 4 level, but there is a marked drop-off when it comes to the hardest grade 5 level items, especially in Jigawa, Kaduna and Kano.

Figure 5. Teacher test scores by grade level and state

As noted in section 2.2.3, there is a possibility of bias in the test results because 10% of the selected teachers did not attend. The teachers who attended the tests were somewhat better qualified on average than those who did not, suggesting that there may indeed have been some self-selection, with some lower-qualified teachers avoiding the test (Table 13). However, teachers who sat the test did not score significantly higher in the other competence criteria than those who did not sit it. We also tested whether this bias was greater in ESSPIN or non-ESSPIN schools, and found no evidence of any difference.⁷ This suggests that overall the estimates for teacher scores presented above are over-estimates of the true level among the population of teachers, but that the difference in test scores between ESSPIN and non-ESSPIN schools is unlikely to be affected by this selection issue.

Table 13. Academic qualifications of teachers according to whether they sat English and mathematics tests in CS2

%	Did not sit tests	Sat tests
Less than SSCE	25.9	18.5
SSCE/WASC	5.9	4.3
OND / diploma	8.2	6.9
NCE	56.6	67.6
BA / BSc / HND / LLB	3.4	2.7
Total	100	100

SSCE: Senior Secondary Certificate of Education; WASC: West African Senior School Certificate; OND: Ordinary National Diploma; NCE: National Certificate of Education; BA: Bachelor of Arts; BSc: Bachelor of Science; HND: Higher National Diploma; LLB: Bachelor of Laws.

⁷ We estimate a logistic regression model where the likelihood of sitting the test depends on whether the teacher is in an ESSPIN school and on his or her academic qualification, with an interaction term between these two variables. The coefficient on the interaction term is not significant, suggesting that the relationship between likelihood of sitting the test and highest academic qualification is independent of whether the teachers come from ESSPIN or non-ESSPIN schools.

3.1.3 Summary and discussion

In summary, the findings regarding teachers suggest no significant change in teacher competence overall between 2012 and 2014. Teachers in the six states improved in terms of use of teaching aids and the frequency with which they praised, rather than reprimanded, students, but English and mathematics teachers appeared to worsen in terms of their ability to recognise curriculum benchmarks. However, teachers in ESSPIN schools, and especially those who individually received training from ESSPIN, do better in most criteria than those in non-ESSPIN schools. This is consistent with a positive impact of ESSPIN on teacher competence, although we cannot rule out confounding factors in identifying a causal link: teachers who received ESSPIN training may have been better than others prior to receiving ESSPIN training. Indeed, this result is not new: this was also the case at the time of the first round of the survey. Breaking the results down by state, we still find differences in the expected direction, although they are smaller in magnitude.

These two findings – lack of a marked improvement over time overall, and large difference between ESSPIN and non-ESSPIN schools – can be reconciled by remembering that ESSPIN intervention reached some 16% of schools in 2012/13, most of these in Lagos and Kwara. Thus, the relatively good performance of ESSPIN-trained teachers had only a limited impact in pulling up overall state averages during 2012 to 2014. (We consider the further scale-up in 2013/14 too recent to have had an impact by the time of the survey, which was carried out towards the end of the same school year).

There is also evidence that teachers benefiting from ESSPIN interventions improved faster (or at least, worsened more slowly) between 2012 and 2014. However it is not clear whether simply being present in an ESSPIN school is the key factor, or if this difference is associated with teachers being directly trained by ESSPIN; our different methods of analysis give inconsistent results in this regard.

In 2014, 68% of teachers across the six states met our original standard of teacher competence. Only 26% met the new, stricter standard that we have developed for CS2, and which includes a requirement to score 50% or above in English and mathematics content knowledge tests. For many of the teachers not meeting the overall strict competence standard, the reason for this is likely to be because their English content knowledge scores are low, particularly on material pitched at primary grades 4 and 5, and especially in questions that aim to test their writing skills. ESSPIN-trained teachers generally achieve better results in these items than other teachers, but the scores remain under 40%, suggesting scope for additional work in this area.

3.2 Head teacher effectiveness

The ESSPIN logframe defines head teacher effectiveness in terms of seven criteria (Box 5). These reflect both activities carried out by the head teacher and behaviour across the teachers and pupils, such as agreement on what time the school opens (criterion 4), presence in class at the beginning of the school day (criterion 5), and appropriate break and lesson durations (criteria 6 and 7).

Box 5. Logframe standard for head teacher effectiveness

A head teacher must ensure that five out of seven of the following criteria are met in order to meet the head teacher effectiveness standard:

- 1) Carried out two or more lesson observations in the past two weeks
- 2) Held four or more professional development meetings since the start of the 2011/12 or 2013/14 school year (NB: survey took place more than nine months into the school year)
- 3) School has a teacher attendance book and head teacher recalls at least two actions taken to promote teacher attendance
- 4) Clear school opening time: more than 50% of pupils sampled agree on the school opening time and more than 50% of teachers sampled agree on the school opening time
- 5) More than 50% of classes are in their classroom with their teacher within 30 minutes of school opening time
- 6) Length of morning break is 35 minutes or less, except in Enugu when it must be 15 minutes or less
- 7) More than 50% of lessons observed finished within five minutes of a standard 35 minute lesson duration (i.e. between 30 and 40 minutes long)

Overall, head teacher effectiveness has not significantly improved or worsened between 2012 and 2014 (Table 14). More head teachers were carrying out lesson observations and holding professional development meetings in 2014 than in 2012, and a higher proportion of lessons were of an appropriate duration (30–40 minutes), but fewer could demonstrate that they had taken action to promote teacher attendance. Other criteria had not changed significantly, and, overall, in 2014 fewer than 20% of head teachers met the effectiveness standard.

Table 14. Head teacher effectiveness in 2012 and 2014

	2012 (CS1)	2014 (CS2)	
(1) Lesson observations	9.2	20.3	+
(2) Professional development meetings	11.4	20.9	+
(3) Action on teacher attendance	83.1	53.2	-
(4) Clear opening time	49.8	53.9	
(5) In class on time	72.4	62.1	
(6) Appropriate morning break	78.9	75.9	
(7) Appropriate lesson length	30.3	51.3	+
Number of criteria fulfilled (/7)	3.4	3.3	
Effective head teacher (5/7 criteria met)	13.6	19.5	

Note. + = significant improvement between 2012 and 2014; - = significant worsening between 2012 and 2014 (using a t-test; $p < .05$)

Focusing on the 2014 data, head teachers who have received leadership training from ESSPIN are more effective than those who have not (Table 15) in terms of lesson observations, professional development meetings, teachers and pupils being in class on time, and their overall effectiveness score. However, appropriate lesson length was significantly worse among ESSPIN-trained head teachers than others.⁸ A small number of head teachers were in schools that had received the ESSPIN intervention package but did not themselves report having received any ESSPIN training (column ii in Table 15). These head teachers were not significantly more effective than the control group.

⁸ The introduction of 60-minute literacy / numeracy sessions may have affected this result.

Going beyond the seven criteria associated with our head teacher effectiveness standard, we examine whether ESSPIN-trained head teachers are able to attract more qualified teachers to their schools, and whether teacher absenteeism differs between ESSPIN and non-ESSPIN schools (Table 16). Teachers sampled in schools where the head teacher had received ESSPIN training were indeed more qualified than those in other schools, both in terms of academic and teaching qualifications, although some of this difference is likely to pre-date the ESSPIN intervention (see chapter 4). Teacher absenteeism – based on head teachers' own records in the form of a teacher attendance register – appears to be somewhat lower in ESSPIN schools, but the difference is not significant. On an average day around one in four teachers were recorded as being absent from school.

Table 15. Head teacher effectiveness in CS2, by intervention group

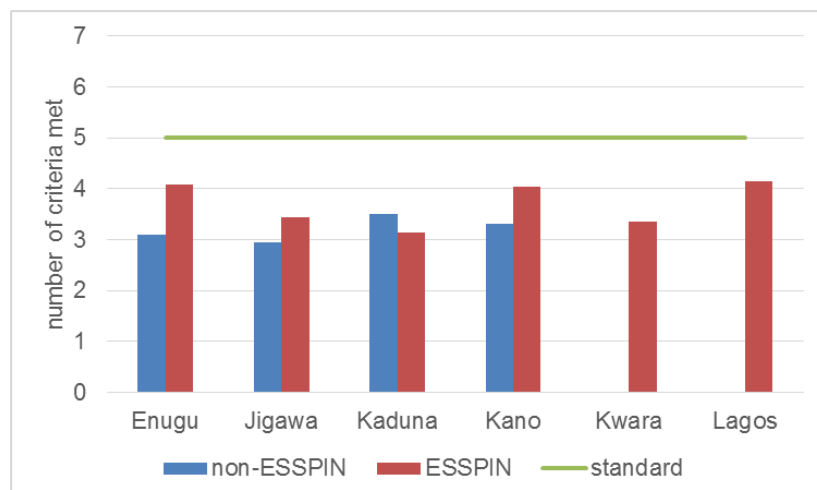
	(i) Non-ESSPIN	(ii) ESSPIN school	(iii) ESSPIN-trained
(1) Lesson observations	14.9	22.9	37.6 +
(2) Professional development meetings	12.3	22.7	48.7 +
(3) Action on teacher attendance	53.0	29.7	59.0
(4) Clear opening time	55.0	37.3	53.8
(5) In class on time	56.6	66.9	79.1 +
(6) Appropriate morning break	76.2	84.1	73.0
(7) Appropriate lesson length	57.1	36.1	36.5 -
Number of criteria fulfilled (/7)	3.2	3.0	3.9 +
Effective head teacher (5/7 criteria met)	14.4	18.8	37.1 +

Note. + indicates a significant difference from the results in non-ESSPIN schools ($p < .05$)

Table 16. Teacher qualifications and absenteeism

	(i) Non-ESSPIN	(ii) ESSPIN school	(iii) ESSPIN-trained
Average academic qualification of teachers	1.6	1.6	2.3 +
Average teaching qualification of teachers	0.8	1.0	1.4 +
Proportion of absent teachers (average over past five days)	30.7	24.2	24.9

Note. The academic qualification score was calculated by assigning each teacher one point if their highest qualification was SSCE/WASC, two for OND / diploma, three for NCE, four for BA / BSc / HND / LLB, and five for MA / MSc. The teaching qualification score was calculated by assigning each teacher one point for NCE, two for grade II or equivalent, 3 for Post-Graduate Diploma in Education (PGDE), four for Bachelor of Education (BEd) or equivalent degree in education, and five for Master in Education (MEd) or equivalent. + indicates a significant difference from the results in non-ESSPIN schools ($p < .05$)

Figure 6. Number of head teacher effectiveness criteria met, by state and intervention group

As in the previous section on teacher competence, we also examined change over time in head teacher effectiveness, to see whether schools that received more ESSPIN intervention between 2012 and 2014, and head teachers who received more leadership training, improved faster than comparators. The results suggest that, as in the previous analysis, there was little overall change in head teacher effectiveness. However, the trend in ESSPIN schools was significantly more positive compared to the change in non-ESSPIN schools (Table 17).⁹ Using an alternative type of analysis of the same question – regression analysis with continuous intervention variables – school improvement is again associated with a significant treatment effect, although we found no such specific treatment effect for leadership training (whether or not we adjusted for the year in which the head teacher was appointed to his or her current school).

Table 17. Head teacher effectiveness difference in differences (comparison of means)

Number of criteria fulfilled (out of 7)	(i) Non-ESSPIN	(ii) ESSPIN school		(iii) ESSPIN-trained	
2012	3.5	3		3.4	
2014	3.2	3.2		4.2	
Difference	-0.3	0.2	+	0.8	+

Note. + indicates a significant difference in differences compared to the non-ESSPIN schools ($p < .05$).

⁹ Similar results are obtained when we adjust the intervention categories to take into account the date when the head teacher was appointed to his or her current school.

Table 18. Head teacher effectiveness difference in differences (regression)

Regression on number of criteria fulfilled (out of 7)		Intervention variable					
		School improvement		Training		Training (adjusted for start date)	
Time (CS2 v. CS1)	Coefficient	-0.24		-0.22		-0.10	
	SE	0.15		0.18		0.18	
Intervention	Coefficient	-0.04		0.01		0.01	
	SE	0.11		0.02		0.02	
Treatment	Coefficient	0.55	*	0.02		0.03	
	SE	0.25		0.03		0.03	
	Sample size	1102		1099		972	

Note. * indicates a significant coefficient ($p < .05$).

In summary, head teacher effectiveness appears to be improving over time across the ESSPIN states, and is better in ESSPIN schools than non-ESSPIN ones. There is also evidence that head teacher effectiveness improved faster among head teachers who benefited more from ESSPIN interventions. We are not able conclusively to disentangle whether this effect should be attributed to the specific component of leadership training or to the overall package of ESSPIN interventions aiming to improve school quality – the outcome varies depending on the mode of analysis.

3.3 School development planning

The definition of effective school development planning depends on five criteria (Box 6). Overall, few schools in ESSPIN states reach this standard, but the situation has improved between 2012 and 2014 (Table 19). In particular, more schools were creating SDPs that included three or more activities aiming specifically to strengthen teaching and learning (criterion 3), and the average number of criteria fulfilled, and the percentage of scores reaching the standard, rose significantly. Changes in the other criteria were in the right direction but not statistically significant.

Box 6. Logframe standard for effective school development planning

The school must meet criterion 1 and criterion 2 listed below, and at least two out of three of the remaining criteria, in order to meet the effective school development planning standard

- 1) Written evidence of school self-evaluation process for current school year
- 2) SDP for current school year available
- 3) SDP contains three or more activities which aim to strengthen teaching and learning¹⁰
- 4) Physical evidence of four or more activities stated in SDP having been carried out¹¹
- 5) Cashbook is up-to-date (balanced in the last 60 days)

Table 19. SDP effectiveness in 2012 and 2014

	2012 (CS1)	2014 (CS2)	
(1) Written evidence of school self-evaluation process (%)	20.4	24.3	
(2) SDP available (%)	20.4	26.2	
(3) SDP contains three or more activities to strengthen teaching and learning (%)	9	13.2	+
(4) Evidence that four or more activities stated in SDP carried out (%)	4.7	5.9	
(5) Cashbook up-to-date (%)	13.1	18.3	
Number of SDP criteria fulfilled (/5)	0.6	0.9	+
School meets effective school development planning standard (%)	3.8	7.4	+

Note. + = significant improvement between 2012 and 2014; - = significant worsening between 2012 and 2014 (using a t-test; $p < .05$)

At the time of CS2, ESSPIN schools were also better than non-ESSPIN schools in a number of SDP indicators. This is the case for four of the five individual criteria, the overall number of criteria met, and the proportion of schools meeting the overall standard. Overall, 20% of ESSPIN schools reach the school development planning standard, while only 3% of non-ESSPIN schools do.

¹⁰ Interviewers were given the following instructions to measure the number of activities that aim to strengthen teaching and learning:

Interviewer: 'Strengthening teaching and learning' can include activities such as:

- *promoting pupil and teacher attendance and punctuality;*
- *buying resources for the classroom;*
- *increasing the amount of lesson observations;*
- *minor improvements to the quality of classrooms e.g. furniture, blackboards.*

How many activities in the SDP involve strengthening teaching and learning?

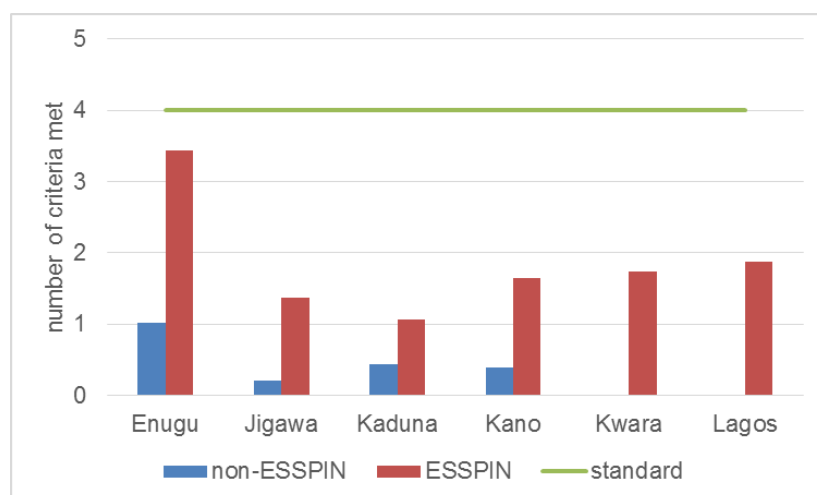
Interviewer: Ask the head teacher to point out the relevant activities in the plan; only count an activity if he/she can explain clearly how it strengthens teaching and learning.

¹¹ Interviewers listed all of the activities in the SDP. For each activity, they then asked whether the activity had been carried out, and if the head teacher said that it had been carried out, they asked to see physical evidence, such as receipts, written records, or objects (e.g. new desks).

Table 20. SDP effectiveness in CS2, by intervention group

	(i) Non-ESSPIN	(ii) ESSPIN	
(1) Written evidence of school self-evaluation process (%)	12.4	56.6	+
(2) SDP available (%)	10.9	67.8	+
(3) SDP contains three or more activities to strengthen teaching and learning (%)	5.2	34.8	+
(4) Evidence that four or more activities stated in SDP carried out (%)	1.4	18	+
(5) Cashbook up-to-date (%)	16.2	24.1	
Number of SDP criteria fulfilled (/5)	0.5	2.0	+
School meets effective school development planning standard (%)	2.8	19.9	+

Note. + indicates a significant difference from the results in non-ESSPIN schools ($p < .05$).

Figure 7. Number of school development planning criteria met, by state and intervention group

Have ESSPIN schools improved faster than non-ESSPIN schools in terms of school development planning? As in the preceding sections, we use two different methods: comparison of means with a categorical intervention variable, and regression analysis with a continuous intervention variable. The former method (Table 21) suggests significantly more positive improvement in ESSPIN than non-ESSPIN schools. In the case of the latter method (Table 22) the treatment effect coefficient is still positive but not significantly so. Overall, this provides suggestive, rather than conclusive, evidence in favour of faster improvement in school development planning in ESSPIN schools than non-ESSPIN schools.

Table 21. SDP effectiveness difference in differences (comparison of means)

	(i) Non-ESSPIN	(ii) ESSPIN	
2012 (CS1)	0.6	1.0	
2014 (CS2)	0.7	2.0	
Difference	0.1	0.9	+

Note: the figures show the number of criteria fulfilled, and so are on a scale from 0 to 5. + indicates a significant difference in differences compared to the non-ESSPIN schools ($p < .05$).

Table 22. SDP effectiveness difference in differences (regression)

		school improvement	
time (CS2 v. CS1)	coefficient	0.20	*
	SE	0.10	
intervention	coefficient	0.51	*
	SE	0.10	
treatment	coefficient	0.21	
	SE	0.16	
	N	1282	

Note: the figures show the number of criteria fulfilled, and so are on a scale from 0 to 5. * indicates a significant coefficient ($p < .05$)

In summary, school development planning appears to be improving in ESSPIN states, and is much better in ESSPIN than in non-ESSPIN schools. There is suggestive evidence that the pace of improvement may be faster in schools that had more ESSPIN intervention between 2012 and 2014, but this depends on the method used.

3.4 School inclusiveness and SBMCs

This section discusses four related sets of indicators relating to inclusiveness and SBMCs: the school's inclusiveness; SBMC functionality; whether the SBMC is inclusive of women; and whether the SBMC is inclusive of children. Further detail on these is provided in the companion Gender and Inclusion Report.

3.4.1 School inclusiveness: meeting the needs of all pupils

The overall standard for school inclusiveness in ESSPIN depends on four criteria (Box 7). The proportion of schools with two or more activities in the SDP that aim to improve access for disadvantaged children (criterion 2) has increased. However the proportion of schools meeting criteria 1 and 4 has decreased significantly. For criterion 1 – the head teacher names more than three actions that he or she has taken to improve pupil attendance – the decline is so large that some degree of measurement error is suspected, although the question format did not change between CS1 and CS2. It is also possible that pupil attendance improved between CS1 and CS2, so that head teachers are now less likely to see a need for further action. Overall, fewer schools met the inclusiveness standard in CS2 than in CS1.

Box 7. Standard for school inclusiveness (meeting needs of all pupils)

The school must meet at least three of the four criteria listed below in order to meet the school inclusiveness standard. The standard is partially met if two criteria are met.

- 1) Head teacher states three or more actions¹² that he/she has taken to improve pupil attendance
- 2) SDP contains two or more activities which aim to improve access
- 3) More than 50% of teachers observed provided evidence of using two or more assessment methods (marked class test, marked pupil workbook, or graded examination paper)
- 4) More than 50% of teachers observed met the spatial inclusion criterion (defined as engaging with at least one pupil from four different areas of the classroom during a lesson) and more than 50% of teachers observed met the gender inclusion criterion. The latter is defined as engaging with boys and girls proportionally to their presence in the classroom within a 10% margin: for example, if the class contains 50% girls then teachers who engage with girls in between 60% and 40% of total engagements will meet the criterion.

Table 23. School inclusiveness in 2012 and 2014

	2012 (CS1)	2014 (CS2)	
(1) Three or more actions on attendance (%)	57.9	39.1	-
(2) Two or more activities in SDP on access for disadvantaged children (%)	5.4	11.9	+
(3) >50% of teachers use two or more assessment methods (%)	70.7	62.3	
(4) >50% of teachers spatially inclusive and >50% are gender inclusive (%)	33.4	23.4	-
Number of inclusiveness criteria fulfilled (/4)	1.7	1.4	-
Inclusiveness score	72.2	63.7	-
School partially met inclusiveness standard (2–4 criteria out of 4)	60.4	46.5	-
School fully met inclusiveness standard (3–4 criteria out of 4)	18.8	12.7	

Note. The inclusiveness score is a total ranging from 0 to 100 and is calculated as follows: $20(\frac{s_1}{7} + \min(1, \frac{s_2}{5})) + \frac{s_3}{3} + \frac{s_4}{6} + s_5$, where s_1 is the number of actions to improve attendance; s_2 is the number of activities in the SDP to improve access for disadvantaged children; s_3 is the average number of assessment methods used by sampled teachers; s_4 is the average number of classroom zones participating in the lesson during lesson observations, and s_5 is the gender equity score (see below). + = significant improvement between 2012 and 2014; - = significant worsening between 2012 and 2014 (using a t-test; $p < .05$)

Focusing on CS2 schools, schools in which we would expect an improvement due to ESSPIN intervention are significantly more inclusive than non-ESSPIN schools, in terms of activities to improve access for disadvantaged children (criterion 2), use of different assessment methods (criterion 3), and in terms of the overall inclusiveness score and the proportion of schools meeting standards. Overall 25% of ESSPIN schools and only 8% of non-ESSPIN schools fully met the inclusiveness standard. Nearly two-thirds of ESSPIN schools partially met the standard (meeting two of the four criteria), but only two-fifths of non-ESSPIN schools did so.

Teachers in ESSPIN schools used more assessment methods, involved children from more parts of the classroom in lessons, and came closer to ensuring equitable participation of boys and girls in lessons. Very few non-ESSPIN schools had an SPD that contained an activity that aims to improve

¹² This was incorrectly stated as *more than three* actions in the CS1 report.

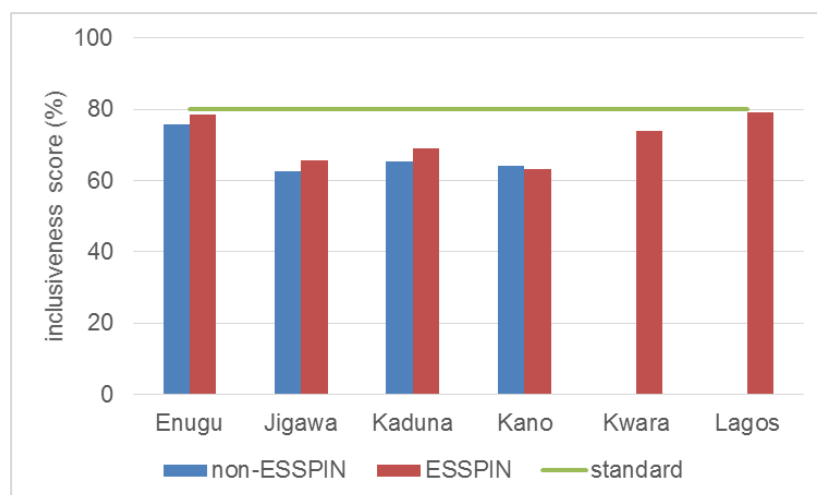
access for disadvantaged children. Over 50% of ESSPIN schools did, and over 30% had two activities or more.

Table 24. School inclusiveness in CS2, by intervention group

	(i) Non-ESSPIN	(ii) ESSPIN	
<i>Inclusiveness criteria</i>			
(1) Three or more actions to improve attendance	39.4	38.3	
(2) Two or more activities in SDP to improve access for disadvantaged children	4.4	32.1	+
(3) >50% of teachers use two or more assessment methods	55.0	81.4	+
(4) >50% of teachers spatially inclusive and >50% are gender inclusive	20.6	30.4	
<i>Overall inclusiveness standard</i>			
Number of inclusiveness criteria fulfilled (/4)	1.2	1.8	+
Inclusiveness score	61.3	69.7	+
School partially met inclusiveness standard (2–4 criteria out of 4)	40.1	63.6	+
School fully met inclusiveness standard (3–4 criteria out of 4)	8.0	25.1	+
<i>Detailed</i>			
Number of actions to improve attendance	2.3	2.3	
Number of activities on access for disadvantaged children	0.1	1.1	+
Average number of assessment methods used	1.1	1.9	+
Average number of zones participating in lessons	3.5	3.9	+
Average gender equity score (0=completely unequal, 100=perfectly equal)	80.6	86.7	+

+ indicates a significant positive difference between non-ESSPIN and ESSPIN schools ($p < .05$). The gender equity score for a teacher is $100 - 100 \times \text{abs}(\frac{g}{g+b} - \frac{G}{G+B})$ where g is the number of girls who participate, b is the number of boys who participate, G is the number of girls present in the class, and B is the number of boys present in the class. It is expressed as a percentage score. For a lesson where the proportion of girls and boys participating is exactly equal to the proportion of girls and boys sitting in the lesson, the gender equity score will be 100; for a lesson where no boys participate or no girls participate, the score will be zero.

Figure 8. Inclusiveness score, by state and intervention group



Note. An inclusiveness score of around 80% or more is needed to meet the overall standard.

Examining the difference in differences between ESSPIN and non-ESSPIN schools over CS1 and CS2, both comparison of means (Table 25) and regression analysis (Table 26) confirm that the worsening over time in the total inclusion school was reduced in the ESSPIN schools, an effect which is statistically significant.

Table 25. School inclusiveness difference in differences (comparison of means)

Inclusiveness score	(i) No expected improvement	(ii) Expected improvement	
2012 (CS1)	71.4	74.7	
2014 (CS2)	62.3	71.3	
Difference	-9.4	-3.4	+

Note. + indicates a significant positive difference in difference compared to the non-ESSPIN schools ($p < .05$).

Table 26. School inclusiveness difference in differences (regression)

Regression on inclusiveness score			
Time (CS2 v. CS1)	Coefficient	-9.29	*
	SE	1.31	
Intervention	Coefficient	2.20	*
	SE	0.95	
Treatment	Coefficient	3.44	*
	SE	1.03	
	N	1253	

Note. * indicates a significant coefficient ($p < .05$).

3.4.2 How well do SBMCs function?

There are nine criteria in the standard for SBMC functionality (Box 8). In general, SBMC functionality appears to have improved between 2012 and 2014, although the proportion of schools meeting the criteria remains low. The average school met 2.3 of the nine criteria in 2012, but in 2014 met three of the criteria – a statistically significant improvement. Most of the criteria for SBMC functionality rely on the ability to provide written or photographic evidence, or at least oral recollection of a specific event. Consequently, the criteria may reflect the quality of record keeping of the SBMC, more than the particular aspects of functionality they aim to measure. Examining individual criteria, SBMCs have particularly improved in terms of awareness-raising activities, networking, women's committees, and contribution of resources for the school. There was a decline in the proportion of SBMCs with written evidence of the chairperson having visited the school.

Two additional inclusiveness-related criteria not included in the CS1 report are also examined in this section: whether the SBMC did anything to support commonly excluded groups, and whether it raised issues of children's exclusion from school with the community, LGEA or state government. Both of these types of SBMC action on inclusion took place more in 2013/14 than in 2011/12.

Box 8. Logframe standard for SBMC functionality

The school must meet at least five of the nine criteria listed below in order to meet the SBMC functionality standard for the school year¹³:

- 1) Two or more SBMC meetings have taken place since the start of the school year (written evidence)
- 2) SBMC conducted awareness-raising activities (written or oral evidence)
- 3) SBMC took steps to address exclusion (written or oral evidence)
- 4) SBMC networked with CBOs, traditional or religious institutions, or other SBMCs (written or physical evidence)
- 5) SBMC interacted with local government education authorities on education service delivery issues (written or physical evidence)
- 6) SBMC women's committee exists (written or physical evidence)
- 7) SBMC children's committee exists (written or physical evidence)
- 8) SBMC contributed resources for the school (written or physical evidence)
- 9) SBMC chair visited the school at least three times since the start of the school year (written evidence)

Table 27. SBMC functionality in 2012 and 2014

	2012 (CS1)	2014 (CS2)	
(1) Two or more meetings this school year	28.7	27.1	
(2) Conducted awareness-raising	35.3	47.5	+
(3) Addressed exclusion	26.7	40.1	+
(4) Networked with CBOs/institutions/other SBMCs	15	55.6	+
(5) Interacted with LGEA	19.7	21.1	
(6) Has a women's committee	13.1	26.6	+
(7) Has a children's committee	19	21	
(8) Contributed resources for school	39	54.5	+
(9) Chair visited school three or more times	25.2	14.8	-
Standard G: functioning SBMC	21.7	30.9	+
Number of SBMC functionality criteria met (out of nine)	2.3	3.3	+
<i>Additional criteria</i>			
Action for commonly excluded groups	13.9	23.8	+
Raised issue of children's exclusion	4.8	19.3	+

Note. + = significant improvement between 2012 and 2014; - = significant worsening between 2012 and 2014 (using a t-test; $p < .05$)

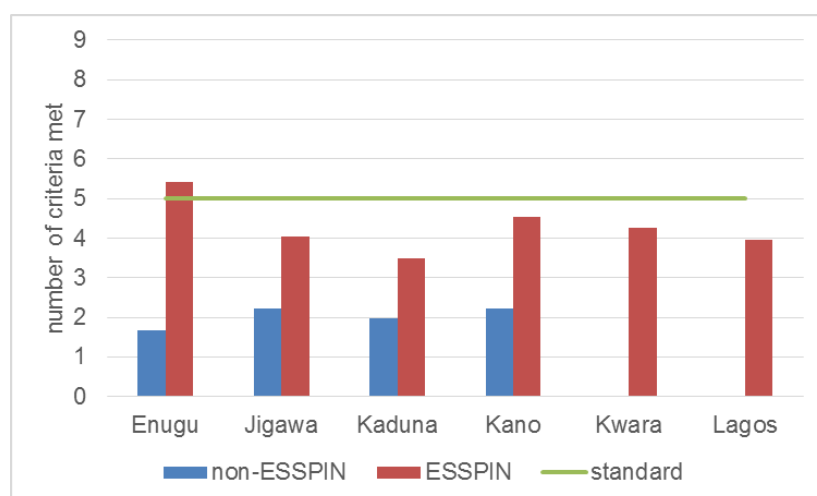
Examining differences between ESSPIN and non-ESSPIN schools in 2014, there were significant differences in the expected direction across all nine criteria. Overall, only 13% of non-ESSPIN schools, but 62% of ESSPIN schools, met the standard for a functional SBMC.

¹³ A slightly different standard with 10 criteria was used in CS1. The new standard with nine criteria was applied to both the CS1 and CS2 data.

Table 28. SBMC functionality in CS2, by intervention group

	(i) Non-ESSPIN	(ii) ESSPIN	
(1) Two or more meetings this school year	13.7	62.6	+
(2) Conducted awareness-raising	38.6	71.1	+
(3) Addressed exclusion	32.7	59.8	+
(4) Networked	48	75.8	+
(5) Interacted with LGEA	16.1	34	+
(6) Has a women's committee	14.9	57.7	+
(7) Has a children's committee	9.5	51.7	+
(8) Contributed resources for school	47.5	73.1	+
(9) Chair visited school three or more times	5.7	39	+
Standard G: functioning SBMC	17.3	67	+
Number of SBMC functionality criteria met	2.5	5.3	+
<i>Additional criteria</i>			
Action for commonly excluded groups	22.4	27.0	
Raised issue of children's exclusion	15.0	29.8	+

Note. + indicates a significant positive difference from the results in non-ESSPIN schools ($p < .05$)

Figure 9. Number of SBMC functionality criteria met, by state and intervention group

Comparing the change in the means over time and across intervention groups, SBMCs in ESSPIN schools appear to have improved more than those in non-ESSPIN schools when we compare means with a categorical measure of intervention (non-ESSPIN vs. ESSPIN) (Table 29) but the effect is not significant when we use regression analysis and a continuous measure of intervention (less ESSPIN vs. more ESSPIN) (Table 30). This result requires further investigation and may suggest that the improvement in SBMCs does not increase in a simple linear way with additional output 3 intervention. The Gender and Inclusion Report will examine this issue in more detail and with respect to output 4 interventions, which may have had more effect on SBMC functionality than the output 3 interventions.

Table 29. Difference in differences of SBMC functionality (comparison of means)

Number of SBMC functionality criteria met	(i) Non-ESSPIN	(ii) ESSPIN	
2012 (CS1)	2.1	2.9	
2014 (CS2)	3.0	4.9	
Difference	0.8	2.0*	+

Note. + indicates a significant positive difference in difference compared to the non-ESSPIN schools ($p < .05$).

Table 30. Difference in differences of SBMC functionality (regression)

Regression on number of SBMC functionality criteria met			
Time (CS2 v. CS1)	Coefficient	0.90	*
	SE	0.25	
Intervention	Coefficient	0.74	
	SE	0.45	
Treatment	Coefficient	0.49	
	SE	0.64	
	Observations	1264	

Note. * indicates a significant coefficient ($p < .05$).

3.4.3 How inclusive are SBMCs of women?

As in the report on CS1, we also examine the extent to which SBMCs are inclusive of women's and children's concerns. The standard on SBMC women's inclusiveness has four criteria (Box 9). There was an improvement between CS1 and CS2 in the proportion of schools where a female member raised an issue during an SBMC meeting, and in the proportion of SBMCs with a women's committee that had met recently (Table 31). Overall, however, the improvement in women's inclusiveness was not large enough to be statistically significant.

Box 9. Logframe standard for SBMCs' inclusiveness of women

The school must meet at least three of the four criteria listed below in order to meet the SBMC women's inclusion standard for the last school year:

- 1) At least one woman attended two or more SBMC meetings (written evidence)
- 2) A female member of SBMC raised at least one issue at SBMC meetings (written evidence or oral evidence from female member of SBMC)
- 3) At least one issue raised by a female member at an SBMC meeting led to action (written, physical or oral evidence from female member of SBMC)
- 4) At least one SBMC women's committee meeting took place¹⁴

¹⁴ This criterion has been slightly altered since CS1 – CS1 had also required the women's committee to have a female leader.

Table 31. SBMCs' women's inclusion in 2012 and 2014

	2012 (CS1)	2014 (CS2)	
(1) At least one woman attended two or more meetings (%)	19.5	17.9	
(2) Female member raised an issue (%)	26.8	31.5	
(3) Issue raised by female member led to action (%)	28.3	14.8	-
(4) Women's committee met (%)	7.8	26.6	+
Number of criteria met	0.6	0.9	+
Meets standard (3/4 criteria)	15.4	15.8	

Note. + = significant improvement between 2012 and 2014; - = significant worsening between 2012 and 2014 (using a t-test; $p < .05$)

There were large differences in women's inclusion between ESSPIN and non-ESSPIN schools, which are significant and in the expected direction across the four criteria (Table 32). Nearly half of ESSPIN schools, but very few non-ESSPIN schools, met the standard for women's inclusiveness. Difference in differences analysis suggests that ESSPIN schools are also improving faster than non-ESSPIN schools in terms of the extent to which SBMCs are inclusive of women (Table 33 and Table 34).

Table 32. SBMC women's inclusion in 2014, by intervention group

	(i) Non-ESSPIN	(ii) ESSPIN	
(1) At least one woman attended two or more meetings (%)	4.9	49.5	+
(2) A female member raised an issue (%)	16.8	66.6	+
(3) Issue raised by female member led to action (%)	4.2	39.7	+
(4) Women's committee met (%)	13.3	58.3	+
Number of criteria met	0.3	2.1	+
Meets standard (3/4 criteria)	2.4	48.3	+

Note. + indicates a significant positive difference from the results in non-ESSPIN schools ($p < .05$)

Table 33. Difference in differences of SBMC women's inclusion (comparison of means)

Number of women's inclusion criteria met (out of 4)	(i) Non-ESSPIN	(ii) ESSPIN	
2012 (CS1)	0.6	0.8	
2014 (CS2)	0.7	1.8	
Difference	0.1	1.0	+

Note. + indicates a significant positive difference in differences compared to the non-ESSPIN schools ($p < .05$).

Table 34. Difference in differences of SBMC women's inclusion (regression)

Regression on number of women's inclusion criteria met			
Time (CS2 v. CS1)	Coefficient	0.15	
	SE	0.11	
Intervention	Coefficient	0.18	
	SE	0.11	
Treatment	Coefficient	0.48	*
	SE	0.10	
	Observations	1208	

Note. * indicates a positive coefficient ($p < .05$)

3.4.4 How inclusive are SBMCs of children?

There are four criteria within the standard on SBMC inclusiveness of children. There was a significant increase between CS1 and CS2 in the proportion of SBMCs where children raised an issue, and a small but significant increase in the overall number of criteria met, but not in the proportion of children meeting the standard, which remained low, at 6%.

Box 10. Logframe standard for SBMC inclusiveness of children

The school must meet at least three of the four criteria listed below in order to meet the SBMC children inclusiveness standard for the last school year:

- 1) At least one child attended two or more SBMC meetings (written evidence)
- 2) A child member of SBMC raised at least one issue at SBMC meetings (written evidence or oral evidence from child member of SBMC)
- 3) At least one issue raised by a child member at an SBMC meeting led to action (written, physical or oral evidence from child member of SBMC)
- 4) At least one SBMC children's committee meeting took place and committee has a trained facilitator¹⁵

Table 35. SBMC inclusion of children in 2012 and 2014

	2012 (CS1)	2014 (CS2)	
(1) A child attended two or more meetings (%)	11.8	8.8	
(2) A child raised an issue (%)	13.6	20.6	
(3) Issue raised by child led to action (%)	11.5	7.3	
(4) Children's committee met and it has a trained facilitator (%)	2.4	14.3	+
Number of criteria met	0.3	0.5	+
Meets standard (3/4 criteria) (%)	5.7	6.2	

Note. The large increase in criterion 4 may be due to a relaxation of the evidence requirement (see footnote 15). + = significant improvement between 2012 and 2014; - = significant worsening between 2012 and 2014 (using a t-test; $p < .05$).

As with women's inclusiveness, there are large positive differences between ESSPIN and non-ESSPIN schools (Table 36). Overall 18% of ESSPIN schools and fewer than 2% of non-ESSPIN schools met the standard for SBMC children's inclusiveness. Moreover, regardless of the method used, ESSPIN schools improved faster in terms of SBMC children's inclusiveness between 2012 and 2014 (Table 37 and Table 38).

¹⁵ In CS1 this criterion required written evidence in the form of minutes of at least one children's committee meeting held in the past school year. This requirement was dropped for CS2 as it was considered unlikely that children's committees would keep good minutes, and that failure to keep minutes does not mean the committee is not functioning.

Table 36. SBMC children's inclusion in 2014, between intervention groups

	(i) non-ESSPIN	(ii) ESSPIN	
(1) A child attended two or more meetings (%)	2.3	25.2	+
(2) A child raised an issue (%)	11.0	44.5	+
(3) Issue raised by child led to action (%)	2.7	19.1	+
(4) Children's committee met and it has a trained facilitator (%)	7.0	32.6	+
Number of criteria met	0.2	1.2	+
Meets standard (3/4 criteria) (%)	1.6	17.7	+

Note. + indicates a significant positive difference from the results in non-ESSPIN schools ($p < .05$)

Table 37. Difference in differences in SBMC children's inclusion (comparison of means)

Number of criteria fulfilled (out of four)	(i) Non-ESSPIN	(ii) ESSPIN	
2012 (CS1)	0.3	0.4	
2014 (CS2)	0.4	1.1	
Difference	0.1	0.7	+

Note. + indicates a significant positive difference in difference compared to the non-ESSPIN schools ($p < .05$).

Table 38. Difference in differences in SBMC children's inclusion (regression)

Regression on number of children's inclusion criteria met			
Time (CS2 v. CS1)	Coefficient	0.14	
	SE	0.08	
Intervention	Coefficient	0.11	
	SE	0.07	
Treatment	Coefficient	0.33	*
	SE	0.08	
	Observations	1264	

Note. * indicates a significant coefficient ($p < .05$).

Overall, schools in ESSPIN states appear to be getting worse in terms of our overall inclusiveness indicators, yet getting better in terms of several other inclusion-related indicators, including SBMC functionality, SBMC action on exclusion, and some criteria relating to inclusiveness of SBMCs in respect of women and children. Focusing on the CS2 data, ESSPIN schools are doing better than non-ESSPIN schools on a wide range of indicators. We find evidence of ESSPIN schools improving faster (or at least, worsening more slowly) for overall school inclusiveness, SBMC functionality, and SBMC inclusiveness of women and children. The forthcoming report on Gender and Inclusion will explore these results in more detail.

3.5 School quality

Overall school quality is measured as a combination of the standards on teacher competence, head teacher effectiveness, school development planning, and SBMC functionality. A quality school is defined as one that meets the teacher competence standard and at least two of the other standards (Box 11). Comparison of school quality between CS1 and CS2 suggests that there has been a large increase in the proportion of schools that meet the overall school quality standard, from 4% to 10% (Table 39). We also use a 'quality score' indicator, which is an average of the continuous indicators developed in the previous sections for teacher competence, head teacher

effectiveness, school development planning, and SBMC functionality. Using this variable there is also a significant positive change in the average score between CS1 and CS2.

Box 11. Logframe standard for school quality

The school must meet at least three of the four output standards listed below in order to meet the school quality outcome standard, with teacher competence having to be one of those three.

- 1) Teacher competence standard (more than half the teachers sampled in each school must be competent)
- 2) Head teacher effectiveness standard
- 3) School development planning effectiveness standard
- 4) SBMC functionality standard

The version of this standard used in CS1 did not rely on teacher content knowledge tests. For CS2, we introduce a second, more strict version of the standard, in which teachers must get above 50% in literacy and numeracy tests to be classed as competent (see section 3.1 and Box 3 above).

Table 39. School quality in 2012 and 2014

	2012 (CS1)	2014 (CS2)	
Meets three or four standards (old version)	3.9	9.8	+
Quality score (old version)	38.3	42.5	+

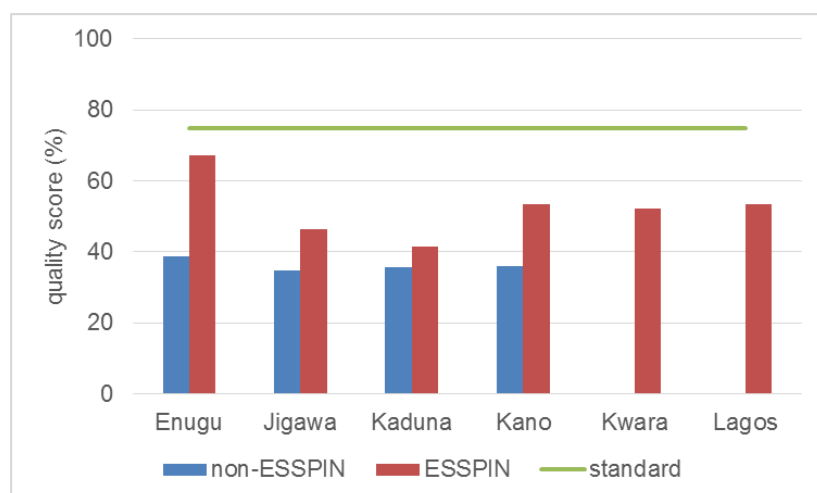
Note. + = significant improvement between 2012 and 2014; - = significant worsening between 2012 and 2014 (using a t-test; $p < .05$).

Within CS2, the proportion of schools meeting the overall school quality is dramatically higher among schools where an improvement would be expected due to ESSPIN interventions, compared to those where no improvement would be expected (Table 40). Only around 1% of non-ESSPIN schools meet the stricter quality standard brought in for CS2, compared to 16% of ESSPIN schools. Using our continuous indicator of school quality, scores are also much higher in ESSPIN schools than in non-ESSPIN schools.

Table 40. School quality in CS2, by intervention group

	(i) Non-ESSPIN	(ii) ESSPIN	
Meets three or four standards (old version)	3.2	27.2	+
Meets three or four standards (new version)	1.5	16.7	+
Quality score (old version)	36.5	56.2	+
Quality score (new version)	34.8	55.3	+

Note: The new (CS2) version of the quality score and school quality standard reflect the strict version of the teacher competence standard, where teachers are required to pass literacy and numeracy tests, as well as fulfilling other criteria. + indicates a significant positive difference between ESSPIN and non-ESSPIN schools ($p < .05$).

Figure 10. School quality score, by state and intervention group

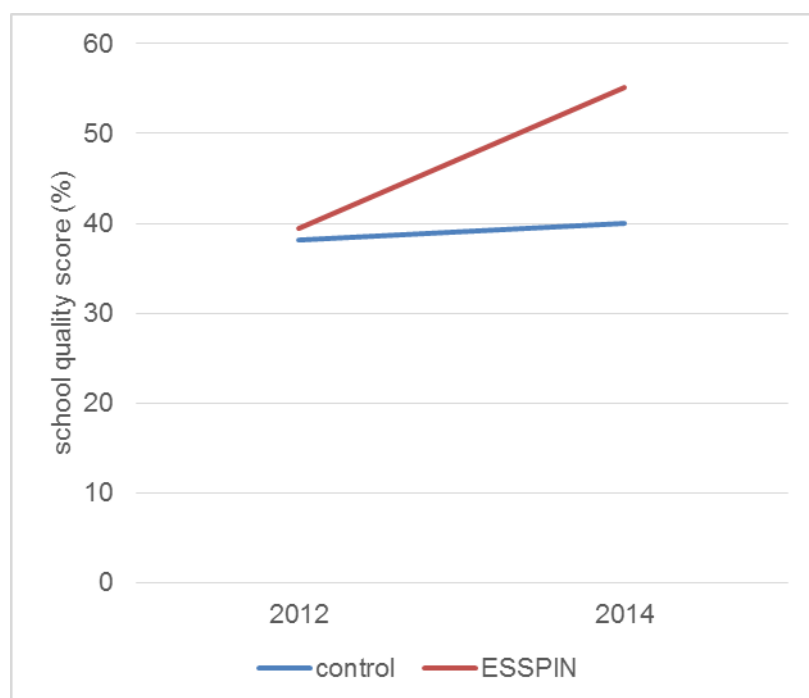
Note. A school quality score of around 75% is needed to meet the CS2 standard on school quality.

The two methods of comparing difference in differences both suggest that the pace of overall quality improvement has been faster in schools that received more ESSPIN intervention between 2012 and 2014. The treatment effect is large and significantly positive using both methods of analysis (Table 41 and Table 42; Figure 11). Control schools have had close to no change in school quality, while ESSPIN schools have improved dramatically, during 2012–2014.

Table 41. School quality difference in differences (comparison of means)

Quality score (old version)	(i) Non-ESSPIN	(ii) ESSPIN	
2012 (CS1)	38.1	39.4	
2014 (CS2)	40.0	55.2	
Difference	1.9	15.8	+

Note. + indicates a significant positive difference in difference compared to the non-ESSPIN schools ($p < .05$).

Figure 11. School quality in 2012 and 2014, in control and ESSPIN schools

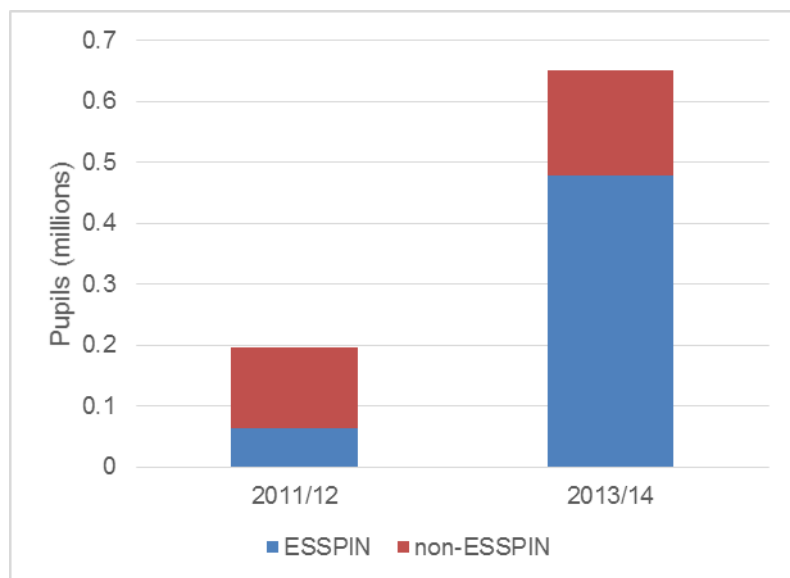
Source: based on means in Table 41.

Table 42. School quality difference in differences (regression with continuous intervention variable)

Regression on quality score (old version)			
Time (CS2 v. CS1)	Coefficient	2.82	
	SE	1.48	
Intervention	Coefficient	3.5	*
	SE	1.28	
Treatment	Coefficient	5.64	*
	SE	2.39	
	Observations	1027	

Note. * indicates a significant coefficient.

Using the CS1 and CS2 estimates of numbers of children in schools passing the quality threshold, combined with census data on enrolments, we can estimate that the number of children in good quality schools in the six states rose from under 200,000 in 2011/12 to over 650,000 in 2013/14 (Figure 12). Of this increase of 450,000 in the number of children attending good quality schools, 90% were in ESSPIN schools.

Figure 12. Number of children in a good quality school

Source: estimates based on annual school census, ESSPIN intervention database, and survey results.

3.6 Pupil learning achievement in English literacy and numeracy

3.6.1 Main analysis

Test results for the ESSPIN states as a whole appear to have declined significantly between CS1 and CS2 (Table 43), except in grade 2 literacy, which has not changed significantly. But children in schools that we expected to have a higher quality due to ESSPIN intervention did indeed have higher test scores than those in control schools (Table 44). The pattern has been roughly the same for both boys and girls. Within states, the pattern is more complicated (Figure 13). Overall, pupils who have been more exposed to schools with ESSPIN intervention are achieving higher test results, but the pattern is not uniform, and the differences between intervention groups within states are small compared to the differences between states. The proportion of pupils achieving ESSPIN logframe indicators (see Annex A) in each test remained very low: around 10% for grade 2 numeracy, and under 5% for each of the other tests. Average test scores remain below 50% even in ESSPIN schools.

How do we explain this pattern of results? It should be noted first that the differences between ESSPIN and non-ESSPIN schools in Table 44 partly reflect differences between, on the one hand, Lagos and Kwara, where ESSPIN was rolled out to all schools by 2012/13, and, on the other, the other states. The more detailed analysis below reveals that there are still significant differences when we control for state (this is shown in Figure 13 and examined statistically in Chapter 4 below), but they are much smaller in magnitude. Secondly, there may be general changes in education in the states that are causing declining test results. In particular, enrolment increases may be placing an increasing strain on school resources (see section 1.1 above), and the profile of children who enter school may also be changing, so that, for example, children from poorer or less educated family backgrounds may make up a larger proportion of enrolment. Thirdly, the scale of ESSPIN, except in Kwara and Lagos, was still (by 2012/13) too small relative to the number of schools in each state to have outweighed any general trend of declining learning outcomes (see section 1.2 above). Finally, we cannot rule out measurement error. Although great efforts were

made to keep CS1 and CS2 test instruments comparable, the process of instrument revision and differences in fieldwork monitoring procedures could have influenced the results¹⁶.

Table 43. Test scores and proportion of children reaching logframe indicator in 2012 and 2014

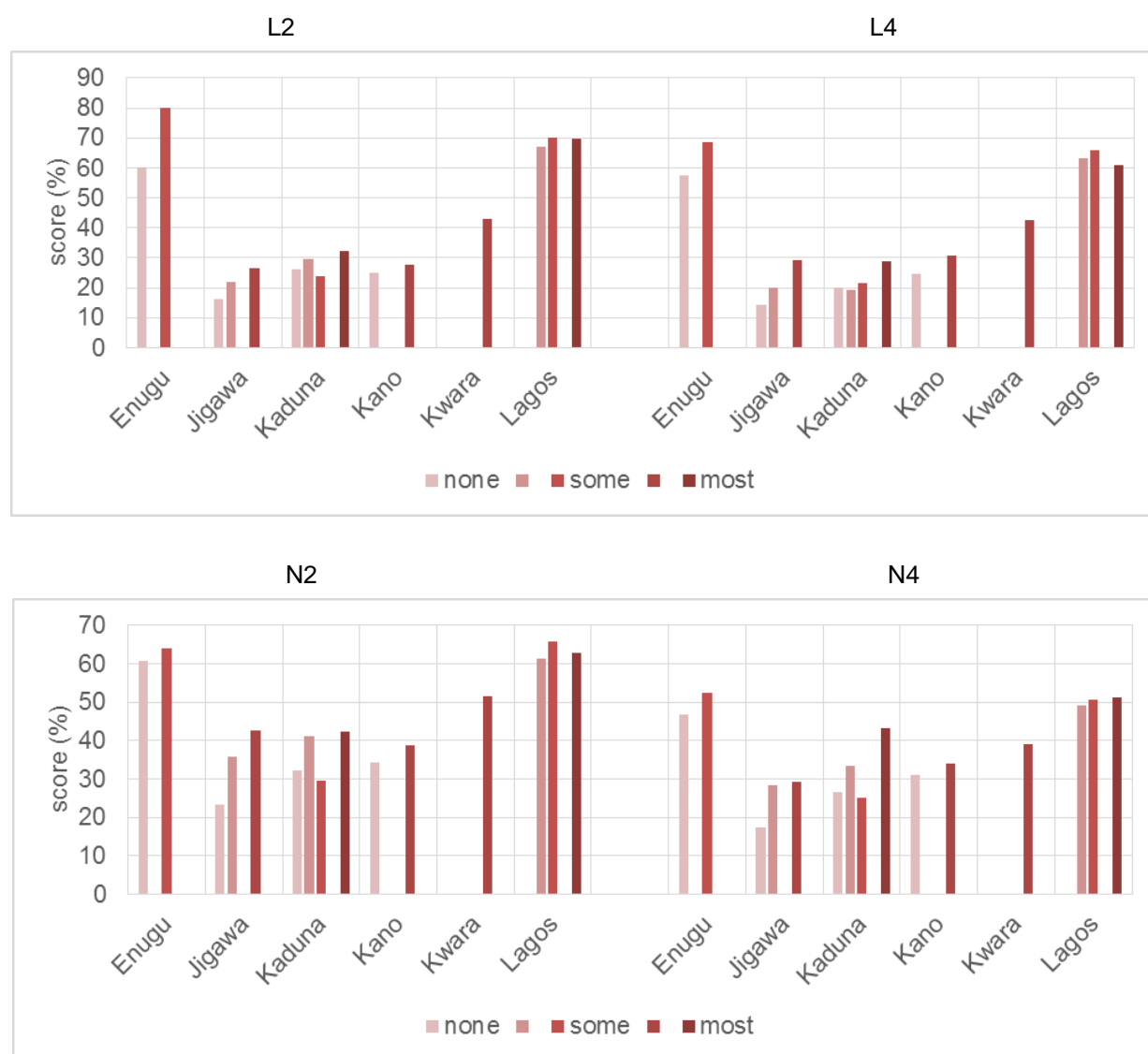
		boys			girls			total		
		2012	2014		2012	2014		2012	2014	
Test score (%)	L2	29.2	29.9		31.3	30.3		30.3	30.1	
	L4	35.7	28.9		33.6	30.2		33.7	29.5	-
	N2	48.3	38.2		47.3	37.3		48	37.8	-
	N4	35.3	32.5		36.8	32.5		36.1	32.5	-
Logframe indicator (%)	L2	1.5	2.4	+	7.3	2.7		4.2	2.5	
	L4	2.9	1.6		2.3	1.7		2.7	1.6	
	N2	11.9	5.2		12.1	8.2	-	12.1	6.5	
	N4	3.7	1.4		8.3	2.7		6	2	-

Note. L2 = grade 2 literacy; L4 = grade 4 literacy; N2 = grade 2 numeracy; N4 = grade 4 numeracy. + = significant improvement between 2012 and 2014; - = significant worsening between 2012 and 2014 (using a t-test; $p < .05$).

Table 44. Test scores and proportion of children reaching logframe indicator in 2014, by intervention group

		boys			girls			total		
		non-ESSPIN	ESSPIN		non-ESSPIN	ESSPIN		non-ESSPIN	ESSPIN	
Test score (%)	L2	27.4	38.2	+	28.4	39.4	+	26.1	40.8	+
	L4	29.5	39.0	+	28.8	40.9	+	24.7	40.1	+
	N2	40.3	49.5	+	40.2	48.6	+	34.3	47.1	+
	N4	32.0	38.7	+	32.8	39.7	+	29.6	38.7	+
Logframe indicator (%)	L2	1.8	2.8		4.6	5.0		1.9	4.4	+
	L4	2.0	2.6		1.5	3.4	+	0.9	3.3	+
	N2	7.0	11.3	+	9.3	12.7		5.2	10.2	+
	N4	2.1	3.3		5.4	4.4		1.2	3.8	+

¹⁶ Questions were identical, but there were some changes in administration. In particular, cascading training of data collectors was used in CS1, i.e. a small group of officers were trained centrally, who then trained their colleagues in each state. In CS2, a single training event was used to ensure greater consistency in administration of the test, and data collection was also monitored more closely. These changes were necessary to ensure the data quality in CS2, but may have led to data collectors administering tests more strictly in CS2 than in CS1. See Annex B.

Figure 13. Test scores by state and ESSPIN intervention

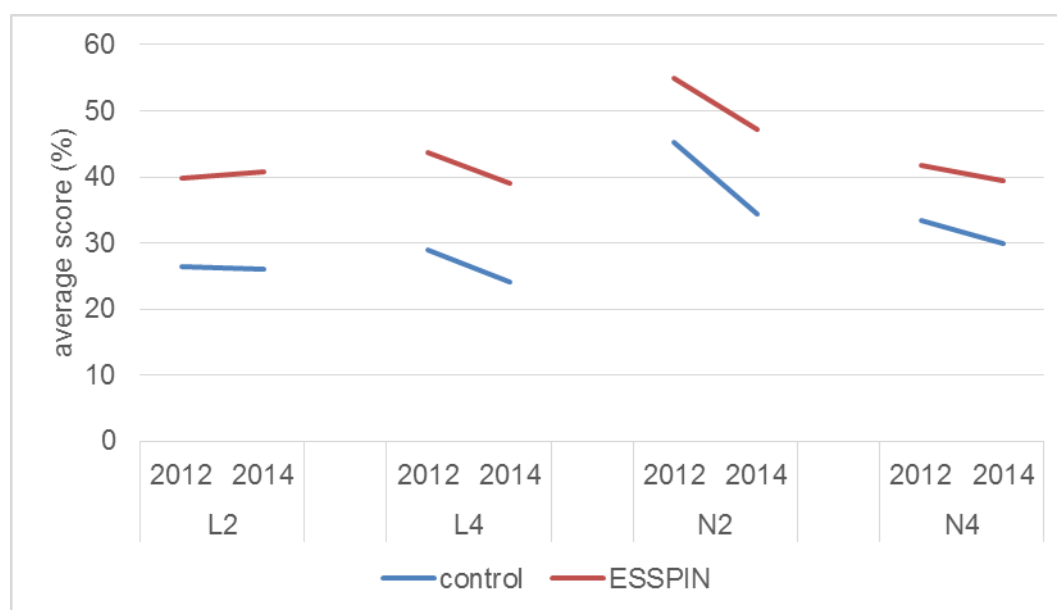
Note. Darker colours represent greater exposure to ESSPIN intervention.

We do, however, find evidence supportive of an ESSPIN impact when we look at the difference in differences. Using two different methods – comparison of means with a categorical intervention variable (Table 45) or regression with a continuous intervention variable (Table 46) – results deteriorated less in ESSPIN schools than in non-ESSPIN schools, suggesting an impact of ESSPIN in helping to offset a state-wide worsening trend. However, the difference is only statistically significant using the regression method and for literacy results. In chapter 4 we extend this analysis to test whether this finding is robust to controlling for state and school characteristics.

Table 45. Pupil test score difference in differences (comparison of means)

		(i) Non-ESSPIN	(ii) ESSPIN	
L2	2012 (CS1)	26.4	39.8	
	2014 (CS2)	26.1	40.8	
	Difference	-0.3	1.0	
L4	2012 (CS1)	29.0	43.7	
	2014 (CS2)	24.7	40.1	
	Difference	-4.3	-3.7	
N2	2012 (CS1)	45.3	54.9	
	2014 (CS2)	34.3	47.1	
	Difference	-11.0	-7.7	
N4	2012 (CS1)	33.4	41.8	
	2014 (CS2)	29.6	38.7	
	Difference	-3.7	-3.1	

Underlying the difference in difference results is a worsening in average test scores in L4, N2, and N4 (Figure 14). In each case the worsening appears to have been slightly less severe for ESSPIN schools than for control schools. For L2, pupils' test scores in control schools have worsened very slightly, while ESSPIN school test scores have improved very slightly, by less than one percentage point in each case.

Figure 14. Pupil test scores in ESSPIN and control schools, in 2012 and 2014

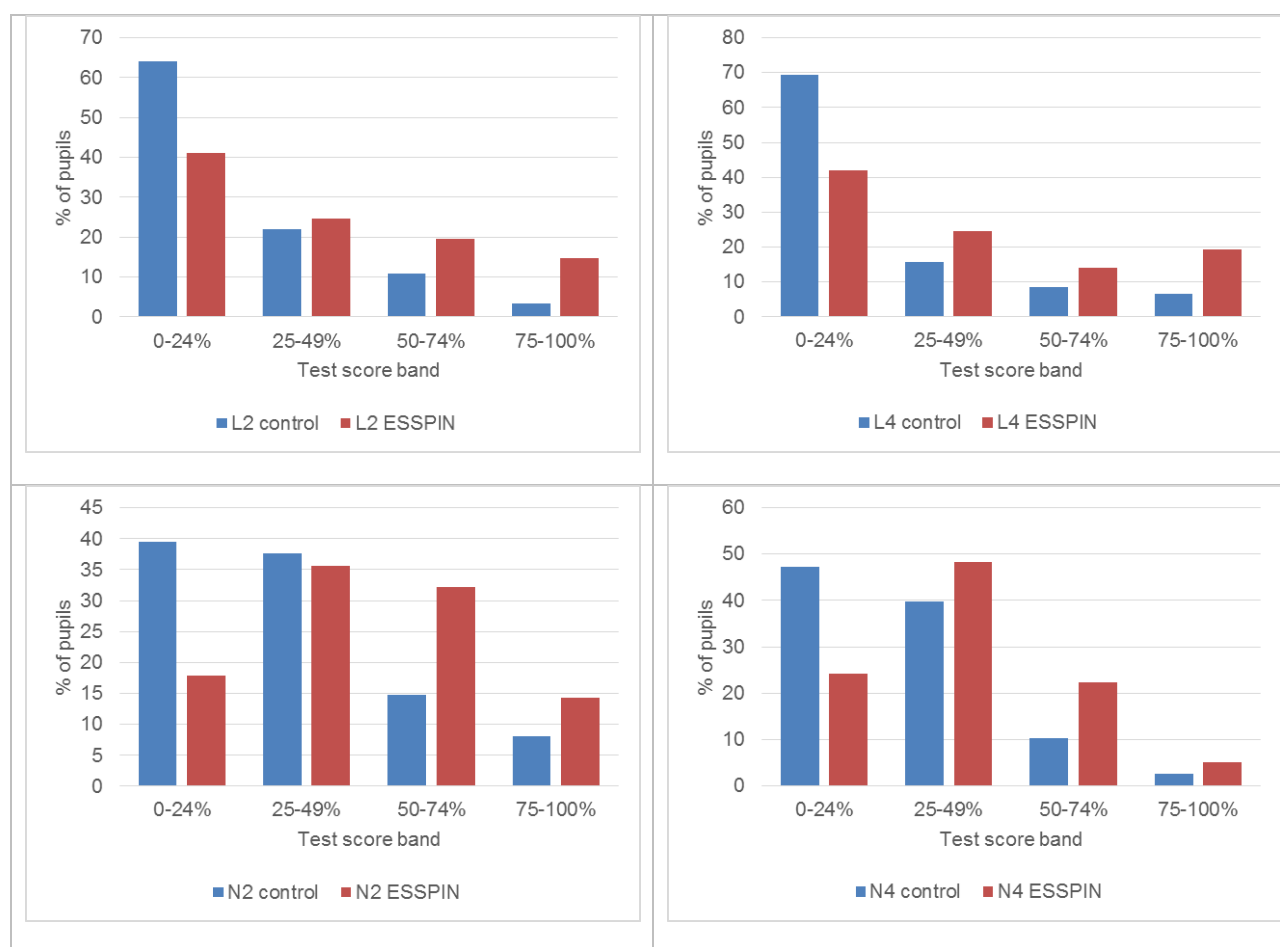
Note. Based on averages in Table 45.

Table 46. Pupil test score difference in differences (regression)

		L2		L4		N2		N4	
Time (CS2 v. CS1)	Coefficient	0.78		-3.92		-8.6	*	-5.4	*
	SE	2.14		2.04		2.23		2.26	
Intervention	Coefficient	5.45	*	2.55	*	4.25	*	1.73	*
	SE	2.07		0.82		1.71		0.63	
Treatment	Coefficient	2.86	*	1.65	*	2.11		0.75	
	SE	1.12		0.65		1.1		0.64	
	Observations	1297		1293		1297		1292	

3.6.2 Distribution of test scores and sub-scale scores

Looking at the distribution of test scores (Figure 15) reinforces the findings summarised above regarding differences between ESSPIN and non-ESSPIN schools. Over 60% of children in non-ESSPIN schools continue to score under 25% in English literacy tests. In ESSPIN schools there are still more children scoring in this band than in any of the other bands (25%–49%, 50%–74%, or 75%–100%), but the proportion is much lower than that in non-ESSPIN schools. For numeracy, most children in ESSPIN schools are in the middle bands, while most children in non-ESSPIN schools remain at the lower end of the scale.

Figure 15. Distribution of test scores by ESSPIN intervention in CS2

Comparisons of sub-scales – groups of items within the tests that correspond to a particular learning concept or level – appear to have worsened in most cases between CS1 and CS2 (Table 47). Number concepts – with questions such as counting to 10 or from 100 – are among the areas of strongest performance, although even here the average mark is under 50% in CS2. Grade 2 children achieved scores under 25% for addition and subtraction, and very few grade 2 children could master writing tasks. In grade 4, children scored around 50% in items aimed at grade 1 or 2, and under 20% in items aimed at grade 4. The fairly consistent decline across different sub-scales suggests that measurement error may not be the main factor in the overall reduction in test scores between CS1 and CS2. Measurement error resulting from minor changes in test administration would tend to affect relatively few test items in isolation, and so is difficult to square with such a consistent pattern of reductions.

Table 47. Detailed scores in test sub-scales, 2012 vs. 2014 (%)

Test	Sub-scale	2012 (CS1)	2014 (CS2)	
L2	Writing	19.8	13.1	-
	Reading	26.9	27.2	
	Score in grade 1 items	36	36.7	
	Score in grade 2 items	25.1	25.1	
N2	Number concepts	60.3	47.3	-
	Addition and subtraction	37.9	22.2	-
	Score in grade 1 items	59.1	51.4	-
	Score in grade 2 items	36.9	23.2	-
L4	Writing	24.2	17.6	-
	Reading with comprehension	23.7	17.9	-
	Score in grade 1/2 items	49.4	44	-
	Score in grade 3 items	28.4	24.8	
	Score in grade 4 items	21.5	14.4	-
N4	Number concepts	45.1	42.4	
	Addition and subtraction	38.2	27.8	-
	Multiplication and division	27.5	16.9	-
	Score in grade 1/2 items	50.5	53.8	+
	Score in grade 3 items	34.6	26	-
	Score in grade 4 items	20.9	13.1	-

Note. - indicates a significant worsening in sub-scale scores; + indicates a significant improvement ($p < .05$).

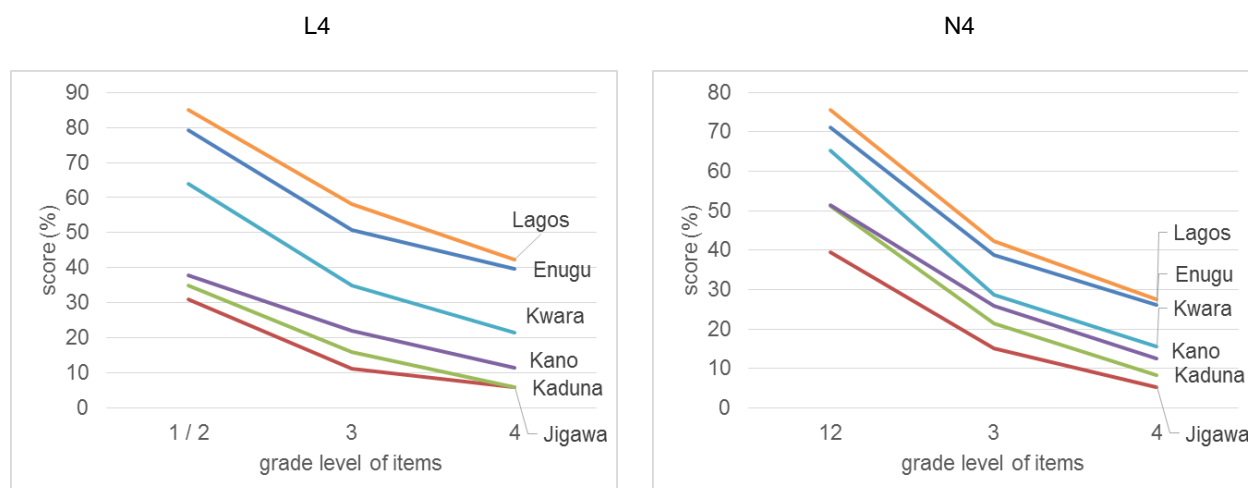
Children in ESSPIN schools achieved slightly higher scores in every sub-scale (Table 48); only in one case (grade 2 number concepts) does this reach statistical significance. Scores in Lagos and Kwara, where ESSPIN has been rolled out to all schools, are much higher, but the lack of a control group within those states prevents us from being able to attribute all of this difference to ESSPIN. In Enugu, Jigawa, Kaduna and Kano (taken together), only around a third of grade 2 pupils could answer grade 1 level questions, and little more than one-in-five could answer grade 2 level questions. In grade 4, pupils in ESSPIN schools scored 54% in grade 1 and 2 numeracy items, compared to under 50% for children in non-ESSPIN schools. In grade 4 literacy there was less difference between ESSPIN and non-ESSPIN schools in these four states: both groups only scored around 40% in the grade 1 and 2 items, and they only scored 12% for grade 4 items. By contrast, in Kwara and Lagos, grade 4 pupils scored over 70% in the grade 1 and 2 items, although, again, only a minority could complete grade 3 and 4 level items.

Examining the patterns by grade level for individual states (Figure 16), grade 4 children in Lagos score around 85% in grade 1 and 2 literacy items and around 75% in grade 1 and 2 numeracy items, but under 50% in grade 4 items in either test. But English literacy appears to be a particular issue in Jigawa, Kaduna and Kano, where grade 4 children score under 40% in grade 1 and 2 level questions, and under 15% in grade 4 level questions, suggesting that the majority cannot answer even the easiest questions correctly. In these three states, numeracy skills appear to be stronger than literacy, while the reverse is true in Enugu, Kwara and Lagos.

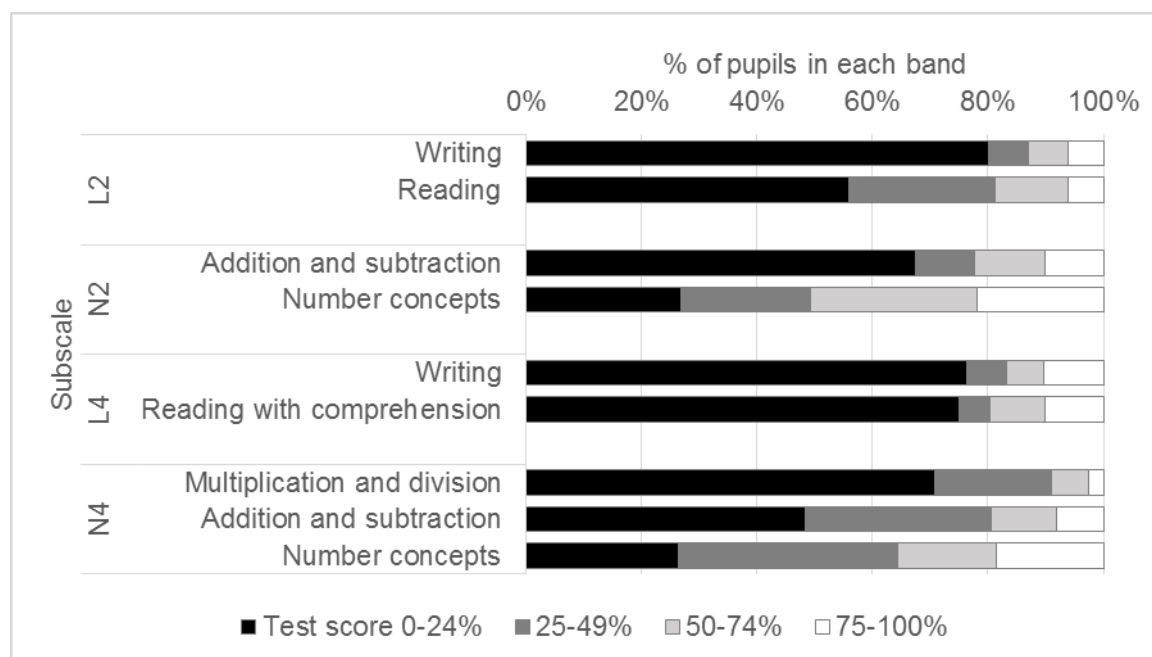
Table 48. Detailed scores in test sub-scales, ESSPIN vs. non-ESSPIN

Test	(%)	Enugu, Jigawa, Kaduna, Kano			Lagos, Kwara
		Non-ESSPIN	ESSPIN		ESSPIN
L2	Writing	9.5	8.6		41.5
	Reading	24.2	26.7		46.7
	Score in grade 1 items	32.7	35.8		63.1
	Score in grade 2 items	21.1	21.9		55.0
N2	Number concepts	42.8	51.3	+	70.2
	Addition and subtraction	18.4	19.6		50.0
	Score in grade 1 items	47.7	52.7		73.6
	Score in grade 2 items	19.8	23.8		43.8
L4	Writing	12.4	13.4		47.7
	Reading with comprehension	13.7	14.6		42.3
	Score in grade 1/2 items	38.1	39.9		77.5
	Score in grade 3 items	20.6	20.9		49.7
	Score in grade 4 items	10.8	11.8		34.8
N4	Number concepts	39.2	44.0		55.3
	Addition and subtraction	22.9	25.9		53.3
	Multiplication & division	14.8	15.2		28.5
	Score in grade 1/2 items	49.8	54.4		71.8
	Score in grade 3 items	23.7	25.5		37.3
	Score in grade 4 items	11.1	12.5		23.1

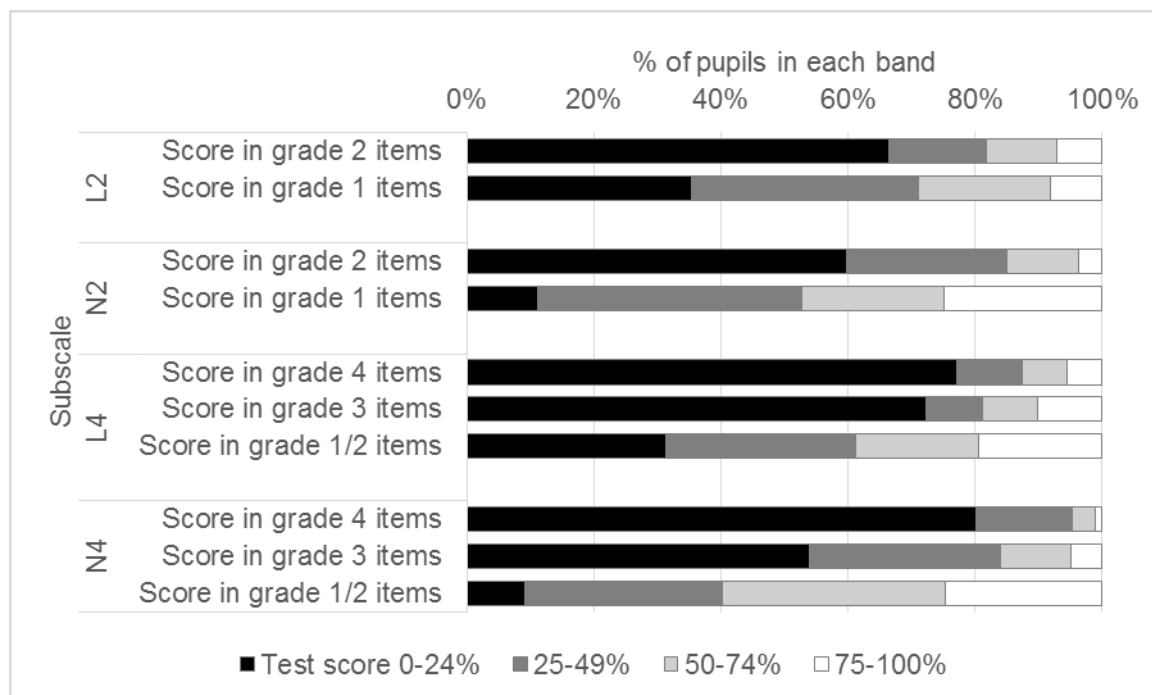
Note. + indicates a significant positive difference between ESSPIN and non-ESSPIN schools in sub-scale scores ($p < .05$). Lagos and Kwara are shown separately because there is no control group in those states. Detailed breakdowns are presented in Table 55 in Annex E.

Figure 16. Test scores by grade level of item and state in CS2 (grade 4 tests)

At both grade 2 and grade 4, more than three-quarters of pupils in the six states as a whole score under 25% in writing questions (Figure 17).¹⁷ A similarly large proportion of pupils score under 25% in grade 4 reading questions and grade 4 multiplication and division questions. Around half of children in grade 2 score over 50% in grade 1 level numeracy items, and around 60% of children in grade 4 score over 60% in grade 1/2 level numeracy items (Figure 18). This suggests that most children are learning numeracy skills, but lag behind the expected level for their grade. A similar pattern applied in literacy, but with lower scores all around: only around one-quarter of grade 2 students scored more than 50% in grade 1 items, and around 40% of grade 4 students scored more than 50% in grade 1/2 items.

Figure 17. Distribution of pupils by test score quartile for each learning domain

¹⁷ Figure 20 and Figure 21 in Annex E give breakdowns of these distributions by ESSPIN and non-ESSPIN status.

Figure 18. Distribution of pupils by test score quartile and grade level of questions

3.6.3 Summary and discussion

In summary, test results appear to be worsening over time in ESSPIN states, although measurement error cannot entirely be ruled out as a partial source of this change. Pupils benefiting from ESSPIN intervention have higher test results than those in schools that have not received ESSPIN intervention. The magnitude of the difference is not large within states, but is significant when we pool results across states. In the following chapter we confirm that this result is not just an product of differences between the states, by repeating the exercise with schools matched within the states. Moreover, there is evidence that pupil test results in schools with more ESSPIN intervention during the study period improved faster, or at least worsened more slowly—providing further support for a positive ESSPIN impact on learning outcomes.

Learning levels continue to vary greatly between states, and remain very low in some states. Few children achieve scores above 50% in questions pitched at the grade at which they are studying. There is likely to be a lag of *at least* a year for school improvement interventions to translate into better pupil learning outcomes, and of at least four years for the maximum effect of these interventions to be felt among pupils at grade 4 (since the maximum effect will come when pupils finishing grade 4 have been exposed to improved school quality throughout their school life). This means that the large scale-up of ESSPIN during 2013/14, in particular, is unlikely so far to have had an impact on state-wide pupil learning, which in turn helps to explain why we see worsening pupil test results across the states as a whole, combined with significant differences between ESSPIN and non-ESSPIN schools.

4 Controlling for confounding school characteristics and changes in enrolment

The previous chapter found significant differences between ESSPIN and non-ESSPIN schools in terms of pupil learning outcomes. However, it is still possible that ESSPIN schools, and the pupils in them, had different characteristics at the outset. Such differences in school characteristics can act as ‘confounders’ in our attempt to identify ESSPIN impact, because we cannot tell (using the methods in the previous chapter) if the true reason for a different outcome is the ESSPIN intervention or the school characteristics.

One possible confounder is the rise in pupil enrolment. In section 1.1 we noted that there have been large increases in pupil enrolment in the six ESSPIN states, and that this poses challenges for school improvement, especially if rising enrolment has not been matched by increases in the resources, classrooms and teachers. If enrolment increases have also differed between ESSPIN and non-ESSPIN schools, then this could bias our estimates of ESSPIN impact.

In this chapter, we first note general trends in pupil enrolment in the six ESSPIN states, and then examine whether enrolment and other school characteristics differ between ESSPIN and non-ESSPIN schools. In the following sections, we use a number of statistical methods to control for potential confounding in the analysis of pupil test results, and identify ESSPIN impact more robustly than in section 3.6.

4.1 Differences between ESSPIN and non-ESSPIN schools

Are the schools selected by ESSPIN for its interventions typical of the schools in each state? We attempt to answer this question by using information about a number of school characteristics taken from the annual school census in 2009/10 and 2013/14. The results (Table 49) suggest that in Enugu, ESSPIN schools are quite similar to typical schools in the state. In Jigawa and Kaduna, ESSPIN schools are, on average, closer to the Local Government Authority (LGA) headquarters, older, and larger, but with lower pupil–teacher ratios than the average. Urban schools and double-shift schools are over-represented, while nomadic and Islamic schools are under-represented. In Jigawa, teachers in ESSPIN schools are also more likely to have an academic diploma or degree; and ESSPIN schools are more likely than the state average to have had a parent–teacher association and SBMC in 2009/10.

In Kano and Lagos no non-ESSPIN schools remain, but we can compare those which have had ESSPIN interventions for longer to those that have had relatively little (and recently introduced) ESSPIN intervention. In Kano, the schools with more ESSPIN intervention (column *c* in the table) were older, much more likely to be urban and to be double-shift, but less likely to have had a parent–teacher association in 2009/10. They were bigger in terms of classrooms, teachers and pupils, but with lower pupil–teacher ratios than schools in Kano with more recent ESSPIN intervention. In Lagos, similarly, the ESSPIN schools in the initial stage (column *d* in the table) were disproportionately urban and larger. However, teachers in those schools were less likely to have a higher teaching qualification (PGDE, BEd or MEd) and there was no significant difference in terms of other types of qualification.

Thus, there are substantial differences in a number of characteristics between schools that received more and schools that received less ESSPIN intervention, in Jigawa, Kaduna, Kano and Lagos. Although ESSPIN selected schools include a wide range of school types, overall larger, urban, and longer-established schools are represented in larger proportions than would be expected if the schools had been chosen at random. School selection for ESSPIN was directed locally, by state and local governments, and the school selection process in each case is not

documented in great detail (but see ESSPIN 2013, Annex B). Government bodies may have sought to maximise impact by choosing larger schools – in the hope of reaching larger numbers of pupils and teachers – or by ensuring that schools were reachable by local staff. Alternatively, they may have been trying to represent a range of different schools within each LGA. Any such selection process would potentially result in selected schools differing from non-selected ones.

The differences in terms of some of these attributes between ESSPIN and non-ESSPIN schools potentially bias our estimations of ESSPIN impact. Where cross-sectional results in CS2 are better for ESSPIN schools than for non-ESSPIN schools, we cannot confidently attribute the difference to ESSPIN, because it could also be due to these other differences in attributes. Difference in differences helps to control for differences in school attributes, by removing any time-invariant differences. However, it still relies on an assumption of *parallel trends*: that the trend in ESSPIN and non-ESSPIN schools would have been the same if the ESSPIN intervention (between CS1 and CS2) had never happened. As we know that ESSPIN schools had both different attributes in 2009/10, and better learning outcomes and other indicators in 2011/12, this assumption can be questioned. It is therefore necessary to try and control statistically for the differences in school attributes.

In Kaduna and Kano, there has also been a rise in enrolment between 2009/10 and 2013/14 in schools with more ESSPIN intervention, which has significantly outstripped the rise in enrolment in schools with less intervention. As noted in section 1.1, there have been large increases in enrolment across the board in the three northern states (Jigawa, Kaduna and Kano). There are several possible explanations for why this increase was larger in ESSPIN-supported schools in Kaduna and Kano. ESSPIN supports schools to improve access, especially for children from disadvantaged backgrounds. An improvement in school functioning due to ESSPIN could also have encouraged more children to enter the school. Having improved access, schools may then have struggled to maintain quality. There were substantial increases in the pupil–teacher ratio in all states except Enugu, and in Kano and Lagos the pupil–teacher ratio increased fastest in schools with the most ESSPIN intervention—an effect which may to some extent have offset any effects of an improvement in school quality due to ESSPIN.

As already noted, increased pupil enrolment could be part of the explanation for reduced learning outcomes between CS1 and CS2. Furthermore, this may have affected ESSPIN schools more than non-ESSPIN schools. Such a large growth in enrolment is also likely to have involved a change in the composition of pupils entering government schools—for example, with more first-generation learners entering the system. Unfortunately we do not have data on these characteristics of the learners, so we restrict ourselves in the following analysis to considering the scale of the enrolment increase.

Table 49. Characteristics of ESSPIN and non-ESSPIN schools, by state

	Enugu		Jigawa			Kaduna			Kano			Lagos	
	a	bc	a	bc		a	bcd		b	c		c	d
Distance from LGA HQ	12.7	10.1	18.9	15.8	*	36.6	23.8	*	10.7	9.9		7.5	4.5
Age of school	48.7	54.8	29	35	*	21.1	26.9	*	22.1	28.1	*	44	46.2
Urban (%)	12.2	15.2	7.1	20.2	*	6.5	18	*	27.1	53.6	*	77.4	90.6
Nomadic (%)	1.3	1	10.5	4.1	*	5.9	4	*	3.9	1.6		0.1	0
Islamic (%)	0	0	5.2	4.4		1.1	0.1	*	52.1	42.9		0	0
Double-shift (%)	0.9	0.5	0.9	1.5		1.6	6.5	*	6	21.8	*	0.1	0
Had parent–teacher assoc. in 2009/10 (%)	94.1	93	95.1	97.3	*	95.8	94.8	*	97.7	92.1	*	87.5	92.7
Had SBMC in 2009/10 (%)	74.4	75	92.5	95.5	*	83.8	86.5		78.8	81.1		58.9	64.2
Pupil–teacher ratio in 2009/10	21	18	47.3	40.4	*	38.1	30.3	*	63.9	45.5	*	29	31.6
No. classrooms 2009/10	5.6	5.9	4.4	4.5		4.9	5.1		5.2	7.2	*	10.8	12.4
No. teachers 2009/10	10.5	12.3	4.4	8.2	*	7.2	15.3	*	8.6	15.1	*	13.7	16.6
Enrolment 2009/10	198	204	183	302	*	213	343	*	390	565	*	384	496
% change in enrolment 2009/10–13/14	6.4	-5.9	30.2	28.6		40.7	47.7	*	40.7	61	*	8	4.4
% of teachers with academic diploma / degree	39.9	46.1	32.3	40.3	*	44.5	45.7		47.6	50.9		40.8	34.7
% of teachers with PGDE, BEd or MEd	20.3	23.4	3	3.3		4.6	4.3		5.6	6		33	29.7
% of teachers with NCE, Grade II or equivalent	78.6	74.9	77.7	76.3		78.8	79.7		63	64.6		65.7	69.4

Note. a = no ESSPIN intervention; b = minimum intervention; c = medium intervention; d = maximum intervention (see Annex C). Source: annual school census, 2009/10 and 2013/14. * indicates a significant coefficient in a linear or logit regression of years of full ESSPIN intervention on the variable of interest. The figures shown are averages calculated at the school level, and so may differ from averages calculated at state level.

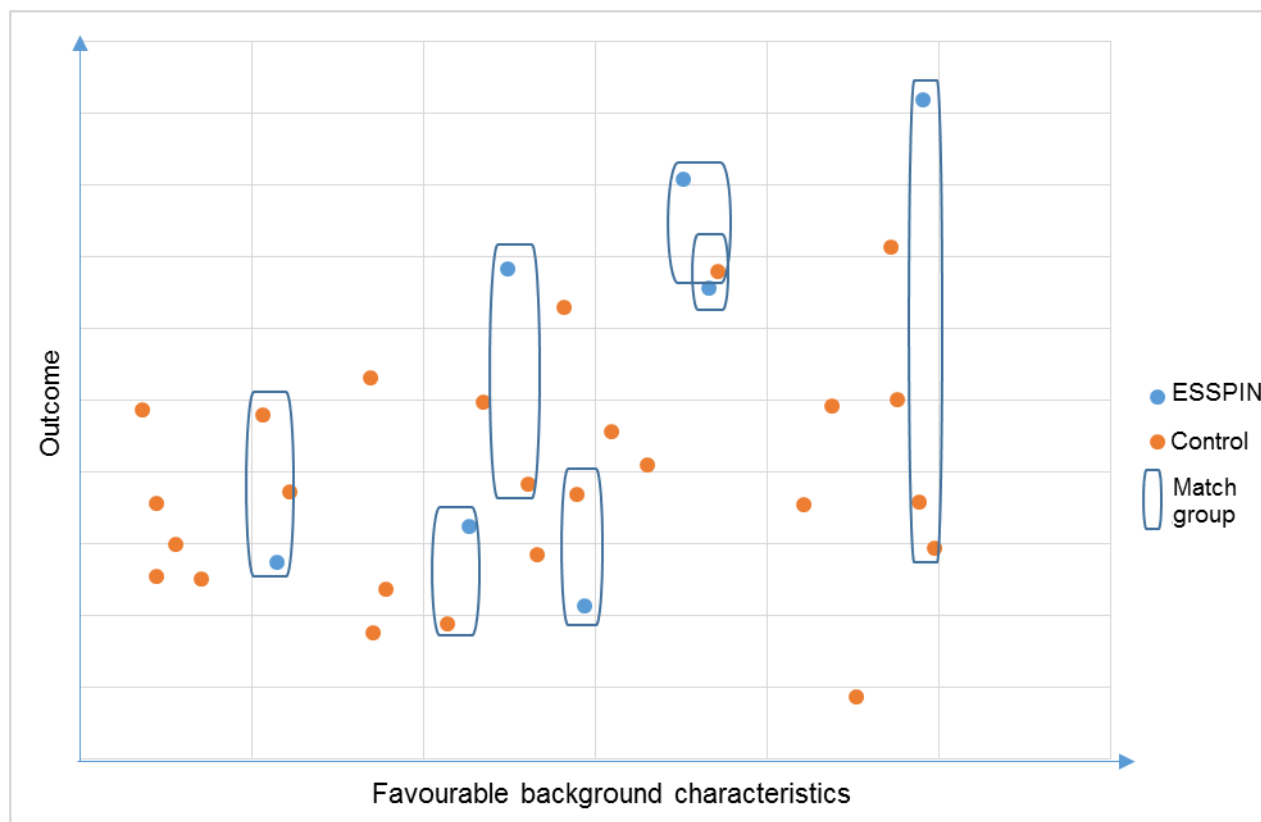
4.2 Controlling for school and pupil characteristics

In order to make more meaningful comparisons between ‘treatment’ schools – those receiving ESSPIN output 3 intervention in the relevant time period – and controls – those that did not receive such intervention – we use a number of statistical techniques that match schools and pupils according to their characteristics. The characteristics that we examine are taken largely from the 2009 annual school census, so that it is unlikely that the ESSPIN SIP could have influenced these characteristics. For characteristics that do not change over time – such as location and the year that the school was founded – we use data from the 2013 annual school census.

We use a combination of ordinary least squares regression analysis and propensity score matching. Regression analysis estimates the correlation of learning outcomes with ESSPIN intervention, conditional on school characteristics. In propensity score matching, non-ESSPIN schools that cannot be matched with a treatment school – because their characteristics are too

different – are either ignored, or assigned a low weight in the analysis. We use a statistical model to estimate the relationship between school characteristics and the likelihood of being an ESSPIN school (known as the school's propensity score). This allows us to compare ESSPIN schools to control schools that have characteristics more typical of ESSPIN schools, and vice versa (see Figure 19).

Figure 19. How matching techniques work (roughly)



4.2.1 Timing of ESSPIN intervention and learning outcomes in 2014

Analysing the effects of ESSPIN intervention is made complicated by the diversity of ESSPIN intervention between and within states. Among the schools that have received ESSPIN intervention, the timing and duration of the intervention varies considerably. How does the timing and duration of intervention affect the impact? We explore this using a regression analysis that compares each combination of timing and duration of intervention with schools that had no intervention (Table 50). We try two versions of the regression: the first controls for differences between the states, while the second controls for both the state and differences in school characteristics.

Impact appears to be highest for schools that received the intervention in 2011/12 and 2012/13, and those that received it in 2009/10 and 2010/11. Improved outcomes in 2014 even among the pilot schools (2009/10 and 2010/11) suggests that the effects of ESSPIN may be durable after the direct intervention has stopped. The schools that had a full package of intervention all the way from 2009/10 to 2012/13, which are located in Kaduna and Lagos, are also significantly better than control schools, but some of this effect appears to disappear once we control for schools' characteristics. As expected, interventions in 2013/14 are too recent for their effects on pupil learning to be felt in terms of a significant difference to the control group in the 2014 Composite Survey.

Table 50. Difference in test scores by timing of ESSPIN intervention

Years of output 3 intervention	Test scores in 2014 (percentage point difference compared to control schools)							
	L2		L4		N2		N4	
Controlling for state								
2013/14 only	1.3		3.6		3.4		1.8	
2012/13 and 2013/14	5.2	*	4.5	*	5.8	*	8.2	*
2011/12 and 2012/13	11.9	*	14.3	*	8.3	*	7.7	*
2009/10 and 2010/11	8.6	*	13.2	*	14.0	*	12.7	*
2009/10, 2010/11, 2013/14	4.3		11.5	*	5.8		5.6	
2009/10–2012/13	9.8	*	10.1	*	6.6	*	8.9	*
Controlling for state and other characteristics								
2013/14 only	-0.5		-0.5		-1.0		-3.3	
2012/13 and 2013/14	3.1		-0.7		2.3		4.6	*
2011/12 and 2012/13	10.5	*	7.8	*	3.4		1.2	
2009/10 and 2010/11	2.8		7.0	*	10.2	*	7.6	*
2009/10, 2010/11, 2013/14	-3.3		4.0		-3.5		-3.1	
2009/10–2012/13	8.5	*	4.8		2.3		3.6	

Note. The comparison group is the group of schools with no intervention. Estimates are based on regressions with state dummies and other school characteristics (an index of classroom infrastructure; distance from the LGA headquarters; the age of the school; urban/rural location; whether the school has a parent–teacher association; whether it is double-shift; the percentage of teachers with diplomas; the pupil–teacher ratio; and the number of classrooms) as covariates. For grade 4 pupils, we also control for pupil wealth. * indicates a significant coefficient ($p < .05$)

4.2.2 Are learning outcomes better in ESSPIN schools in 2014, controlling for school and pupil characteristics?

As explained in section 3.6 above, pupil test results are on average better in ESSPIN schools than non-ESSPIN schools. Is this finding robust to statistical controls for the fact that ESSPIN schools are disproportionately located in some states more than others, and for the different school and pupil characteristics described in section 4.1 above? We find significant differences between ESSPIN and control schools after using several techniques to control for state and school characteristics (Table 51). Controlling for the characteristics of schools tends to reduce the estimated effect but does not eliminate it—although it becomes non-significant in some cases, particularly for numeracy. Our preferred specification controls for state, school characteristics, pupil wealth (for grade 4 students only), and changes in the pupil–teacher ratio (row 9 in Table 51). (As noted above, enrolments have been increasing rapidly in some states, especially in ESSPIN-supported schools, and if the number of teachers did not increase at the same pace then this is likely to have had a negative impact on learning outcomes, potentially offsetting any positive impact of ESSPIN intervention.) This specification can only be calculated for the four states which retain a control group: Enugu, Jigawa, Kaduna and Kano. In those states, we find that pupils in ESSPIN schools have literacy scores approximately six to seven percentage points higher than those in control schools. The difference is not significant for grade 2 numeracy, but for grade 4 numeracy we find a difference of around two percentage points in favour of ESSPIN schools.

Table 51. Estimates of the effect of ESSPIN intervention on learning outcomes in CS2

Model	L2		L4		N2		N4	
(1) Control for state	7.27	*	9.49	*	5.66	*	6.74	*
(2) Control for state and school characteristics	4.22	*	5.29	*	1.72		3.28	*
(3) Matched (near-neighbour matching)	3.40	*	4.07	*	1.62		4.12	*
(4) Matched (propensity score matching)	2.48		4.52		2.09		1.26	
(5) Matched and controlled for state	5.22	*	5.6	*	2.63		1.47	
(6) Matched and controlled for school characteristics	3.03	*	5.63	*	3.05	*	2.17	*
(7) Matched, controlled for state and school characteristics	6.96	*	6.12	*	1.19		2.23	*
(8) Matched, controlled for state, school characteristics, and changes in enrolment between 2009/10 and 2013/14	6.98	*	6.01	*	1.18		2.03	
(9) Matched, controlled for state, school characteristics, and changes in pupil–teacher ratio between 2009/10 and 2013/14	7.2	*	6.26	*	1.00		2.19	*

Note. When controlling for state we only include Enugu, Jigawa, Kaduna and Kano, as the other two states do not have a control group for comparison. * indicates a positive coefficient in the regression ($p < .05$).

4.2.3 Have learning outcomes improved faster in schools with ESSPIN intervention, controlling for school characteristics?

In section 3.6 above, we controlled for outcomes at baseline by analysing whether learning outcomes are improving faster (or deteriorating more slowly) in ESSPIN schools than in other schools. We noted that there were some signs of positive difference in differences, but that this depended on which model was used. As noted in section 3.6, difference in differences analysis assumes parallel trends between different types of schools. In this section we try to control for the possibility that schools were set on different trends to start with, by matching and using regression controls for state and school characteristics.

Schools that received an intervention that affected pupils in the school between 2012 and 2014, when the surveys were conducted, should have improved more than schools that did not receive such intervention. This analysis is made difficult by uncertainty about how long it takes for the intervention to feed through into pupil learning outcomes. For example, schools that received an intervention in 2009/10 and 2010/11 might be continuing to improve through processes started by that intervention, such as a more active and effective head teacher, even if they have not received any intervention during our period of interest, 2012–2014. On the other hand, schools that did receive an intervention in 2013/14 are unlikely so far to have raised learning outcomes as a result, and may even have falling learning outcomes—for example, if teachers were occupied with training and temporarily reduced their teaching time.

To reduce this problem in regard to identifying impact, we restrict this part of the analysis to Enugu, Jigawa and Kaduna states, where there is a clean comparison between a control group that has never had an intervention, and an ESSPIN group that has had an intervention within the 2012–14 period (Table 52). We ignore schools that have had intervention only in 2013/14, and schools that had an intervention during the pilot phase.

We then apply a similar set of matching and regression techniques to analyse change over time as we used to analyse the CS2 results in the previous section (Table 53). The results are somewhat surprising: the difference between the change over time in ESSPIN schools and the change over

time in control schools is generally not significant, except for grade 4 literacy, where it is significant and negative—that is, ESSPIN schools have improved less rapidly than control schools.

Table 52. Identification of control and ESSPIN groups for difference in differences analysis

	Control group	ESSPIN group	Ignoring
Enugu	No intervention	Intervention in 2011/12 and 2012/13	None
Jigawa	No intervention	Intervention in 2012/13 and 2013/14	Pilot (2009/10 and 2010/11) schools
Kaduna	No intervention	Intervention in 2011/12 and 2012/13 or in 2012/13 and 2013/14 ¹⁸	Schools that only had an intervention in 2013/14; schools included in the pilot phase

Table 53. Estimates of the effect of ESSPIN intervention on changes in learning outcomes between 2012 and 2014

Model	L2	L4	N2	N4
(1) Control for state	-3.18	-3.63	-2.49	-0.12
(2) Control for state and school characteristics	-3.47	-6.6 *	-5.65	-1.97
(3) Matched (propensity score matching)	-6.74	-11.68 *	-7.26	-5.13
(4) Matched and controlled for state	-6.26	-11.67 *	-6.54	-5.03
(5) Matched and controlled for school characteristics	-8.11	-12.61 *	-8.11	-5.86
(6) Matched, controlled for state and school characteristics	-6.11	-11.85 *	-5.76	-5.13
(7) Matched, controlled for state, school characteristics, and changes in enrolment between 2009/10 and 2013/14	-6.11	-11.8 *	-5.73	-5.1
(8) Matched, controlled for state, school characteristics, and changes in pupil–teacher ratio between 2009/10 and 2013/14	-5.86	-12.1 *	-6.16	-5.81

Note. * indicates a significant coefficient ($p < .05$)

What explains these findings? Looking at the mean test scores for grade 4 literacy within each state (Table 54), these mean scores suggest that the overall findings are mainly driven by Enugu, where control schools improved quite rapidly, while ESSPIN schools remained about the same, and Kaduna, where there was a worsening of test scores in both control and ESSPIN schools, but where the worsening was more severe in ESSPIN schools. As noted in section 4.1 above, there were large enrolment increases in Kaduna between 2009/10 and 2013/14, which may have made it harder for schools to maintain quality. However, this does not explain the apparent worsening in ESSPIN schools that remains even after we attempt to control for changes in the pupil–teacher ratio (row 8 in Table 53 above). This may merit further investigation.

¹⁸ It is possible, although unlikely, that schools that received an intervention during 2011/12 may have already benefited from ESSPIN intervention when the first composite survey was conducted in 2012, which would dilute our estimated impact. We check for this in Kaduna by examining the different intervention groups (no intervention, vs. 2011/12 and 2012/13, vs. 2012/13 and 2013/14) separately, but we find non-significant or negative effects for both intervention groups.

Table 54. Grade 4 literacy test results in control and ESSPIN schools, in 2012 and 2014 (%)

	Control		ESSPIN intervention during 2011/12–2012/13	
	2012	2014	2012	2014
Enugu	46.1	57.0	67.3	68.7
Jigawa	17.6	17.1	22.8	20.4
Kaduna	29.3	21.3	36.1	20.2
Total	28.1	26.5	32.9	21.9

However, there are some reasons why we may need to be wary of the analysis based on changes over time. It may be that it takes more than one year for ESSPIN intervention to feed through into a measurable effect on learning outcomes. Our comparison is inherently noisy because we assess a different random sample of 16 pupils within each school in 2014 than we were assessed in 2012. ESSPIN schools had a higher starting point, and it is not clear whether an increase of, say, 10 percentage points, towards the top of the scale in each test, has the same meaning in terms of pedagogical content as a 10 percentage point change towards the bottom of each scale. Finally, it is not clear that the unobserved characteristics that affect learning outcomes are time invariant. For example, a more motivated head teacher might cause learning outcomes to improve continuously. The effect of such a head teacher (relative to a school with an unmotivated head teacher) would be a continuous change rather than a one-off improvement, so it does not make sense to think of head teacher motivation as a time-invariant school characteristic.

4.2.4 Summary

This chapter has shown that ESSPIN schools differ from non-ESSPIN schools in several characteristics, making it important to check whether these differences pre-dating ESSPIN's interventions might underlie the differences in learning outcomes between the two groups of schools in 2014. It has used several different matching techniques to examine whether differences between ESSPIN and non-ESSPIN schools are robust to controlling for characteristics. Reassuringly, the different methods all point in the same direction with respect to the analysis of CS2 results: ESSPIN schools have better learning outcomes than non-ESSPIN schools even after controlling for differences in characteristics.

When, instead of the cross-sectional analysis of CS2 results, we turn to an analysis of the change over time, between CS1 and CS2, the results are much less clear, sometimes negative, and generally not statistically significant. Thus, we do not have robust evidence that ESSPIN schools have improved faster than non-ESSPIN schools during 2012–2014, when we control for school characteristics. As has already been discussed in the previous chapters, this may reflect the difficulty of measuring change over a relatively short time period, especially among schools that vary in terms of their starting levels.

We are not able to find evidence that changes in enrolment during 2009–2013 have had a significant impact on learning outcomes, and neither is there any evidence that they reduce the effects of ESSPIN programmes. Treatment effects retain the same sign and significance after controlling for indicators of changes in pupil enrolment. This is surprising given our expectation that pupil enrolment increases would be likely to put pressure on schools' attempts to maintain high quality. Further research on this question may still be warranted, including looking at indicators of enrolment change over periods other than 2009–13. It is possible that changes in the profile of the enrolled pupils, more than changes in the numbers enrolling or in the pupil–teacher ratio, matter for learning outcomes.

5 Conclusion and implications of survey findings for ESSPIN programme

This report has examined a wide range of indicators – teacher competence, head teacher effectiveness, school development planning, SBMC functioning and inclusiveness, and pupil learning – in the two rounds of the Composite Survey, in 2012 and 2014. It has asked whether things were getting better over time across the six states; whether ESSPIN schools are currently doing better than non-ESSPIN schools; and whether schools with more ESSPIN intervention during the relevant period are improving faster than non-ESSPIN schools. In addition, it has examined the issue of whether confounding characteristics and pupil enrolment changes could, to some extent, be driving changes in learning outcomes.

Teachers, on average, have not become significantly more or less likely to meet our competence standard in 2014 than they were in 2012. They appeared to improve in two criteria – use of teaching aids and use of more praise than reprimands in the classroom – but English and mathematics teachers' knowledge of benchmarks in their subjects appeared to have worsened. However, teachers in ESSPIN schools, and especially those who individually received training from ESSPIN, do better in most criteria than those in non-ESSPIN schools. There is tentative evidence that teachers benefiting from ESSPIN interventions improved faster (or at least, worsened more slowly) between 2012 and 2014. The bleak situation overall in ESSPIN states is tempered by ESSPIN interventions that appear to be holding up teacher competence to some extent.

We introduced a new teacher content knowledge test for CS2, and find that in ESSPIN schools teachers' knowledge in English and mathematics is much better than in non-ESSPIN schools, although it remains low across the board: teachers in non-ESSPIN schools scored only 42% in an English test pitched at primary-grade level, while those trained by ESSPIN scored 55%. In mathematics, teachers got a majority of questions right, and those trained by ESSPIN scored over 70%, but there were still clearly primary school level numeracy questions that teachers struggled with.

Head teacher effectiveness appears to be improving over time across the ESSPIN states, and is better in ESSPIN schools than in non-ESSPIN ones. There is also evidence that head teacher effectiveness improved faster with more ESSPIN intervention. We are not able conclusively to disentangle whether this effect should be attributed to the specific component of leadership training, or to the overall package of ESSPIN interventions that aim to improve school quality.

School development planning also appears to be improving in ESSPIN states, and is much better in ESSPIN than in non-ESSPIN schools. There is suggestive evidence that the pace of improvement may be faster in schools that received more ESSPIN intervention between 2012 and 2014, but this depends on the method used in the analysis.

We measure **school inclusiveness** using a standard that reflects actions and plans to improve attendance, use of different assessment methods by teachers, participation in lessons by children in different parts of the classroom, and equal participation of girls and boys in lessons. Fewer schools met the inclusiveness standard in 2014 than in 2012. In 2014, 25% of ESSPIN schools, but only 8% of control schools, met the standard. The drop in our inclusiveness score during 2012–2014 was also smaller in ESSPIN than in non-ESSPIN schools, suggesting that whatever factors are driving the decline in our inclusiveness measure are having less of an impact in ESSPIN schools than in others.

SBMCs appear to be functioning better in 2014 than they were in 2012, and much better in ESSPIN schools than in non-ESSPIN schools. In 2014, 62% of ESSPIN schools met the SBMC

functionality standard, compared to only 13% of non-ESSPIN schools. Both ESSPIN and non-ESSPIN schools are improving over time, but ESSPIN schools appear to be improving more quickly. SBMCs in ESSPIN schools were also more likely to be inclusive of women, with 48% meeting the standard, compared to only 2% in non-ESSPIN schools, and also more inclusive of children, with 18% meeting the standard, compared to 2% in non-ESSPIN schools. SBMCs in ESSPIN schools also improved in terms of their inclusiveness during 2012–2014, while there was no significant change over time in non-ESSPIN schools.

The proportion of schools meeting ESSPIN overall **school quality** standards increased significantly, from 3% to 10%. However, there was no change in the average levels of a continuous measure of school quality based on the same criteria. In 2014 there were significant and large differences in quality between ESSPIN and non-ESSPIN schools. Only around 1% of non-ESSPIN schools meet the quality standard, compared to over 30% of ESSPIN schools. The pace of improvement between 2012 and 2014 has also been much faster in schools which received more ESSPIN intervention during that period than in schools which received less. Non-ESSPIN schools have seen little improvement in quality during this period, while ESSPIN schools have improved substantially.

Pupil test results in grade 2 literacy, grade 4 literacy, grade 2 numeracy and grade 4 numeracy, suggest that pupil learning outcomes may be worsening over time in ESSPIN states. Children in ESSPIN schools have significantly better test results than those in non-ESSPIN schools across all four types of test, however. There is some evidence that pupil learning, especially in literacy, may be improving faster in schools that received more ESSPIN intervention than in those that received less. The difference in results in 2014 between ESSPIN schools and non-ESSPIN ones remains even after controlling for possible confounding due to differences in school characteristics, but the results when examining the rate of improvement over time are less positive.

We noted rapid increases in pupil enrolment in ESSPIN states and we suggest that this may be part of the explanation for the slowing pace of improvement, but we are not able to find solid evidence for this hypothesis. This remains an issue that demands further research, as we are only able to control for numbers of pupils, and not for possible changes in the profile of new pupils. Although there are clear differences in learning outcomes between ESSPIN and non-ESSPIN schools, and some scale-up of ESSPIN, during 2012/13, the SIP has not yet reached a scale where it can push up averages for the states as a whole. This could change as a result of the larger scale-up during 2013/14, provided that the programme's effects are not too diluted by the changes to delivery mechanisms that are necessary for reaching such a large scale.

Overall, ESSPIN's outcome and output indicators are mostly improving over time, and are better in ESSPIN schools than in non-ESSPIN schools. However, pupil test results – the main impact indicator for ESSPIN's SIP – are actually worsening over time in the ESSPIN states. Nevertheless, pupil learning is better in ESSPIN than non-ESSPIN schools and this does not appear to be an product of the type of schools that ESSPIN has selected to work in. Several indicators appear to have been improving faster in ESSPIN schools than in non-ESSPIN schools during the period between 2012 and 2014, which is again supportive of an ESSPIN impact in these areas. It is plausible to suppose that the SIP was able to achieve rapid gains in the first part of its cycle by focusing on relatively solvable problems, but is now entering a phase where more entrenched and difficult problems act as barriers to further improvement in learning outcomes. Low teacher literacy in the language of instruction, for example, may take time to address, as will long-term contextual factors, such as persistent conflict in some of the ESSPIN states. It remains to be seen whether, as ESSPIN scales up, its impact will be sufficient to reverse these trends at the state level.

References

- De, S. and Cameron, S. (2015) 'ESSPIN Composite Survey 2: Gender and Inclusion Report'. Oxford Policy Management.
- ESSPIN (2013a) 'Overall findings and technical report of ESSPIN composite survey 1 (2012)'. Report number ESSPIN 060.
- ESSPIN (2013b) Extension of the Education Sector Support Programme in Nigeria, August 2014 – January 2017. Business case for DFID.
- Megill, D. (2014a) 'Recommendations on Sampling Plans for the Second Composite Survey and the GEP-3 Unified Survey (GUS) in Nigeria'. Unpublished note written for ESSPIN and OPM.
- Megill, D. (2014b) 'Final sample design and weighting procedures for ESSPIN Second Composite Survey (CS2) in Nigeria'. Unpublished note written for ESSPIN and OPM.
- RTI International (2014) 'Nigeria reading and access research and activity (RARA): adaptation of Education Sector Support Programme in Nigeria's teacher capacity development strategy'. Document produced for the United States Agency on International Development (USAID)

Annex A Indicators

A.1 Pupil learning logframe indicators

Test items used to calculate the ESSPIN logframe pupil learning indicators

The relevant questions are given below. The *[Italics]* underneath each question set out the criterion used to classify a pupil as demonstrating the particular skill being tested. The question numbers refer to CS1 and were altered slightly for CS2. The content of each question was not altered.

P2 Literacy

Proportion of p2 children who demonstrate skills for reading

comprehension: Proportion of p2 children who correctly answer a p2 curriculum level question on listening comprehension (Q11) and correctly read a sufficient number of words from a p2 curriculum level passage (Q13).

P2 Q11 Listening comprehension

This is an oral question and answer. **Do not** show this page to the pupil.

Ndi has two brothers. Their names are Paul and Raymond. They are older than Ndi. Ndi likes to go to school, because she has many friends there.

Ask the pupil:

11a) How many brothers does Ndi have?

11b) Why does Ndi like to go to school?

[Pupil must answer both parts correctly].

P2 Q13 Reading a passage aloud

Read the following passage aloud:

Good morning. My name is Fatima. I am

seven years old. My brother's name is Sam.

He is five years old. I also have a sister. Her

name is Nandi. We like to read stories. We go

to the market every Saturday. My mother sells

fruit at the market in town.

[Pupil must get 26 or more words correct].

P4 Literacy

Proportion of p4 children who demonstrate ability to read with

comprehension: Proportion of p4 children who correctly read a sufficient number of familiar words at p4 curriculum level (Q21) and correctly read a sufficient number of words from a p4 curriculum level passage (Q23) and correctly answer at least four out of five reading comprehension questions (Q23).

P4 Q21 Pronunciation (reading)

Ask the pupil to read the words as quickly and carefully as they can, going along the row.

back	glass	quick	small	start
fall	vary	shot	bird	miss
animal	calendar	beginning	introduce	medicine
chicken	their	carry	handle	rhyme
apple	banana	grass	yellow	orange
mistake	sugar	tangle	hospital	through

[Pupil must get 15 or more words correct].

P4 Q23 Reading with comprehension

My name is Umar. I live on a farm with my mother, father and sister Fatima.

Every year the land gets very dry before the rains come. We watch the sky and wait.

One afternoon as I sat outside, I saw dark clouds. Then something hit my head, lightly at first and then harder.

I jumped up and ran towards the house. The rains had come at last.

[Pupil must get 34 or more words correct and answer at least four out of five comprehension questions correctly].

P2 Numeracy

Proportion of p2 children who demonstrate the ability to do basic

arithmetic calculations: Proportion of p2 children who correctly answer at least five out of six p2 curriculum level questions on addition and subtraction (Q14) and both multiplication questions (Q15).

P2 Q14 Addition and subtraction of two and three-digit numbers

The pupil will write the answers to the sums on this page in the spaces.

14a) $32 + 16 = \underline{\hspace{2cm}}$

14b) $25 + 7 = \underline{\hspace{2cm}}$

14c) $234 + 342 = \underline{\hspace{2cm}}$

HTU

234

+ 342

14d) $19 - 6 = \underline{\hspace{2cm}}$

14e) $16 - 8 = \underline{\hspace{2cm}}$

14f) $49 - 22 = \underline{\hspace{2cm}}$

[Pupil must get at least five out of six sums correct].

P2 Q15 Multiplication of single-digit numbers

The pupil will write the answers to the two sums on this page in the spaces provided.

Multiply these two numbers together.

15a $3 \times 2 = \underline{\hspace{2cm}}$

15b $4 \times 4 = \underline{\hspace{2cm}}$

[Pupil must get both sums correct].

P4 Numeracy

Proportion of p4 children who demonstrate the ability to do basic arithmetic calculations: Proportion of p4 children who correctly answer p4 curriculum level questions on addition and subtraction (Q25) and multiplication (Q26) and division (Q27).

P4 Q25 Addition and subtraction

a)

	3	2	3
	2	1	4
+	1	6	1
	<hr/>		
	<hr/>		

b)

	3	4	0
	6	4	3
+	6	3	4
	<hr/>		
	<hr/>		

c)

	4	3	2
-		6	1
	<hr/>		



[Pupil must get all three sums correct].

P4 Q26 Multiplication of numbers

a)

	1	4
x		7

b)

	3.	42
x		2

[Pupil must do both multiplications correctly].

P4 Q27 Dividing numbers

a) 4 / 68 = _____

[Pupil must do division correctly].

A.2 Total test scores

Total test scores were calculated on the basis of one mark per question, and using the marking criteria devised for CS1. CS1 scores were recalculated to exclude questions that were not included in CS2 and to apply the same skip patterns to both tests.

Annex B Note on changes to assessments for CS2

B.1 Introduction

The assessment instruments used for the first round of the ESSPIN Composite Survey (CS1) were recognised as being valid tools for assessing pupils' abilities in maths and English literacy, and as having generated useful data. However, a number of criticisms were raised by ESSPIN, in the CS1 review, analysis of the data, and during training of coordinators for CS2. These include:

- Construct validity: based on examination of the content of the questions, assessment experts doubt whether some questions measure what they purport to measure.
- Multidimensionality and redundancy: based on statistical analysis of the results from CS1 using Rasch modelling, the tests appear to measure more than one underlying competency, and some questions have been revealed to be redundant – in that they do not add much power to discriminate between pupils at different levels of ability.
- Length of tests: During piloting for CS2, tests for grade 4 children took over an hour in some cases. This is boring for the child, may cause distress in cases of children who are unable to answer (but are compelled to continue anyway), and is likely to reduce data quality.
- Difficult to administer: Instructions to data collectors were in some cases unclear or inconsistent; for example instructions in the 'pupil books' are sometimes different from those in the 'guide for data collectors'. Data collectors for CS1 were asked to mark questions and calculate question scores in the pupil book itself, which was unnecessary and may have affected data quality.

The use of CAPI for CS2 – instead of the paper test sheets used for CS1 – also represented an opportunity to tailor the tests to pupils' abilities—for example, administering different questions to a pupil depending on how successfully he or she answered earlier questions.

However, there were limits to the extent to which these issues could be addressed given the need for comparability with CS1. Modest changes were therefore made to the guidelines for data collectors, the form of the tests, and the questions.

B.2 Statistical analysis

Analysis of data from CS1 using the set of statistical techniques known as Rasch modelling, by consultant Joshua McGrane, returned the following overall assessment of the instruments.

- L2 was somewhat difficult, lacking questions at the lower end of the ability scale. Few items seemed to be redundant, although questions 8 and 9 and questions 4 and 13 were highly correlated. Deletion of questions did not seem to be advisable. Fit to the Rasch model was reasonable, indicating that this test functions as a single coherent scale.
- In L4 there was substantial 'misfit', suggesting that these items did not fit together into a single coherent scale; there was evidence of multidimensionality. There was a good spread of items at different difficulty levels but a relative lack of 'easy' questions (questions discriminating between the weakest pupils). Several items were recommended for deletion on the grounds that they did not fit the model well and did not seem to add to the discriminating power of the test.
- N2 also did not fit the Rasch model well, suggesting multidimensionality. A few misfitting items that were not highly discriminating were recommended for deletion.

- N4 again did not fit the Rasch model well. However, the items appeared to have been well targeted, with a spread across easier and harder items. Several misfitting items that were not highly discriminating were recommended for deletion.

In short, the test items appeared on the whole to have been fairly well targeted, with a spread of easy and difficult items; but three of the four tests exhibited some misfit from the Rasch model, suggesting that they do not measure a single concept. This should perhaps not be taken as cause for alarm: it suggests that the different learning domains covered in each test are separable and different pupils progress in some faster than in others.

B.3 Question removal and skip patterns

Careful consideration was given before removing any questions from the test. However, given the length of the tests, especially for grade 4, there was a potential strong pay-off, in terms of data quality, as a result of removing questions. The criteria applied for removal were as follows:

- The question is recommended for deletion in Joshua McGrane's report, on the basis that it is 'lowly discriminating' – does not help us distinguish between pupils at different levels, and may also be 'misfitting' – removing it improves the overall fit of the test.
- It is not part of a logframe indicator.
- It is not the sole indicator of a particular learning domain (e.g. grade 2 reading).

On this basis, the following were removed:

- L2: none
- L4: 4 (making a sentence); 14 ('should'); 26 (dialogue)
- N2: 3 (fractions); 17 (capacity); 19 (shapes)
- N4: 2 (sequences); 13 (greater/less than); 18 (length); 19 (area); 21 (graph); 31 (angle); 32 (graph)

Based on further analysis of the CS1 data, some 'skip patterns' were identified. Cases were identified where children scoring less than some threshold in an 'easy' set of questions had a very strong tendency to score zero in a more difficult set of questions. For example, among the 28% of children who were not able to read any letters in an early question in the literacy test, their average scores in reading words (question 4), answering an oral question with a written answer (question 14) and spelling (question 16), were all under 0.03 (i.e. fewer than 3% got the question right).

In particular:

- L2: children who are not able to identify *any* letters in question 3 are allowed to skip questions 4, 14, 16
- L4: children who are not able to identify *any* words in question 3 are allowed to skip questions 8, 11, 12, 13, and 14
- N4: children scoring less than 1.25 points in the first 11 number questions (questions 1, 3, 4, 5, 6, 12, 13, 14, 15, 17) are allowed to skip the harder number questions (questions 23, 27, 28, 29, 30).

It might be argued that even if only 3% of pupils are unable to name letters, but are able to answer a written question with an oral answer, that such pupils should still be allowed to attempt all of the questions so as not to lose this data for the 3%. Against this perspective, it can be argued that some of this 3% likely represents data collector error, and that it is ethically dubious to make

children sit through a long test where they lack the basic ability required for most of the questions, if that test can be shortened for them. Ideally a better test design would ask questions at incremental difficulty levels in different domains, stopping at the point where the pupil is unable to answer. Repeating the CS1 test, but with some questions deleted and some skip patterns inserted, represents a reasonable compromise between comparability and progressive improvement in test quality. In order to compare the CS1 and CS2 test data, the same skip patterns and deletions have been applied to both, using statistical software.

B.4 Pupil background

Three questions were added on pupil background:

- age;
- language spoken at home; and
- (for grade 4 children) what items are owned by the family. The item list is derived from the 2010 Nigerian Living Standards Measurement Survey, with one addition (mobile phone) and several items removed that might be difficult for a 10-year-old to identify.

Data on the answers to these questions will be available for use as covariates in the analysis of the pupil test scores.

B.5 Disability

After discussion with ESSPIN and external disability specialists several questions were added at the beginning of each test. These do not aim to assess disability – which would require a much larger set of questions, with follow-up by specialists – but focus on ability to take the test. In each case, children who do not have a particular ability were not made to sit through questions that required that ability: e.g. children who cannot see were not asked to read a text from the book. Whereas in CS1 the question of disability was not explicitly addressed, for CS2 it was made clear to data collectors that children with disabilities needed to be included in the test, but that where they did not possess particular abilities for answering specific questions, they should not be asked to answer such questions.

The data collector manual included careful instructions for data collectors to use their common sense when applying the ability test questions. For example, a child who responded when their name was called but did not respond in any way to a greeting, would not be dismissed as unable to hear, but presumably was shy or unable to understand the greeting.

B.5.1 Ability to hear

All of the questions in both tests require the pupil to understand spoken instructions. For children who were unable to hear, or understand through some other means such as lip-reading or with the aid of a signing assistant, the CAPI software therefore skipped to the end of the test.

Interviewer: While leading the pupil to the test location, greet the child and ask his/her name, using the local language where possible.

When you first speak to the pupil, if the pupil shows no signs of hearing what you say, confirm with the teacher whether the pupil can hear.

If he or she cannot hear but can understand through some other means – lip-reading or through a signing assistant – continue with the test.

If the pupil is not able to understand you, give them the biscuit, drink and pencil, and thank them for their participation.

B.5.2 Ability to speak

Pupils who appeared unable to speak at all were not given questions that required a spoken answer. In the numeracy tests, their ability to speak was checked in the following way:

Prompt if necessary until you get a reply to your greeting.

Mark whether the pupil responds verbally to your greeting and/or says his name.

In the literacy tests this check on ability to speak was combined with a question that checks whether the child can respond in English:

Greet the child again in English:

Good morning / good afternoon

Prompt if necessary until you get a reply to your greeting.

Mark whether the child responds with 'Good morning/afternoon' or any culturally appropriate greeting; responds verbally but inappropriately; or does not respond at all.

If the pupil responded verbally but inappropriately, they were marked incorrect but the test continued without skipping any questions. If they did not respond at all, they were marked as unable to speak, and spoken questions in the test were subsequently skipped.

B.5.3 Ability to see

Children were asked the following question to gauge whether they could see well enough to take the test. If they could not, they were subsequently asked only questions that could be asked orally and that required an oral response.

I am going to ask you lots of number questions. I will ask you to write or say the answers. You should try your best but do not worry if you cannot answer.

Can you see the book here?

Point to the pupil book on the table.

Mark whether the pupil indicates that he/she can see the book, by looking at it and/or saying yes.

B.5.4 Ability to write

The following question tested whether children had the physical ability to hold a pencil and mark the page. For children who could not do this, questions requiring writing were skipped.

Interviewer: Turn to the 'Drawing' page and give the child the pencil.

I'm going to draw a line between these two dots.

Interviewer: draw a line on the pupil book between the two dots at the top of the page.

Now, can you do the same and draw a line between these two dots?

Interviewer: point to the two dots lower down the page.

Mark whether the pupil draws or writes something in the book, regardless of whether it is a straight line between the two dots or something else.

In practice, more than 99% of the pupils showed all four abilities and took the full test.

Annex C ESSPIN output 3 interventions

The table below shows the ESSPIN output 3 interventions delivered to date in each state. In order to make the variation in interventions across and within states manageable for analysis, each combination of interventions was categorised as none, minimum, medium, or maximum, according to the number of years of continuous intervention.

	Category	2009/10			2010/11			2011/12				2012/13			2013/14				EB	EF
		L	T	SV	L	T	SV	L	T	SV		L	T	SV	L	T	SV			
Enugu	None										CS1							CS2	No	No
	Minimum														6	3	9		No	No
	Medium (1)							6	3	9		6	3	9	3		9		Yes	Yes
	Medium (2)											6	3	9	6	3	9		Yes	Yes
Jigawa	None																		No	No
	Minimum														6	3	9		No	No
	Medium (1)											6	3	9	6	3	9		Yes	Yes
	Medium (2)	5*	5*	9*	10*	5*	9*												Yes	No
Kaduna	None																		No	No
	Minimum														6	3	9		No	No
	Medium (1)											6	3	9	6	3	9		Yes	Yes
	Medium (2)									6		6	3	9	3		9		Yes	Yes
	Maximum	5*	5*	9*	10*	5*	9*	6	3	9		6	3	9	3		9		Yes	Yes
Kano	Minimum														9	9	9		No	No
	Medium	5*	5*	9*	10*	5*	9*								9	9	9		Yes	No
Kwara	Medium	6	3	30	6	3	30			30				30	6	3	30		Yes	No
Lagos	Medium (1)											6	3	9	6	3	9		Yes	Yes
	Medium (2)							6	3	9		6	3	9	3		9		Yes	Yes
	Maximum	5*	5*	9*	10*	5*	9*	6	3	9		6	3	9	3		9		Yes	Yes

L = days of leadership training; T = days of teaching training; SV = school visits; * = pilot.

EB = Expected to be better than comparison schools in 2014; EF = expected to improve faster than comparison schools during 2012–14.

Annex D ESSPIN output 4 interventions

The table below shows the days of output 4 intervention in each state under different headings: SBMC training; women and children participation training; and mentoring visits.

	Category	2010/11			2011/12				2012/13			2013/14				De facto phase	Level of output 3 intervention
		S	P	M	S	P	M		S	P	M	S	P	M			
Enugu	None							CS1							CS2	Control	None
	2c											7		4		Post-CS1	Min
	2b								7		4	r	6	4		Post-CS1	Med
	1 / 2a				7		4		r		4		6			Pre-CS1	Med
Jigawa	None															Control	None/min
	2b											2				Control	Min
	2a											7		4		Post-CS1	Med
	1	7		4	r		4			6	4*			4*		Pre-CS1	Med
Kaduna	None															Control	None/min/med
	2a				7		4		r		4		6	4*		Pre-CS1	Med
	1	7		4	r		4			6	4*			4*		Pre-CS1	Max
Kano	None															Control	Min
	2b											r	6			(Unknown)	Min
	2a											7		4		Post-CS1	Min
	1				7		4		r	6	4			4*		Pre-CS1	Med
Kwara	None															Control	Med
	2b											7				Post-CS1	Med
	2a											4		2		Post-CS1	Med
	1				7		4		r		4		6	4*		Pre-CS1	Med
Lagos	2b											7		4		Post-CS1	Med
	1 / 2a				7		4		r		4	7		4		Pre-CS1	Max/med

Note. S = SBMC training. P = women and children participation training. M = mentoring visits. r = one-day refresher. Mentoring visits were by civil society-government partnership teams, except those marked with an asterisk, which were by social mobilisation officers. There is missing intervention data for Kano group 2b.

Annex E Additional tables and figures: pupil test results in CS2

E.1 Disaggregated scores by sub-scale and state

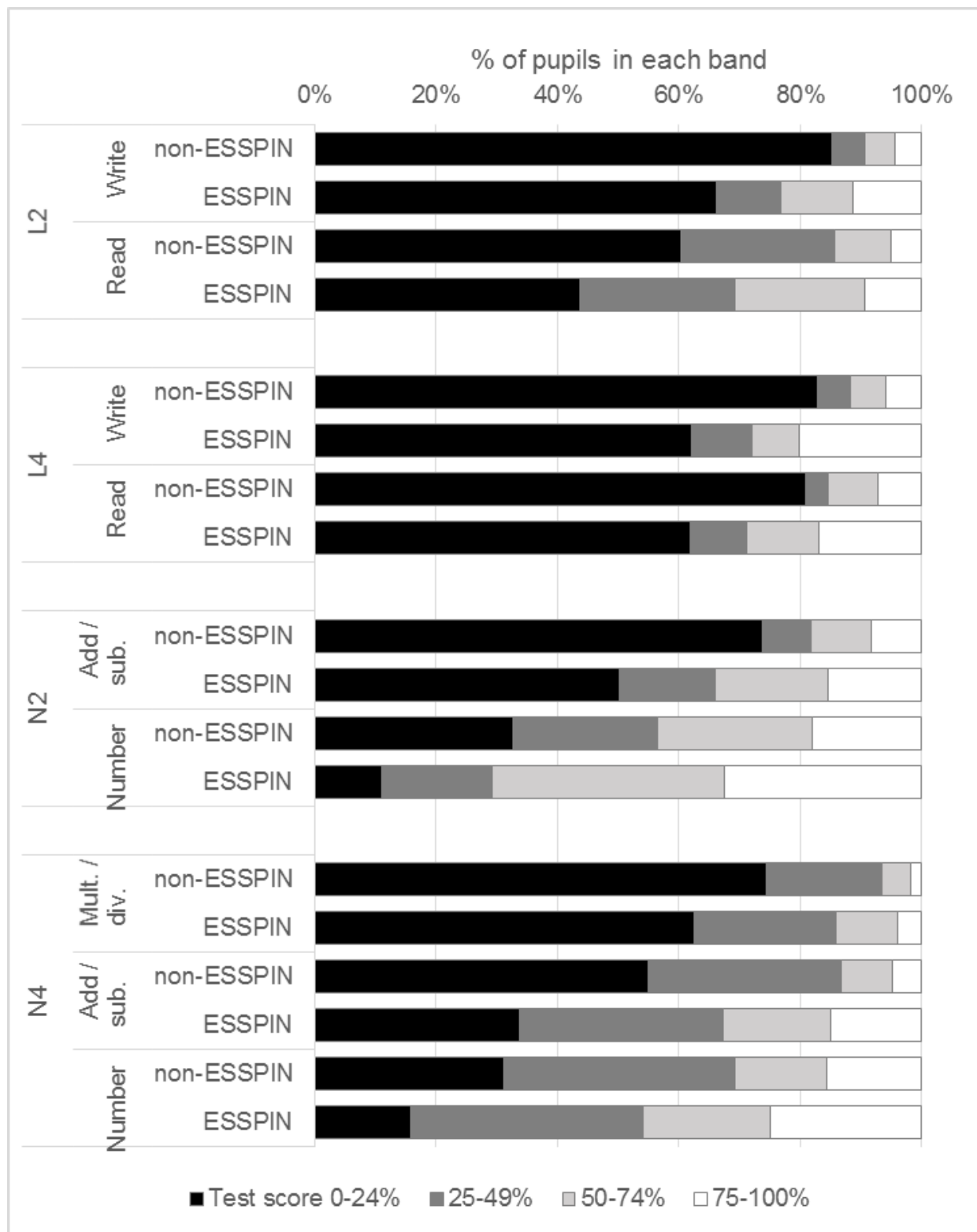
Table 55. Detailed scores in test sub-scales, ESSPIN vs. non-ESSPIN (disaggregated by state)

		Enugu			Jigawa			Kaduna			Kano		Lagos	Kwara
		Non-ESSPIN	ESSPIN		Non-ESSPIN	ESSPIN		Non-ESSPIN	ESSPIN		Non-ESSPIN	ESSPIN	ESSPIN	ESSPIN
L2	Writing	39.0	61.6	+	1.0	4.6	+	7.9	9.8		9.1	7.2	53.8	24.2
	Reading	48.0	75.5	+	10.5	18.8	+	22.1	25.4		25.0	30.9	57.7	31
	Score in grade 1 items	64.6	81.7	+	19.0	29.0	+	31.7	35.7		32.7	38.2	72.6	49.5
	Score in grade 2 items	57.0	78.9	+	13.3	19.0	+	22.6	22.4		19.4	20.1	66.6	38.6
N2	Number concepts	74.0	76.2		28.7	51.8	+	41.4	47.4		43.0	53.0	73.3	65.9
	Addition and subtraction	49.9	53.6		5.5	16.4	+	16.4	19.5		18.5	20.0	59.2	37.1
	Score in grade 1 items	77.3	80.5		34.7	52.0	+	47.2	51.5		47.5	52.6	78.1	67.3
	Score in grade 2 items	44.1	47.6		10.6	23.5	+	17.2	22.2		20.0	24.2	49.7	35.7
L4	Writing	48.5	61.3	+	4.7	11.0	+	5.6	7.9		12.2	18.6	59.8	26.5
	Reading with comprehension	45.1	59.6	+	3.3	9.1	+	7.3	10.6		14.2	20.9	50.3	28.3
	Score in grade 1/2 items	78.4	86.8	+	27.4	37.7	+	34.2	35.9		37.3	43.1	85.1	64.1
	Score in grade 3 items	49.5	64.1	+	8.7	15.7	+	15.6	16.8		21.4	27.1	58.2	34.9
	Score in grade 4 items	38.8	48.1		3.8	9.7	+	5.2	7.3		10.9	16.4	42.4	21.5
N4	Number concepts	65.0	70.6		27.4	46.2	+	35.6	39.9		39.9	45.0	58.6	49.6
	Addition and subtraction	47.5	55.9	+	9.5	20.2	+	18.9	27.3		24.0	27.0	58.6	44
	Multiplication & division	27.3	32.4		4.4	11.2	+	8.5	16.1		17.1	16.4	33.7	19.6
	Score in grade 1/2 items	70.7	76.7	+	33.6	50.2	+	48.9	55.3		51.0	55.3	75.6	65.3
	Score in grade 3 items	38.2	46.6	+	11.7	21.0	+	19.6	24.4		25.5	29.2	42.2	28.8
	Score in grade 4 items	25.9	29.4		2.8	10.0	+	6.1	12.2	+	12.4	13.8	27.5	15.5

Note. See Table 48 above.

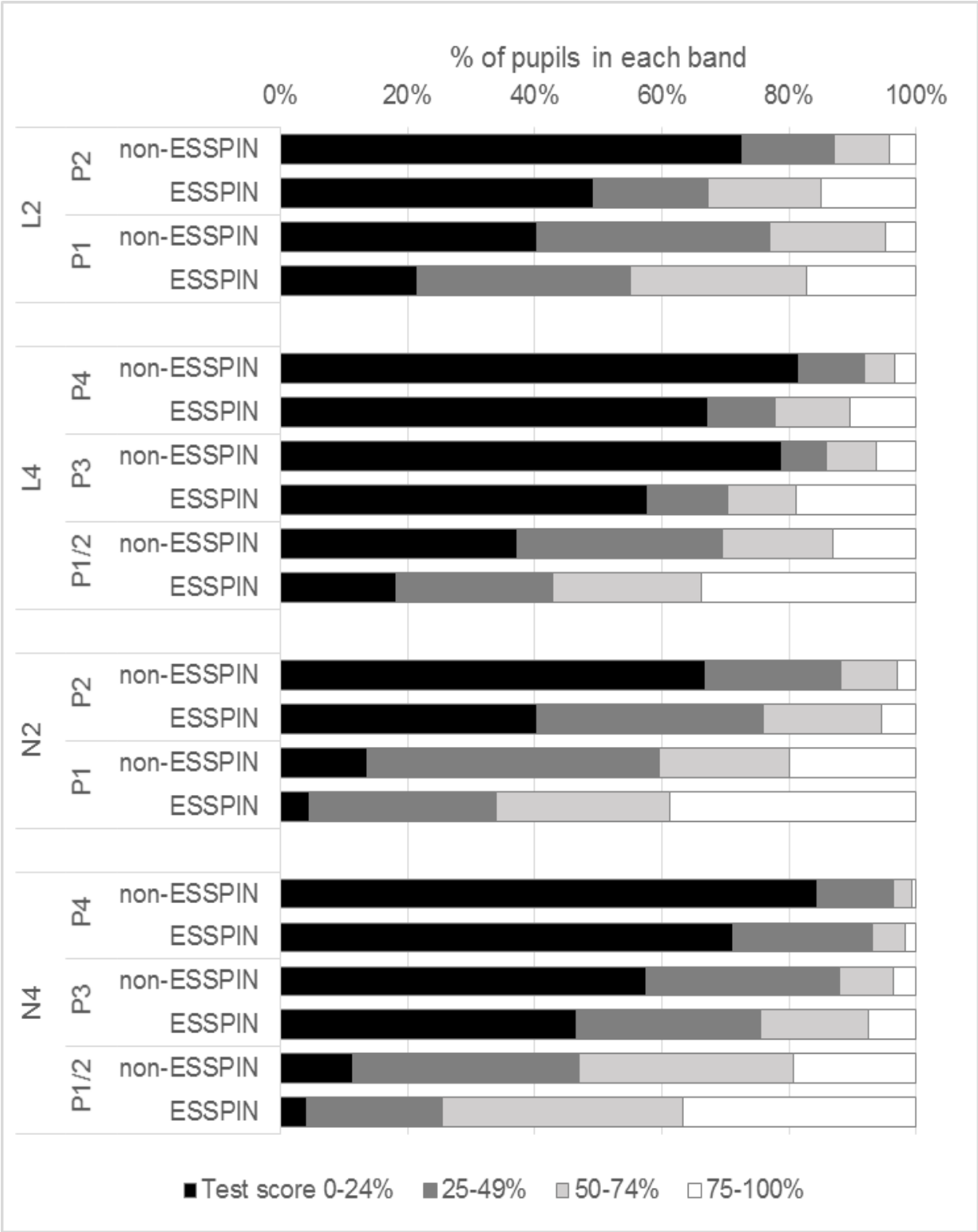
E.2 Distribution of pupils by test score quartiles, by learning domain or grade level

Figure 20. Distribution of pupils by test score quartiles, learning domain and ESSPIN status



Note. See Figure 17 above.

Figure 21. Distribution of pupils by test score quartiles, grade level of questions, and ESSPIN status



Note. See Figure 18 above.

E.3 Proportion of pupils scoring 0%–24%, by state and learning domain

Figure 22. Proportion of pupils scoring 0%–24% in grade 2 literacy, by state and learning domain

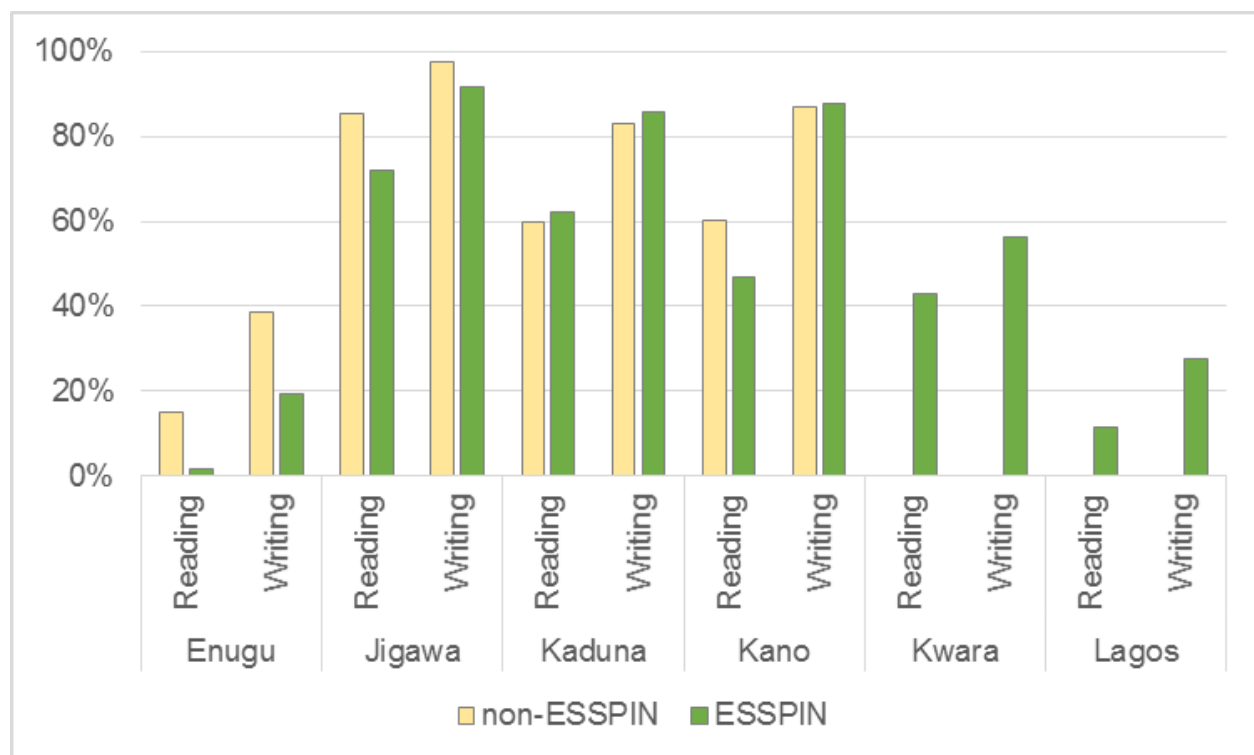


Figure 23. Proportion of pupils scoring 0%–24% in grade 4 literacy, by state and learning domain

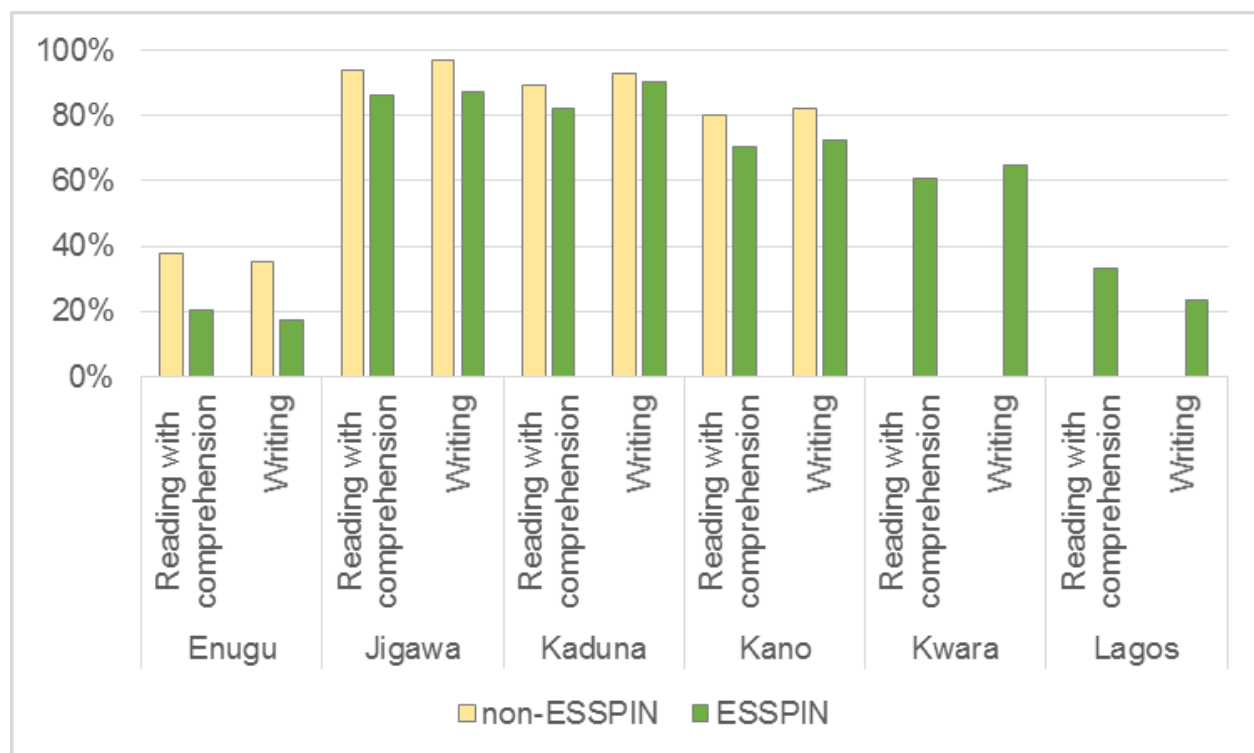
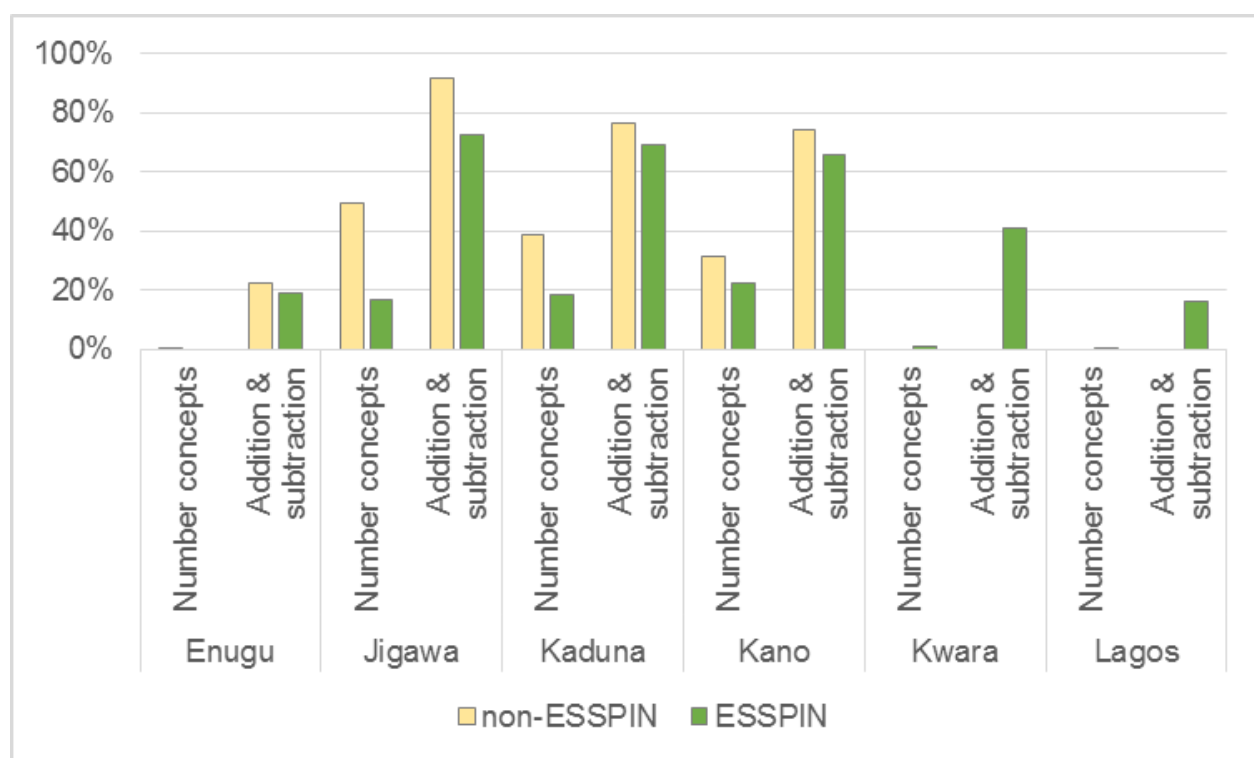
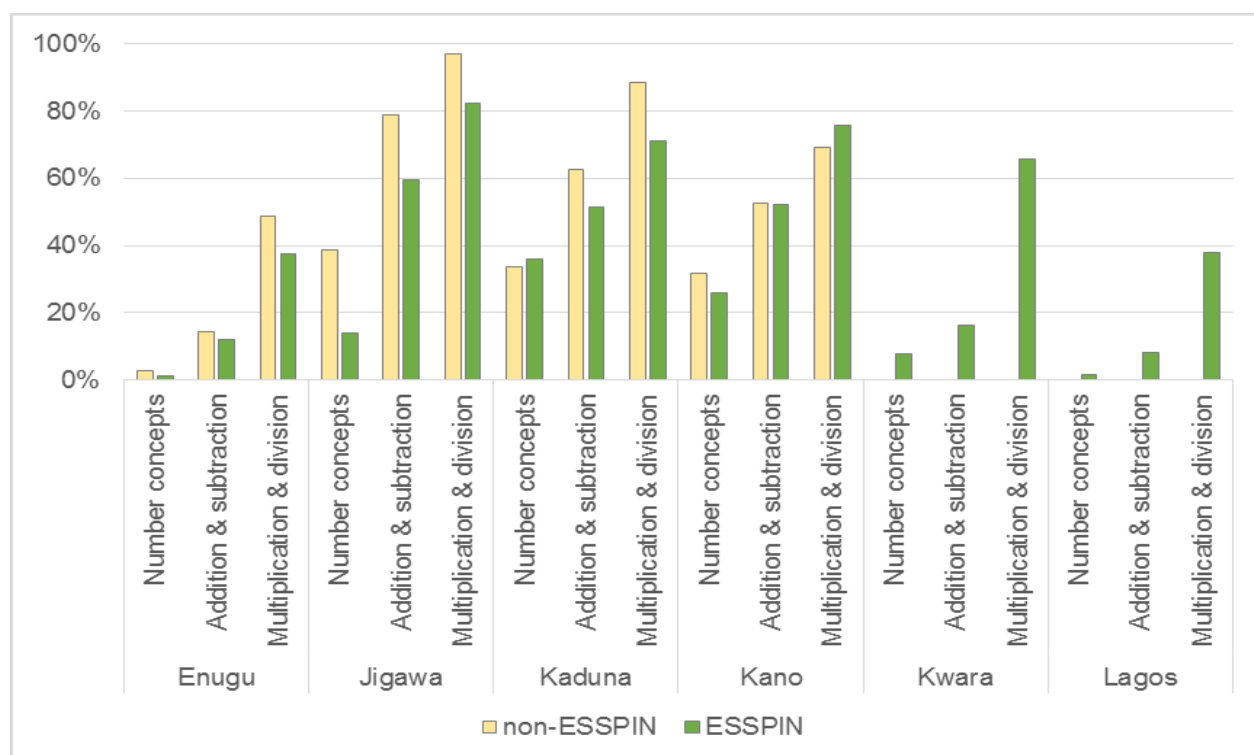


Figure 24. Proportion of pupils scoring 0%–24% in grade 2 numeracy, by state and learning domain**Figure 25. Proportion of pupils scoring 0%–24% in grade 4 numeracy, by state and learning domain**

E.4 Proportion of pupils scoring 75%–100%, by state and learning domain

Figure 26. Proportion of pupils scoring 75%–100% in grade 2 literacy, by state and learning domain

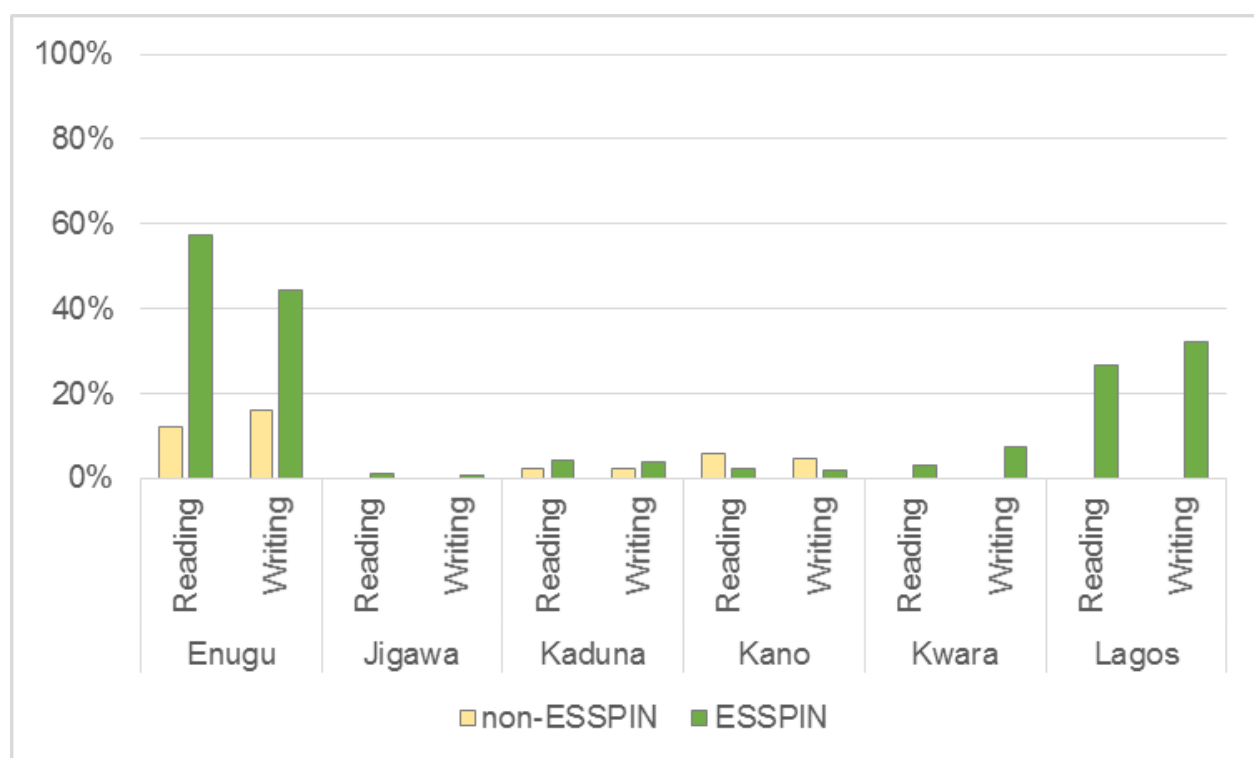


Figure 27. Proportion of pupils scoring 75%–100% in grade 4 literacy, by state and learning domain

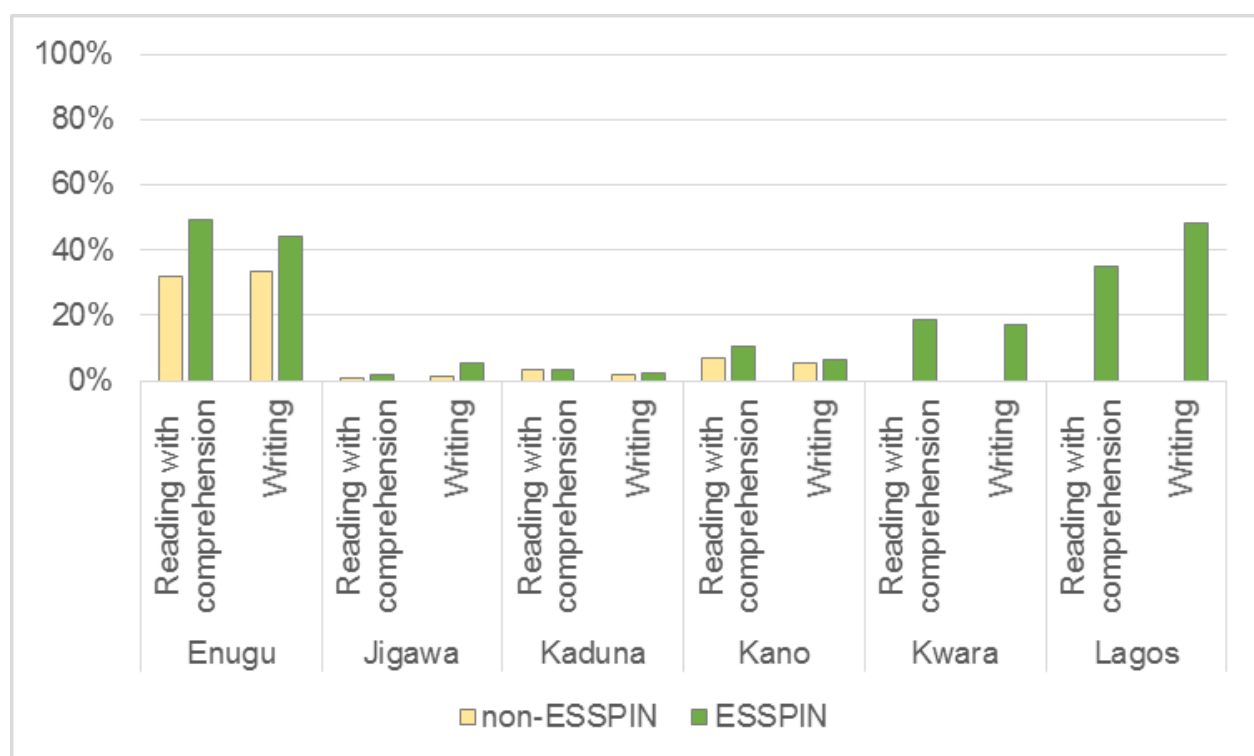
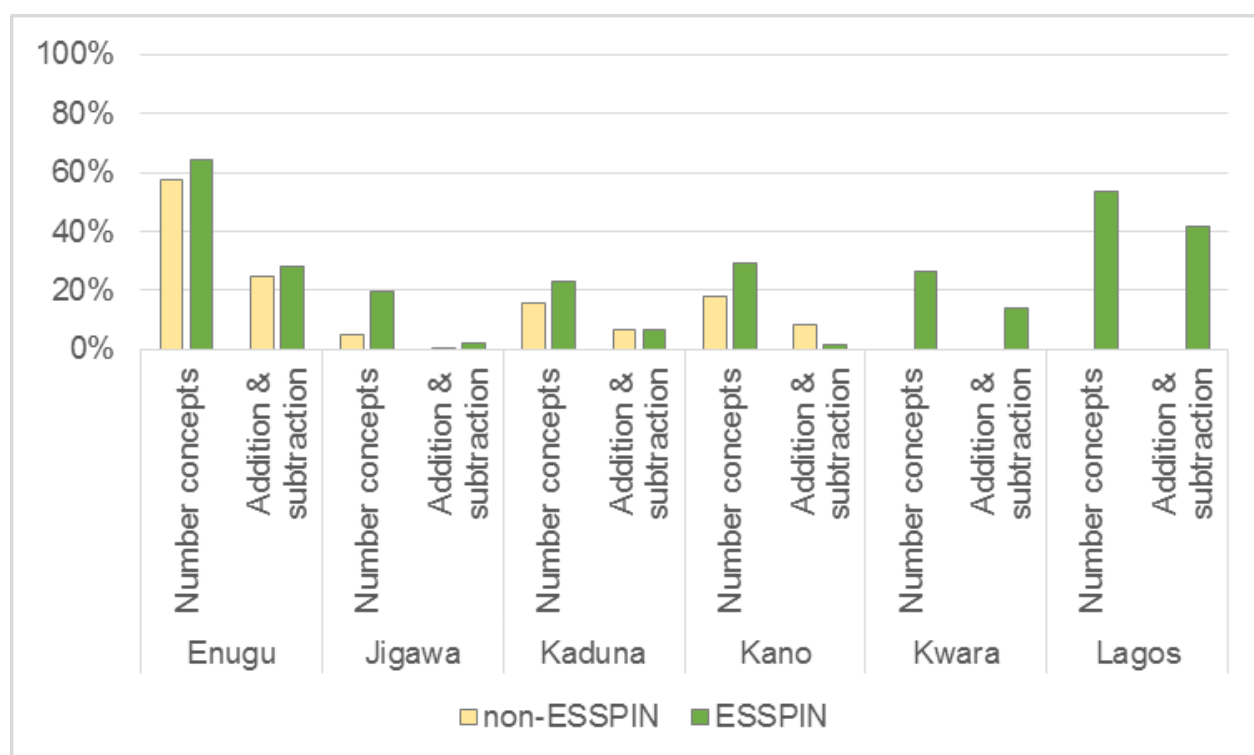
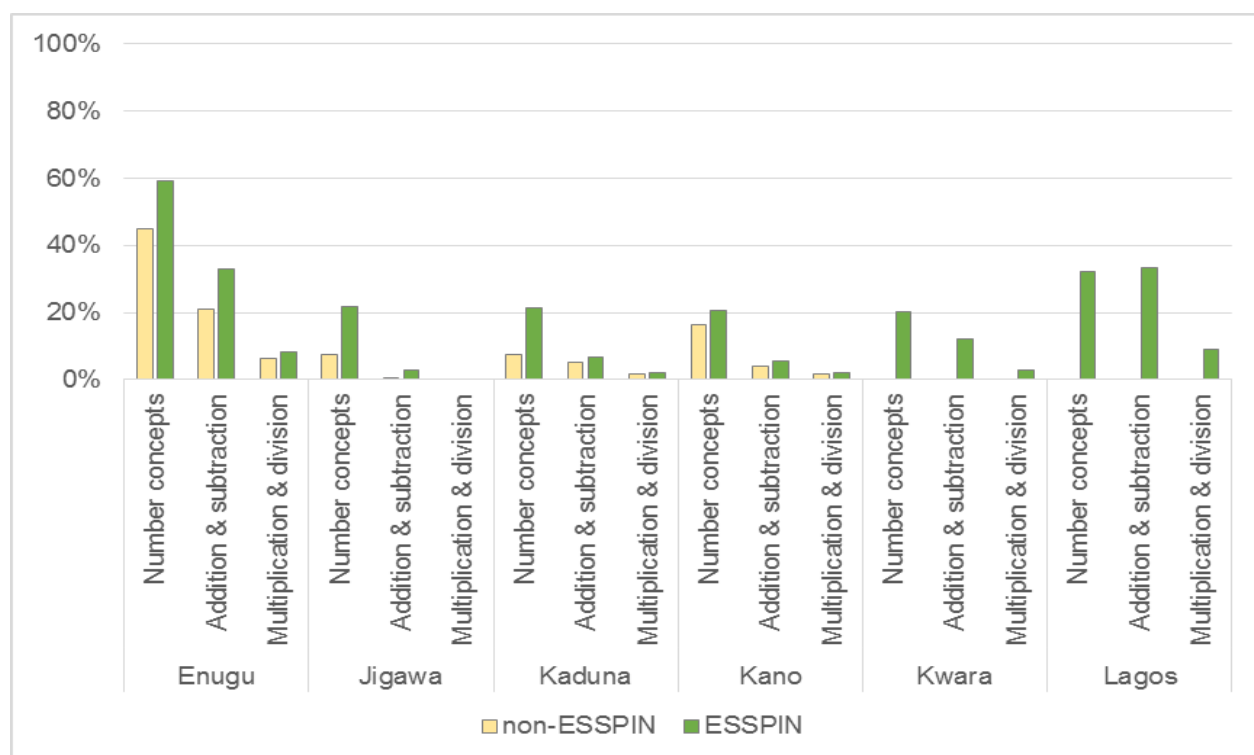


Figure 28. Proportion of pupils scoring 75%–100% in grade 2 numeracy, by state and learning domain**Figure 29. Proportion of pupils scoring 75%–100% in grade 4 numeracy, by state and learning domain**

E.5 Proportion of pupils in each band, by ESSPIN status and state

Figure 30. Grade 2 literacy

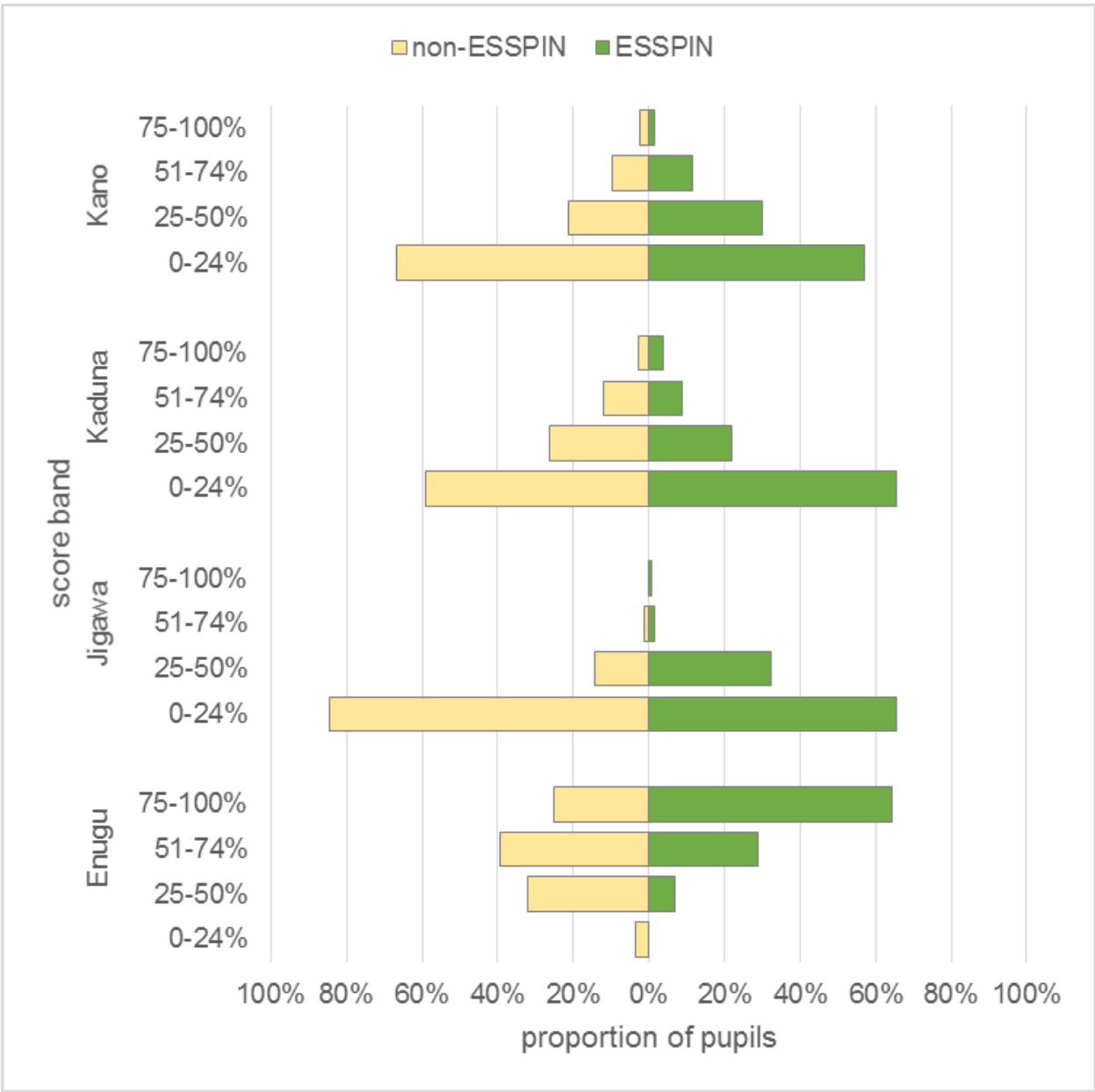


Figure 31. Grade 4 literacy

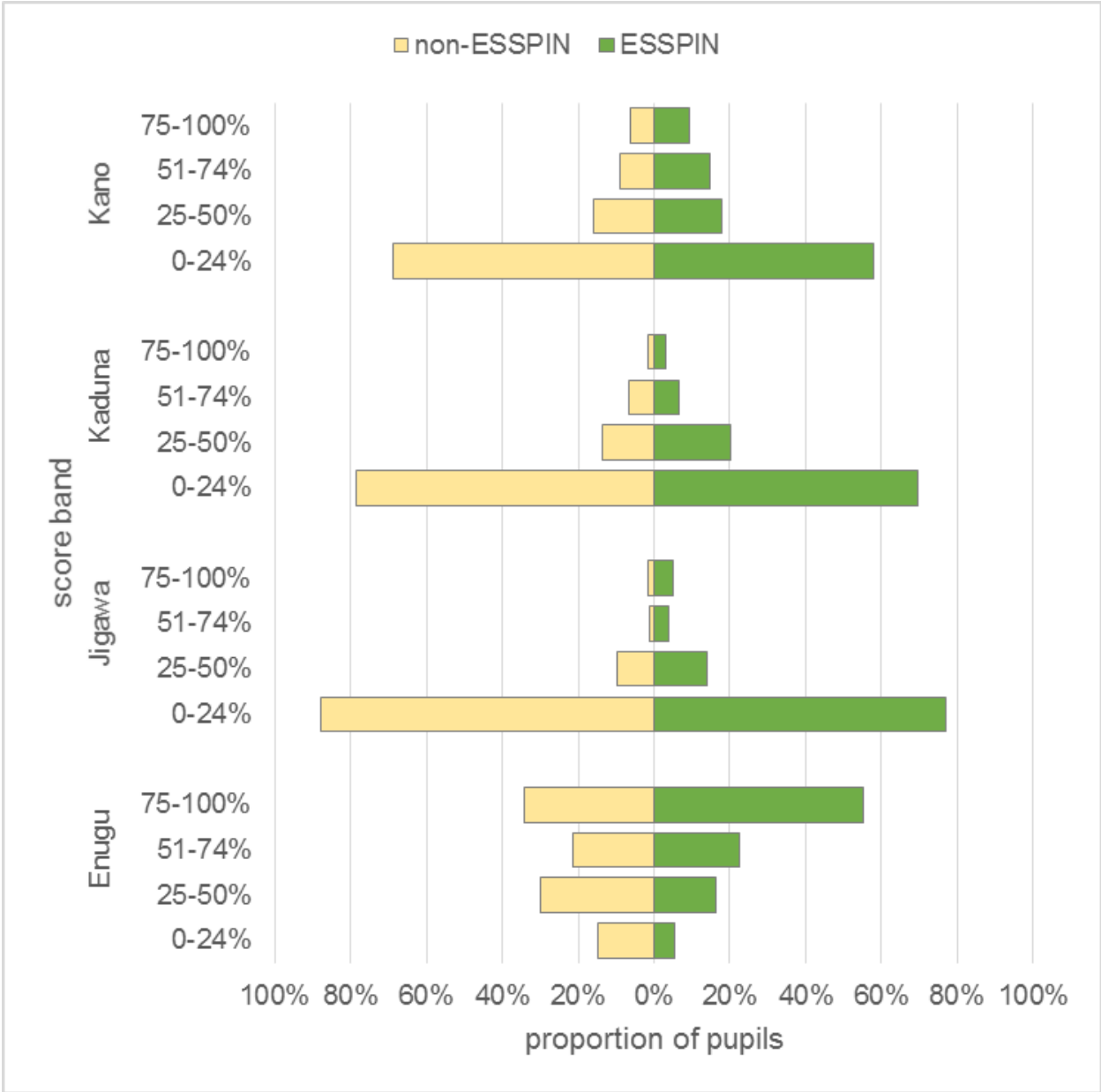


Figure 32. Grade 2 numeracy

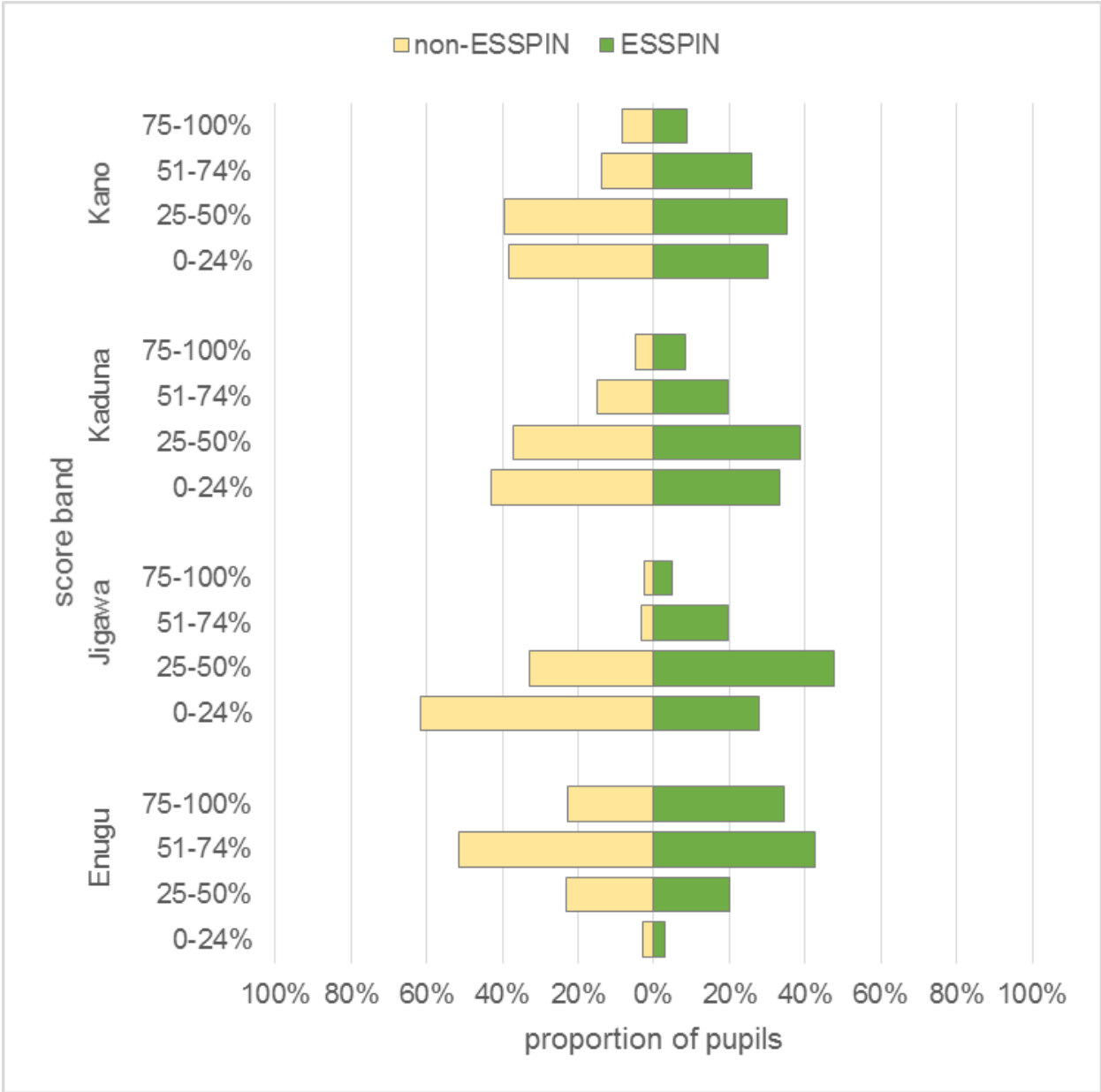
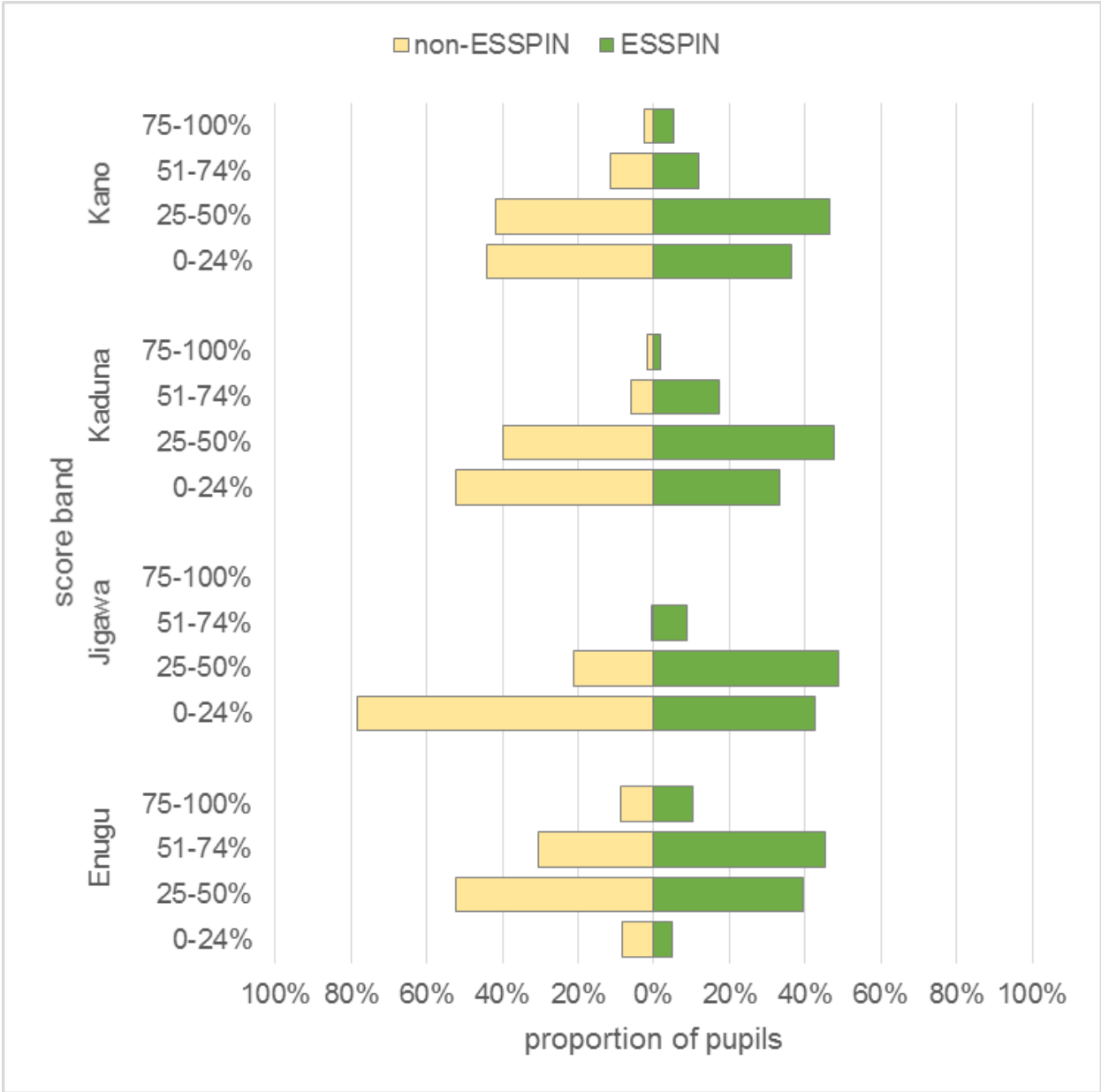


Figure 33. Grade 4 numeracy



E.6 Proportion of grade 4 pupils in each band on grade 1/2 level items only, by ESSPIN status and state

Figure 34. Proportion of grade 4 pupils in each band for grade 1/2 level literacy items, by ESSPIN status and state

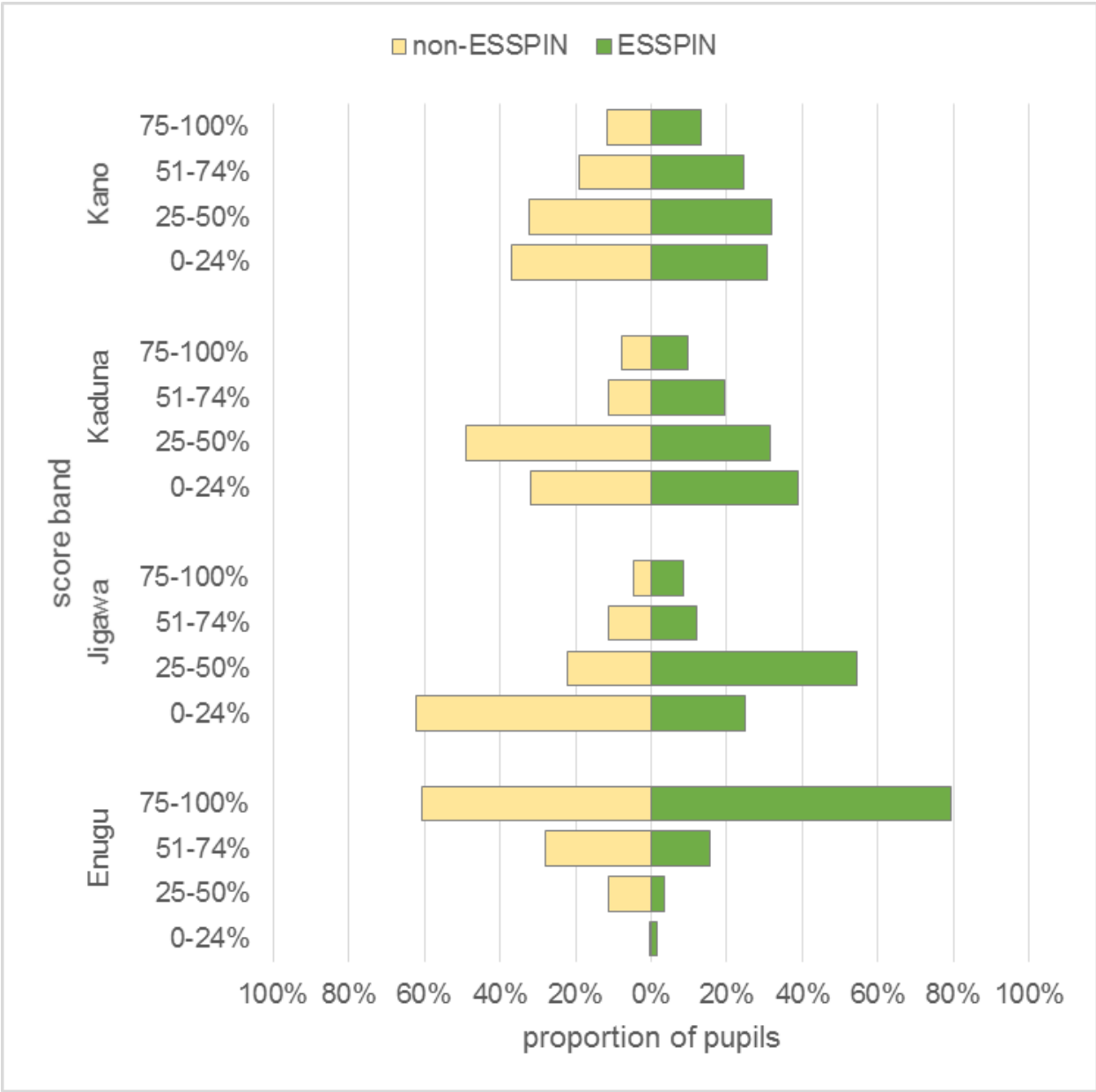
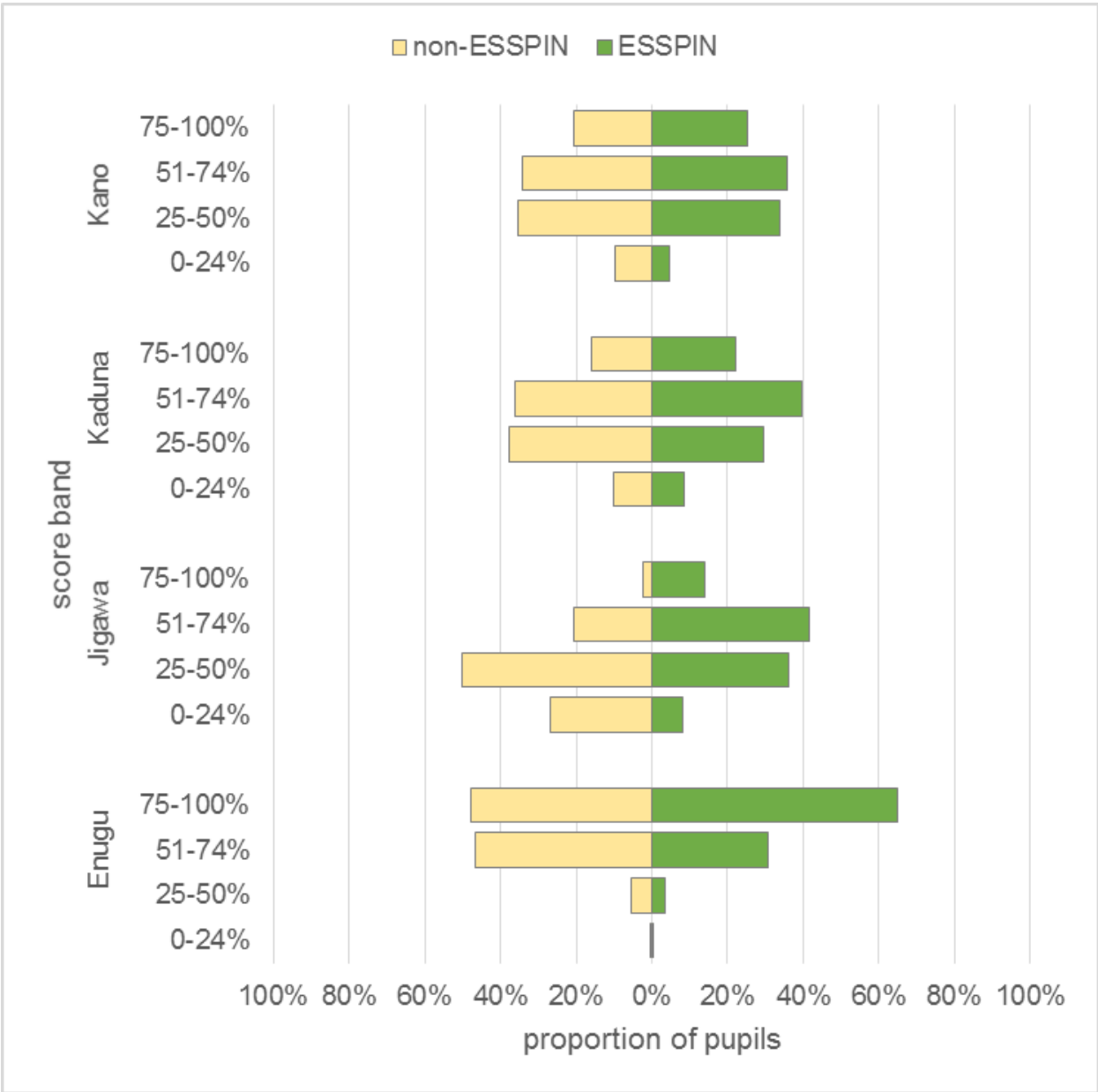
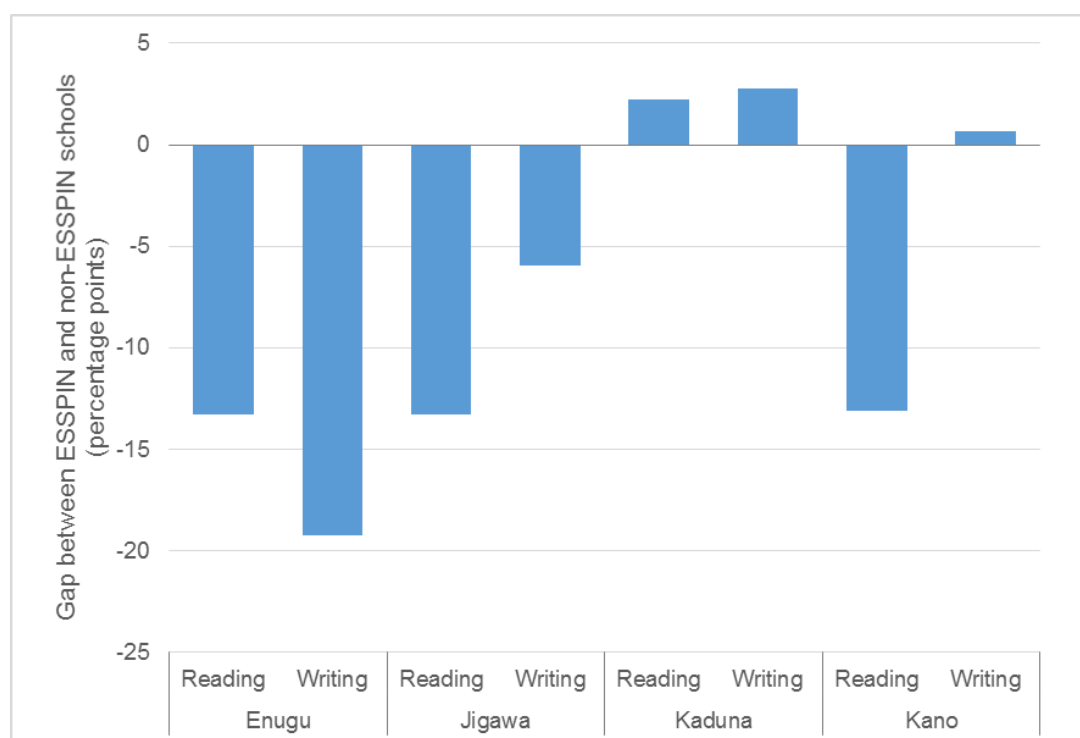


Figure 35. Proportion of grade 4 pupils in each band for grade 1/2 level numeracy items, by ESSPIN status and state



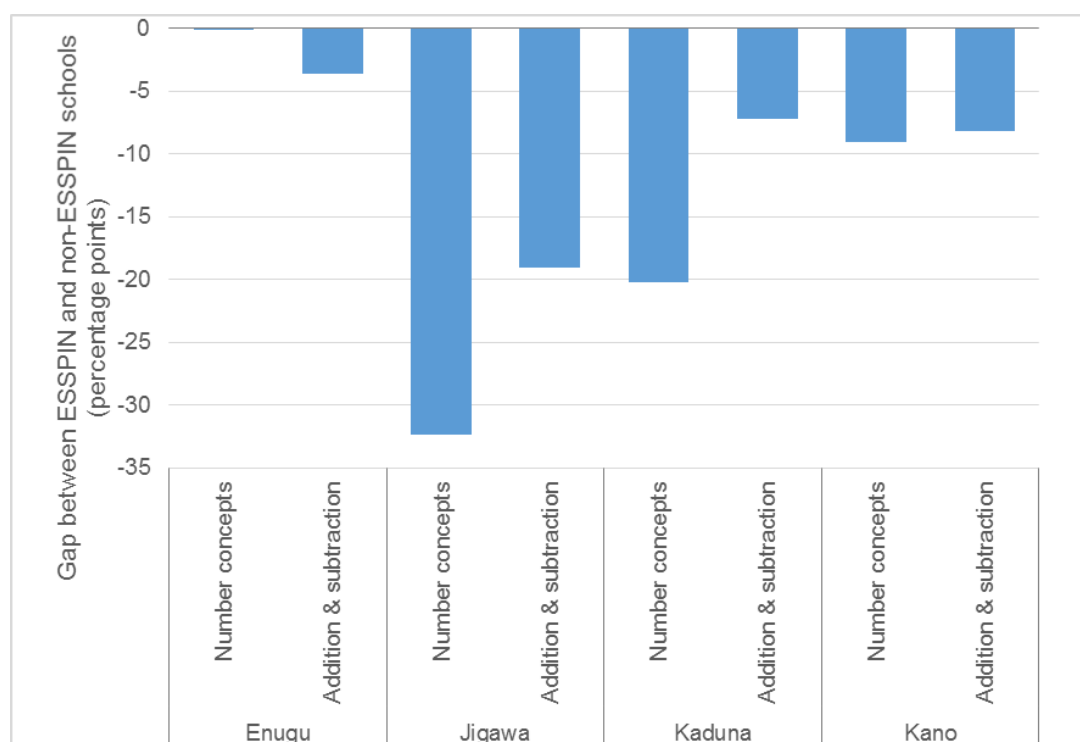
E.7 Gap between ESSPIN and non-ESSPIN schools in the proportion of children scoring under 25%

Figure 36. Gap between ESSPIN and non-ESSPIN schools in the proportion of children scoring under 25%, by state and learning domain (grade 2 literacy)



Note. A negative number indicates a gap in favour of ESSPIN schools.

Figure 37. Gap between ESSPIN and non-ESSPIN schools in the proportion of children scoring under 25%, by state and learning domain (grade 2 numeracy)



Note. A negative number indicates a gap in favour of ESSPIN schools.