Oxford Policy Management

CAMBRIDGE EDUCATION

esspin
Education Sector Support Programme in Nigeria

UKaid
from the British people

ESSPIN Composite Survey 3

Overall report

Stuart Cameron, Katharina Keck, Alia Aghajanian and Zara Majeed

September 2016

Acknowledgements

Ifeatu Nnodu and Shefali Rai managed the Composite Survey 3 (CS3) project. Femi Adegoke of Oxford Policy Management (OPM) Abuja directed the survey team in Abuja. Ekundayo Arogundade and Kelechi Udoh, and their team of state coordinators and monitoring officers, managed the fieldwork. Babatunde Akano and Joshua Moriyonu led data management. Rachel Outhred and Tanya Lone worked on learner assessment analysis and redesign. Caine Rolleston provided helpful comments on an earlier draft of this report. David Megill provided sampling support and survey weights. Reg Allen analysed learner assessments, teacher assessments, and the teacher motivation survey. Rachita Daga and Sourovi De contributed analysis of School-Based Management Committee (SBMC) and inclusion indicators.

From the Education Sector Support Programme in Nigeria (ESSPIN), we acknowledge the contribution of Ron Tuck, who first devised the concept of the Composite Survey; we also acknowledge the various leadership roles undertaken in relation to Composite Surveys 1 and 2 (CS1 and CS2) by Stuart Cameron, Kayode Sanni, John Kay, Lilian Breakell, Simon Thomson, Fatima Aboki, Sandra Graham and Jake Ross. For CS3, Abimbola Osuntusa, Fatima Aboki, Helen Pinnock, Kayode Sanni, Laura McInerney, Manjola Kola, Sandra Graham, Simon Thomson and Jake Ross provided valuable information and support, including comments on an earlier draft of this report. Allan Findlay and Agbenu Onazi provided information on enrolment and school numbers from the Annual School Censuses.

The Composite Surveys would not have been conducted without UK aid funding provided by the UK Department for International Development (DFID), including the guidance and support of Barbara Payne, Emily Oldmeadow and Esohe Eigbike.

We are very grateful to the School Support Officers (SSOs), Social Mobilisation Officers (SMOs), and other local government and state staff who acted as data collectors for the Composite Survey. We are also very grateful to the state government officials who shared information on the context in which ESSPIN has been implemented. Finally, we are extremely grateful to all of the head teachers, teachers and students who kindly took the time to participate in this survey.

The Composite Survey was carried out by Oxford Policy Management (OPM) for the Education Sector Support Programme in Nigeria (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 Shefali Rai. For further information contact Shefali.rai@opml.co.uk or stuart.cameron@opml.co.uk.

Oxford Policy Management Limited
Registered in England: 3122495

Level 3, Clarendon House
52 Cornmarket St
Oxford OX1 3HJ
United Kingdom

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

Executive summary

The Education Sector Support Programme in Nigeria (ESSPIN) (2008–17) seeks to improve learning outcomes for children of basic education age in six Nigerian states: Enugu, Jigawa, Kaduna, Kano, Kwara and Lagos. 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 public schools in the six ESSPIN-supported states. ESSPIN is funded by the UK Department for International Development (DFID) and managed by a consortium led by Cambridge Education. The Composite Surveys have been carried out for ESSPIN by Oxford Policy Management (OPM).

This report presents findings from the first, second and third rounds of the ESSPIN Composite Survey (CS1, CS2 and CS3). These took place in 2012, 2014 and 2016, respectively. The surveys covered a wide range of indicators at the teacher, head teacher, School-Based Management Committee (SBMC), and learner levels. The aim is to understand change in schools over time, and whether schools which receive intervention through ESSPIN are working better than those which do not. The main findings are as follows:

Head teacher effectiveness has not changed significantly over time, for the six states as a whole, although aspects such as the use of lesson observation and professional development meetings have improved. Around 18% of head teachers meet an overall standard for head teacher effectiveness. Head teachers in schools that have had more years of ESSPIN intervention are much more effective than those in schools that have had few years of ESSPIN intervention.

School development planning has improved dramatically since 2014. Around one in five schools across the six states meet the standard for effective school development planning. Schools that have had more years of ESSPIN intervention are much better at school development planning than those that have received fewer years of intervention.

Trends in **school inclusiveness** – measured by aspects such as whether the head teacher has taken action on learners' attendance, and whether teachers engage boys and girls equally – depend on the exact measure used. Fewer schools meet the overall inclusiveness standard than in 2012, but a more nuanced continuous measure of how close they are to meeting it has improved. Schools that have received more years of ESSPIN intervention appear to be slightly more inclusive than those that have received fewer years of intervention, although the difference in the proportion of schools that meet the overall standard is not statistically significant.

SBMCs have become much more functional since 2012 or 2014, and are also better at facilitating the participation of women and children. They are more likely to have conducted awareness-raising about the value of education, and to have raised issues of children's exclusion with the Local Government Educational Authority (LGEA) or state government, than in 2012. ESSPIN's intervention is associated with much better functioning and inclusive SBMCs.

Teachers trained through ESSPIN perform better on English and numeracy content knowledge tests than non-ESSPIN-trained teachers, and are more competent, making more use of better teaching techniques during lesson observations. In particular, they make more use of teaching aids, and assign more individual and group tasks. Since 2014, teacher competence has improved, but teachers' scores on the content knowledge tests have worsened significantly.

Overall **school quality** has improved since 2012, according to our composite measure based on head teacher effectiveness, school development planning, SBMC functionality and teacher competence. Each year of ESSPIN intervention is associated with an increase of around 12 percentage points in the proportion of schools that meet the quality standard. When we control for

quality at baseline and differences between states, the effect of ESSPIN remains, and is estimated at around 5–6 percentage points per year of intervention.

Children's **learning outcomes** have improved for Grade 4 numeracy, but worsened in Grade 2 English literacy and numeracy, with no significant change in Grade 4 English literacy. ESSPIN's intervention is associated with higher scores, even controlling for the state that the school is located in, learners' socioeconomic status, and pre-existing school facilities.

Table 1, Table 2, Table 3, and Table 4 below summarise the key findings.

Table 1: Change over time: key indicators in 2012, 2014, 2016

	See page	2012 (CS1)	2014 (CS2)	2016 (CS3)	Change 2012–16	Change 2014–16
Effective head teacher (%)	21	13.6	14.2	17.8	+4.1	+3.6
School development planning (%)	26	3.8	7.4	18.6	+14.8*	+11.3*
Inclusive (%)	29	18.8	10.5	11.4	-7.4*	+0.9
Functioning SBMC (%)	35	21.7	30.9	56.1	+34.4*	+25.1*
Competent teachers (%)	48	69.7	57.4	66.8	-2.9	+9.4*
Competent teachers (new measure, %)	48		21.0	20.5	n/a	-0.5
Good quality school (%)	63	3.9	8.3	20.2	+16.2*	+11.8*
Good quality school (new measure, %)	63		4.6	5.6	n/a	+1.0
Grade 2 literacy score	70	461	466	447	-13.5	-18.5*
Grade 4 literacy score	70	458	449	448	-10.7	-1.6
Grade 2 numeracy score	70	499	457	445	-53.9*	-12.4*
Grade 4 numeracy score	70	463	449	461	-2.2	+12.2*

Note. * indicates statistical significance ($p < .05$). The new, stricter measure of teacher competence and school quality take into account teachers' test results.

Table 2: Key indicators in 2016, by ESSPIN Output Stream 3 Intervention

	See page	Min. (1 year)	Med. (2–3 years)	Max. (4–5 years)	Estimated effect of one year of full intervention
Effective head teacher (%)	23	14.1	26.0	24.2	+5.2*
School development planning (%)	27	11.1	36.3	28.3	+8.7*
Inclusive (%)	31	7.3	17.4	23.4	+2.3
Good quality school (%)	63	12.1	35.6	36.7	+9.4*
Good quality school (stricter measure, %)	63	1.1	13.2	17.3	+3.6*
Grade 2 literacy score	72	431.6	470.9	496	+9.2*
Grade 4 literacy score	72	428.5	473.4	494.1	+11.5*
Grade 2 numeracy score	72	433.5	459.2	493	+4.4
Grade 4 numeracy score	72	442.1	485.2	512.2	+9.2

Note. * indicates statistical significance ($p < .05$). The stricter measure of school quality takes into account teachers' test results. ESSPIN's Output Stream 3 intervention focuses on head teacher and teacher training and support.

Table 3: Key indicators on SBMCs and inclusion, by ESSPIN Output Stream 4 intervention

	See page	No intervention	Pre-CS1	Post-CS1	Estimated effect of one year of intervention
School meets standard for functioning SBMC (%)	39	49.4	61	70.7	+5.2
SBMCs' work on inclusion	39				
Conducted awareness-raising (%)		67.1	81.3	75.9	4.1
Addressed exclusion (%)		66.3	82.2	67.2	-0.1
Took action for commonly excluded groups (%)		18.6	25.1	17.4	+3.3
Raised issues of children's exclusion (%)		13.9	21.6	13.3	+2.9

Note. * indicates statistical significance ($p < .05$). ESSPIN's Output Stream 4 intervention focuses on improving community participation in school improvement.

Table 4: Teacher competence in 2016, by ESSPIN training

	See page	Non-ESSPIN-trained	ESSPIN-trained	Difference
Competent teachers (%)	50	68.6	65.1	-3.5
Competent teachers (stricter measure, %)	50	18.4	22.6	+4.3*
Teachers' English scale	56	446.0	467.3	+21.2*
Teachers' mathematics scale	56	448	475	+26.3*

Note. * indicates statistical significance ($p < .05$). The stricter measure of teacher competence takes into account teachers' test results.

Table of contents

Annex C (with full details of the regression analysis) is provided in a separate document.

List of tables, figures and boxes

List of abbreviations

ACLED	Armed Conflict Location and Event Data Project
BEd	Bachelor of Education
CAPI	Computer-assisted personal interviews
CBO	Community-based organisation
CS1	Composite Survey 1
CS2	Composite Survey 2
CS3	Composite Survey 3
DFID	Department for International Development (UK)
EMIS	Education Management Information System
ESSPIN	Education Sector Support Programme in Nigeria
HND	Higher National Diploma
IRT	Item response theory
L2	Grade 2 literacy test
L4	Grade 4 literacy test
LGA	Local Government Authority
LGEA	Local Government Educational Authority
N2	Grade 2 numeracy test
N4	Grade 4 numeracy test
NCE	National Certificate of Education
NGN	Nigerian Naira
OPM	Oxford Policy Management
P2	Primary Grade 2
P4	Primary Grade 4
PGDE	Post-Graduate Diploma in Education
PTR	Pupil–teacher ratio
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
UBEC	Universal Basic Education Commission

1 Introduction

The Education Sector Support Programme in Nigeria (ESSPIN, 2008-17) seeks to improve learning outcomes for children of basic education age in six Nigerian states – Enugu, Jigawa, Kaduna, Kano, Kwara, and Lagos. 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. ESSPIN is funded by the UK Department for International Development and managed by a consortium led by Cambridge Education. The Composite Survey has been carried out for ESSPIN by Oxford Policy Management.

The first two rounds of the Composite Survey were carried out in 2012 and 2014. The surveys address five output indicators: teacher competence, head teacher effectiveness, school development planning, SBMC functionality, and inclusive practices in schools. They also address outcome indicators on school quality and pupil learning benchmarks, and an impact indicator, pupil learning achievement. The third round of the Composite Survey (CS3) collects comparable data on these indicators in order to provide information on the extent to which key school-level indicators in the six states have improved during the course of the programme.

1.1 ESSPIN's SIP

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. These activities were designed by the ESSPIN technical assistance team; were piloted and developed in partnership with state and local governments, schools and civil society organisations; and were taken to scale by state governments with steadily decreasing levels of technical assistance support from ESSPIN over the programme lifetime. ESSPIN 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; and (iv) improving inclusion policies and practices in basic education (ESSPIN, 2013b). More recently, the fourth output stream has been altered to focus on increasing community participation in school improvement.

Under the third of these output streams, 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 (Sanni, 2015). 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. 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.

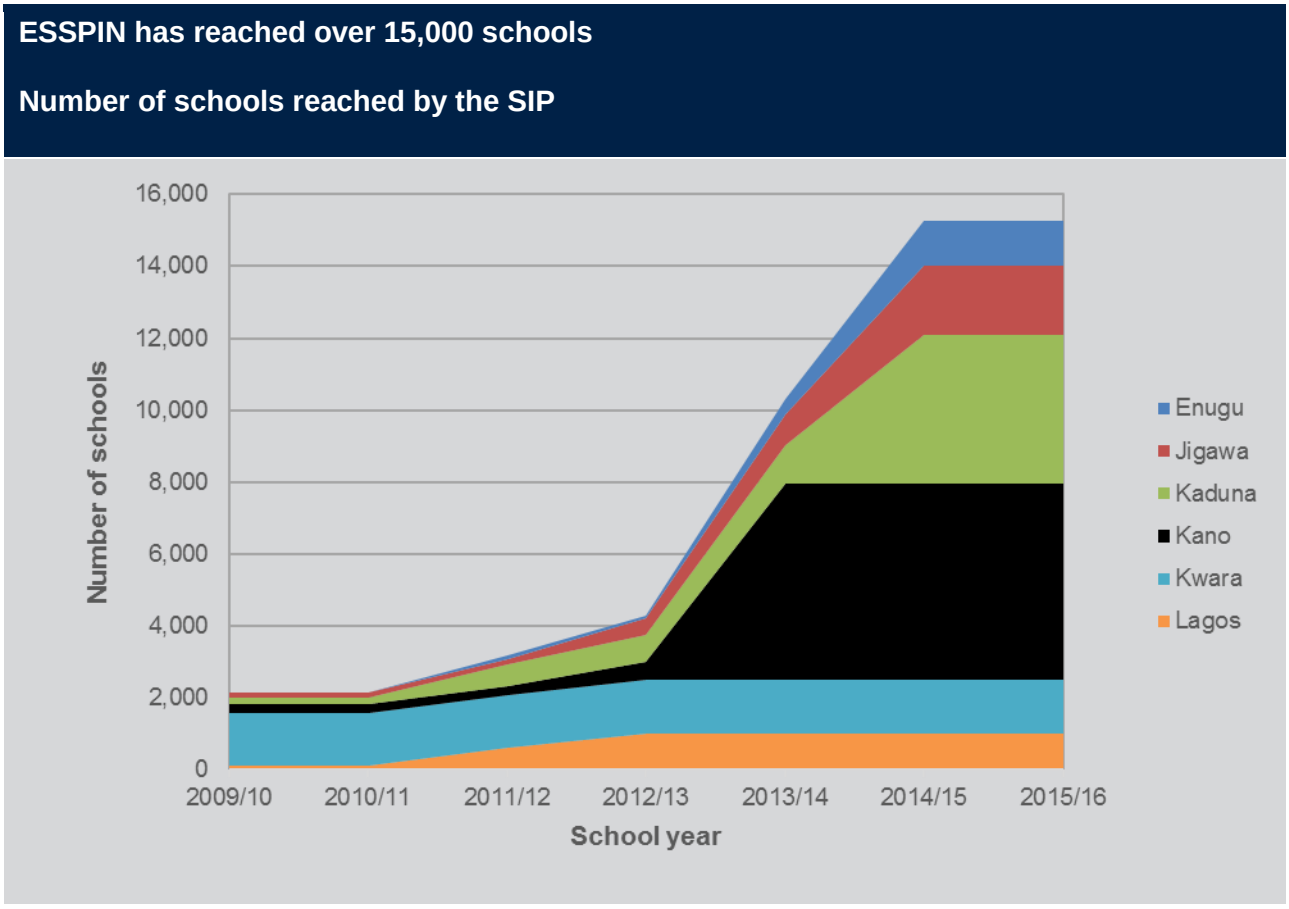
In addition to these interventions, since 2010/11 schools have received support under ESSPIN's fourth output stream (Output Stream 4), originally called improving inclusion policies and practices in basic education, but more recently focused on improving community engagement and learner participation in school improvement. ESSPIN has trained civil society members and government officers from the Department of Social Mobilisation, Social Mobilisation Officers (SMOs), to enable them to train and mentor SBMCs. SBMC members, in turn, have been trained on the roles and responsibilities of SBMCs, school planning and management, communication and leadership, change and relationships management, the participation of women and children in school improvement and education decision-making, resource mobilisation and financial processes, and child protection and participation. This has been complemented by follow-up mentoring visits by SMOs.

The extent to which schools have received each of the intervention components has varied from one year to the next and from state to state (see Annex A and Annex B). In terms of Output Stream 4 intervention, schools in Jigawa and Kaduna first began receiving the intervention in 2010/11, schools in Enugu and Kwara in 2011/12, and schools in Lagos and Kano in 2012/13. In most states, the majority of schools have not yet received a year of full Output Stream 4 intervention, Lagos being the only exception. Moreover, most Output Stream 4 intervention has been provided over the past two years.

In the past few years, ESSPIN's SIP has gone through different rounds of expansion across the six states.

- In Kwara, ESSPIN's Output Stream 3 intervention has been delivered to all schools since the beginning in 2009/2010 (although the intensity has dropped significantly in recent years).
- In Lagos, ESSPIN expanded to all schools during 2012/2013.
- In Kano, ESSPIN expanded to all schools during 2013/2014, although no schools received a full set of interventions in 2014/2015 or 2015/2016 (this is deliberate, as intervention has focused particularly on teacher training, and no leadership training has been delivered as a result).
- In Enugu, Jigawa and Kaduna, ESSPIN expanded to all schools during 2014/2015, although no schools in Kaduna received any intervention during 2015/2016.

Figure 1: Number of schools reached by the SIP



Note. The numbers shown are schools that have received at least one year of full intervention between 2009 and 2015.

Table 5 shows the proportion of schools that received a full package of Output Stream 3 intervention for each year that ESSPIN has been in place. 'Full intervention' in this report means that the school has received some leadership training, some teacher training, and some school

visits during the year, though the amount of each activity may vary. Every school in all six states has received at least one year of full Output Stream 3 intervention at some point in the previous years. However, not all schools that received the full intervention in one year have continued to receive it in all of the subsequent years.

Table 5: Proportion of schools receiving full package of ESSPIN Output Stream 3 interventions (%)

	2009/10	2010/11	2011/12	2012/13	2013/14	2014/15	2015/16	Any year
Enugu	0	0	8	8	25	99	99	100
Jigawa	8	8	0	15	37	100	55	100
Kaduna	4	4	14	19	11	74	0	100
Kano	5	5	0	5	100	0	0	100
Kwara	100	100	0	0	100	100	100	100
Lagos	10	10	60	100	40	100	100	100
Total	0	0	8	8	25	99	99	100

Source: Authors' 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 Annual School Census. 'Any year' refers to the proportion of schools that have received a full package of interventions in one or more years.

The expansion of the programme to all schools in the state required a changed model for delivering training, with the training located closer to schools. 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 SSITs and ESSPIN trained School Support Officers (SSOs) – local government-level staff – who, in turn, delivered the school-level interventions. 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.

Stakeholders perceive that having Local Government Authority (LGA) officers deliver the SIP has brought support closer to the schools, and that a school's needs can therefore be addressed more directly. However, the change in model is likely to have affected the quality of implementation to some extent. Programme staff argue that locating training closer to the schools has longer-term benefits, but that in the shorter term the quality standards of the pilot programme might not be fully upheld as the new, much larger numbers of trainers, who typically have lower qualifications than those in the first wave, develop competencies. This trade-off was necessary because it would not have been feasible for the 200 members of the SSITs to support more than 15,000 schools.

Thus, we may expect that schools which were only reached by ESSPIN under this new model would not improve as much as schools that had received training by the SSITs. However, 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, if we assume that 100% of ESSPIN schools, but only 10% of non-ESSPIN schools, meet a particular quality standard, and if in 2012/13 around 16% of schools across the six states received ESSPIN's Output Stream 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 all

schools in the six states would make a dramatic difference to the average school quality in the states as a whole.

This report will be able to give an indication of whether the scale-up during the 2013/14 and 2014/2015 school years has had these expected effects. By the end of the 2014/2015 school year, each school had had at least one full year of intervention, and we can expect that by the time of the CS3 survey (at the end of the 2015/2016 school year), the intervention would have taken effect in ways that can be measured with our indicators. If the change in delivery model has led to a dilution in the 'ESSPIN effect', one additional year of ESSPIN intervention may lead to only small improvements. On the other hand, by this stage (CS3), we can expect state-level averages to have improved compared to CS1 and CS2, since more schools have since experienced a full year of intervention, and since the intervention received during 2013/14 and 2014/15 has had time to take effect.

In terms of Output Stream 4 intervention, we would expect changes over time to be smaller, because many schools have not yet received sustained Output Stream 4 intervention. We would, however, expect Output Stream 4 intervention to be associated with better performance on those indicators – such as the functionality of SBMCs – that relate closely to Output Stream 4 activities, especially since the Output Stream 4 intervention has intensified in recent years in those schools in which it had already been rolled out. We would expect these schools to have improved compared to schools that have not yet received sustained Output Stream 4 intervention.

However, the effects of the programme depend on both its design and its implementation, delivered in partnership with state governments. In particular, it depends on the timely release of sufficient funds by state governments to support the intensity and continuity required for the School Improvement Model to run effectively as designed and to have the desired impacts. There are also a number of contextual factors that are likely to have affected the functioning of schools in the six states and the programme, including increases in the number of learners and possible changes in the backgrounds that learners come from. Such changes could mask gains in learning outcomes that would otherwise have been recorded.

1.2 Context

1.2.1 Changes in enrolment in ESSPIN states

Information from Annual School Censuses indicates large increases in enrolment between 2009 and 2013, with some 800,000 additional students entering the school system in the six states, an increase of 19% (Table 6). The number of schools listed in the censuses increased by around 10% during the same period.

There is some uncertainty about the magnitude of these changes, because some of the new schools in the censuses appear to have existed previously but were not listed in the censuses. Nevertheless, even within the schools found in both the 2009 and 2013 census (Table 7) there were increases in Jigawa, Kaduna and Kano of 16%–19%. Across the six states total enrolment increased by 12% just within the schools found in both censuses.

Between 2013 and 2014, total enrolments increased slightly further in all states except Enugu, increasing in total by 2.5%. Again, however, this may partly reflect changes in coverage of the censuses. Looking only at the schools listed in both rounds of the census, there were large increases in enrolment in Kaduna and Kwara, but declines in all of the other states. Pupil–teacher ratios (PTRs) initially increased between 2009 and 2013. Between 2013 and 2014, the data were mixed, suggesting both some increases and decreases in PTRs.

It is difficult to anticipate how the changes in enrolment and PTRs may have affected teaching and learning outcomes in the past few years. The increase in enrolment between 2009 and 2013 is likely to have put a strain on existing schools, which are not likely to have had commensurate increases in teachers or resources, but schools may have been able to adapt as enrolment stabilised or decreased during 2014. However, in 2014, average PTRs in three of the states were still extremely high: 49 in Kaduna, 55 in Jigawa and 69 in Kano. Given such high PTRs, one might not expect the teaching and learning environment to have improved much even if there have been small decreases in the PTR between 2013 and 2014. Section 6.2.5 examines how changes in enrolments and PTRs may have affected progress in learning outcomes.

Table 8 presents the numbers of male and female learners enrolled in 2014/15. Lagos had slightly more girls than boys enrolled, and Kano had close to equal numbers of boys and girls enrolled in school, but the other four states fell short of gender equity. In Jigawa there were only 76 girls in school for every 100 boys, and in Kaduna there were 87 girls in school for every 100 boys.

Table 6: Numbers of schools and enrolment in the 2009, 2013 and 2014 Annual School Censuses

	2009/10		2013/14		2014/15	
State	Number	Enrolment	Number	Enrolment	Number	Enrolment
Enugu	1,188	237,548	1222	187,495	1223	177,185
Jigawa	1,868	471,820	1997	550,813	2012	555,395
Kaduna	3,947	979,659	4225	1,153,460	4225	1,180,039
Kano	4,768	1,883,472	5732	2,474,922	5833	2,568,532
Kwara	1,448	199,604	1497	198,759	1528	199,868
Lagos	986	387,581	1007	400,277	1010	406,678
Total	14,205	4,159,684	15,680	4,965,726	15,831	5,087,697

Source: Annual School Census reports

Table 7: Enrolment change between 2009–2014 and 2013–2014

State	Enrolment change (%)			
	2009 to 2013	2009 to 2013 (Schools found in both censuses only)	2013 to 2014	2013 to 2014 (Schools found in both censuses only)
Enugu	-21.1	-23.7	-5.5	-8.3
Jigawa	+16.7	+18.5	+0.8	-2.4
Kaduna	+17.7	+16.2	+2.3	+7.0
Kano	+31.4	+18.1	+3.8	-17.9
Kwara	-0.4	-2.2	+0.6	+19.8
Lagos	+3.3	0.4	+1.6	-1.17
Total	+19.4	12.4	+2.5	-6.48

Table 8: Male and female enrolment in 2014/15

	Male	Female	Girls enrolled for every 100 boys
Enugu	90,637	86,548	95
Jigawa	315,289	240,106	76
Kaduna	631,287	548,752	87
Kano	1,296,091	1,272,441	98
Kwara	104,629	95,239	91
Lagos	199,396	207,282	104
Total	2,637,329	2,450,368	93
Source: Annual School Census reports, 2014/15			

1.2.2 Political and financial context

State governments play the major role in education financing decisions in Nigeria. A recent study of public financial management in education in the six states found that mismatch of available resources and budget allocations, and the complete political discretion of the governor in determining spending releases, often results in a significant difference between the approved budget and actual releases, which weakens budget credibility (Steenbergen *et al*., 2016). Between 2012 and 2014 there was a strong upward trend in budget allocations to the SIP, but this was mainly driven by two states, Kano and Kaduna, where the states made up 38% and 31%, respectively, of the overall SIP budget allocation. The four other states each made up only somewhere between 5% and 11% of overall SIP budget allocations. A comparison of ESSPIN's own spending with states' budget allocations shows that on average states leveraged about 3.1–4.1 Nigerian Naira (NGN) in state budget allocations for every NGN 1 spent by ESSPIN. However, in terms of actual spending the state leveraging rate is somewhere between NGN 0.2–0.4 for every NGN 1 spent by ESSPIN.

The public financial management study also finds that all six ESSPIN states are faced with budget credibility issues, although to different degrees. In some cases, the Medium-Term Sector Strategy and the budget are relatively well aligned (e.g. Kano), but spending differs considerably. In other cases, the planning documents are overly ambitious, budget and spending are not well aligned, or budget data are not publicly available.

Basic education spending is determined by two key contextual factors: overall resource availability and political influences. Research for the public financial management study found that the former was particularly important during 2012–2015, given the fall in oil prices and federal revenues, which in turn led to a fall in federal allocations to states. Given their dependence on federal allocations and weak internally generated revenues, many states have struggled to finance salary and development expenditure in recent months. Political influence is also key to education spending, especially given the governor's discretionary power over all budget releases. Greater proportions of federal funding have been disbursed in ESSPIN states than in other states, suggesting that the ESSPIN states have been more successful at drawing down and utilising federal education intervention funds than non-ESSPIN states.[1]Conflict and violence

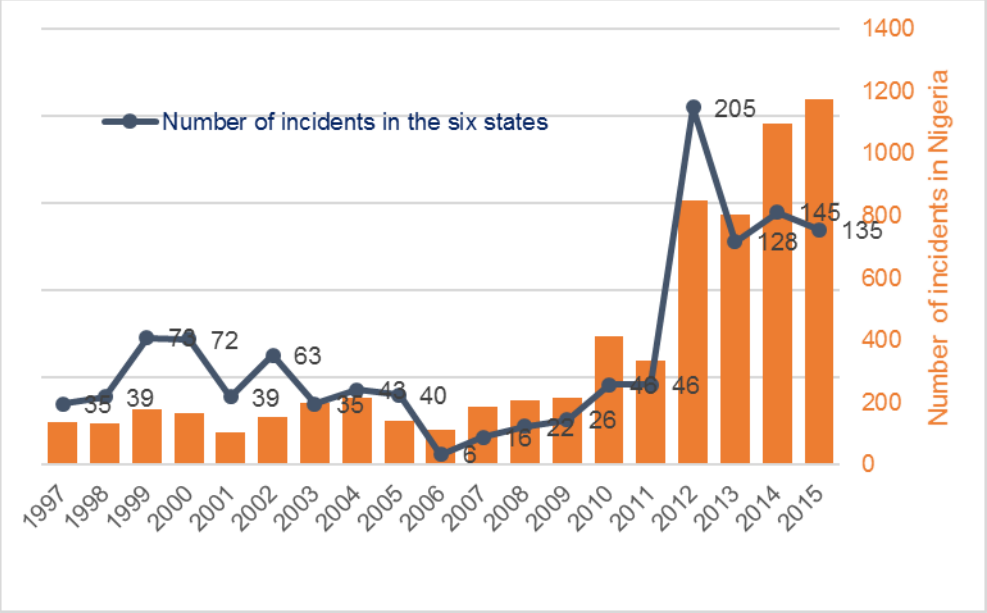
Insecurity and violence are potential barriers to the inclusion of all children in schools, and to the implementation of school improvement programmes in Nigeria. This is the case particularly in the three north-eastern states, where a state of emergency has been declared (Borno, Yobe and

[1] Source: ESSPIN analysis based on data from UBEC (http://www.ubeconline.com/financial_records.php)

Adamawa), but the problem is not limited to those states. Across Nigeria the number of recorded incidents of political violence and conflict have increased dramatically since 1997. In the six ESSPIN states, the incidence of violence also increased during this time, reaching their peak in 2012 and decreasing again somewhat after that (Figure 2). The majority of these incidents have been in Kaduna, Kano and Lagos. The northern ESSPIN states in particular have also been affected by large numbers of internally displaced persons migrating to flee violence associated with the Boko Haram fighting and attacks in the north-east of Nigeria. Kano schools in particular saw a huge surge in enrolments during the height of the insurgency.

Almost two-thirds of fatalities due to violent conflict during 2010–2015 were in Kaduna, where many stakeholders have stated that this has had adverse effects on the education sector. Conflict in Kaduna has included inter-ethnic conflict, raids on villages by armed groups, and clashes between military and civilians. During discussion with stakeholders, it was reported that armed robberies, attacks and kidnappings are common in certain areas in Kaduna. In some cases, whole communities are displaced as a result. Children are taken out of schools by the parents as they move away, which disrupts their learning process. In addition, some schools have become more insecure and there have been reports of attacks not only on communities but also on schools directly. The security situation in Kaduna has led to the closure of a number of primary schools, some of which have remained closed during the whole year. The psychological and emotional effects of violent conflict may be long lasting and varied (Coinco, 2014; Calder, 2016).

Figure 2: Incidence of political violence in Nigeria and in the six states in which ESSPIN works



Source: Authors' calculations based on Armed Conflict Location and Event Data Project (ACLED) (Raleigh *et al*., 2010)

Table 9: Fatalities due to political violence, 2010–2015

	2010	2011	2012	2013	2014	2015	Proportion of all fatalities, 2010–2015
Enugu	2		11	11	8	15	1%
Jigawa			1	9	17	2	1%
Kaduna	20	851	269	140	538	432	64%
Kano	6	19	339	186	359	89	28%
Kwara		0	5	26	11	11	1%
Lagos	15	4	17	35	62	25	4%
Total	43	874	642	407	995	574	100%

Source: Authors' calculations based on ACLED (Raleigh ***et al.***, 2010)

2 Methods

2.1 Evaluation strategy

2.1.1 Classifying the amount of ESSPIN intervention

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: roll-out prior to the 2012/13 school year), and to those where intervention had started more recently (Phase 2: roll-out in 2012/13 or 2013/14). In practice, the roll-out followed a different implementation plan driven by state governments, with the result that by CS3 all schools had received some level of ESSPIN SIP activities.

While faster roll-out and greater reach are signs of success for the ESSPIN programme, it presents a difficulty for evaluation as there is no longer a control group which has received no intervention. The nature, timing and intensity of ESSPIN's intervention varies widely both between and within states (Annex A shows the number of days of leadership training, teacher training and school visits that schools have received).

To simplify the analysis, we focus our analysis on the number of years of full Output Stream 3 intervention that schools have received. Full intervention means that the school has received some leadership training, some teacher training, and some school visits during the year, though the amount of each activity may vary. A further simplification groups the schools into minimum (zero to one years), medium (two to three years), and maximum (four to five years) intervention categories. We also assume that there is a one-year lag between ESSPIN intervention and measurable impact, and have therefore not considered intervention received during 2015/16 when grouping schools into different intervention categories.

Using this categorisation makes it clear what types of comparison will be possible in the analysis, and how these differ by state (Table 10). Only Kaduna has a substantial number of schools in all three categories. Enugu, Jigawa, and Kano have only minimum and medium categories; Lagos has medium and maximum categories; and Kwara has only the maximum category.

Table 10: Total number of schools in the six states by years of Output Stream 3 intervention

Intervention group	Minimum		Medium		Maximum	
Years of intervention	0	1	2	3	4	5
Enugu	0	820	318	91	0	0
Jigawa	0	1,065	436	436	0	0
Kaduna	0	3,361	601	0	165	0
Kano	0	4,973	255	254	0	0
Kwara	0	0	0	0	1,485	0
Lagos	0	6	0	903	0	97

The Composite Survey sample was selected initially based on the intended intervention groups at the time of CS1, using the intended phases of intervention roll-out as at 2011. In practice, the fact that states have controlled this process has led to differences in the way the intervention was really rolled out. Despite this, most categories of intervention are still well represented in our sample.

In certain cases, the small sample of schools within a particular treatment category means that our estimates may be biased or may have low statistical power for cross-group comparisons. For example, in the medium category in Jigawa, only four schools have received two years of intervention, while 71 schools have received three years. As a result, there may be some upward bias in our estimates of differences between medium and minimum intervention schools in Jigawa. We chose not to alter the CS2 sample so that we can make over-time comparisons for all of the schools in the sample (with the exception of the six schools that had to be replaced, as discussed below). This limits the analysis for certain sub-categories, which will be discussed in the associated analysis sections of this report.

For individual outcome indicators, we alter the classification scheme slightly according to the purpose of our analysis. For example, when examining teacher competence within the CS3 survey, we consider two different groups: teachers who are in schools that have received ESSPIN intervention but who have not themselves been trained by ESSPIN[2]; and teachers who have been trained through ESSPIN. When examining SBMC functionality and inclusive practices of SBMCs, we classify schools according to the amount of Output Stream 4 intervention received. Schools are classified as 'no intervention' (five or fewer days of Output Stream 4 intervention received), 'pre-CS1' (started receiving intervention in 2011/12 or prior to this), and post-CS1 (started receiving intervention in 2012/13 or after). Once again, the types of comparison differ by state. For example, all schools in Lagos fall into the post-CS1 intervention group and so no comparisons by Output Stream 4 intervention are possible for Lagos. All other states have schools from at least two of the intervention groups.

Learning outcomes – literacy and numeracy in Grades 2 and 4 – are analysed using item response theory (IRT), providing an overall scale of how well children have scored, as well as a grouping of children into levels corresponding to the level they are expected to reach by the end of each grade. The distributions of children in schools with minimum, medium and maximum intervention levels, and in CS1, CS2, and CS3, are compared. Percentage scores in specific test sub-scales (in literacy these are labelled *receptive*, *fluency in reading and writing*, and *productive*; in numeracy they are *calculation*, *everyday maths*, and *word problems*) are also analysed by intervention group and over time.

Teacher tests – literacy and numeracy – are analysed in a similar manner, also using IRT. We analyse teachers' results in similar ways to learners' results. The analysis includes an overall scale of how well teachers have scored, groupings of teachers according to which grade level they have achieved, and analyses of specific sub-scales within the literacy and numeracy tests.

2.1.2 Modes of analysis

The purpose of CS3 is both to provide insights into the changes state-wide over time in the six states in which ESSPIN operates, and to evaluate whether ESSPIN is having an effect in the specific schools in which 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 combined to give an overall indicator of school quality. Finally, ESSPIN's impact is measured in terms of improved pupil learning

[2] 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 who are in ESSPIN schools and so are expected to have received support from their head teachers and colleagues ('not ESSPIN-trained').

outcomes, which we ascertain through test scores in numeracy and English literacy at Grades 2 and 4. For each of these indicators we present in the following chapter two main types of analysis:

1. Change over time between CS1 and CS3, and between CS2 and CS3, for ESSPIN states as a whole. It is important to monitor change over time in how schools function and how much children are learning, both to inform programmes such as ESSPIN and for broader education policy-making. Trends over time in ESSPIN states are likely to reflect both the presence of the intervention and a number of other economic, social and political factors. ESSPIN aims to improve schools across the states in which it works. If the programme has been successful in this aim, then we would expect – other things being equal – that schools in CS3 have higher output, outcome and impact measures than schools in CS1 and CS2. In practice, however, many other things may not be equal. Changes in enrolment, student profile, state financing, and political commitment may all affect these indicators at the same time. We present these changes over time and, where information is available, consider what may be driving changes aside from ESSPIN's intervention.

We use statistical significance tests (t-tests) to give an indication of whether a difference in results between our samples is likely to reflect a genuine difference in the overall populations. Given two 'populations' or groups of interest that we wish to compare – say, schools in 2012 and schools in 2016 – a common approach is to take a random sample from each group and compare the average performance in one sample to that in the other sample. However, there will be some random variation between the two samples that is due to the set of schools that happened to be sampled. This random variation could result in differences between the two samples even when the two populations are the same. Statistical tests tell us the probability that a difference between the two groups occurred by chance due to random variation in the samples, as opposed to being due to genuine differences in the two populations that the samples were drawn from. When we are looking at change over time, the t-test tells us the probability that a difference between our 2012 and 2016 sample is due to chance variation between the samples, as opposed to reflecting a genuine change over time. A probability (sometimes known as the 'p-value') of 5% or less is often taken to be a good threshold for accepting that there is a genuine change, and we mark the result with an asterisk (*) when this is the case.

2. Differences between the different levels of intervention categories (minimum, medium and maximum) within the CS3 results. We hypothesise that schools that have received more years of full ESSPIN intervention will have higher output, outcome and impact measures than schools which have received fewer years of intervention.

To test this, we use a continuous measure of the years of full intervention that each school has received (one to six), and calculate the estimated effect of having received one additional year of intervention using a simple regression model with dummy variables for each state. This approach comes one step closer to estimating the effect of ESSPIN's intervention, by controlling for one of the main confounding variables (the state). However, this will not be a conclusive indicator of ESSPIN's effect because there are also differences in school and learner background characteristics within states. The overall technical report on CS2 showed that even within states, schools with some characteristics – typically larger, more urban schools – were disproportionately selected for the early phases of ESSPIN roll-out.

Controlling for these school and learner background characteristics is a more difficult statistical exercise, so we only attempt this for our outcome measure, overall school quality, and our impact measure, pupil learning outcomes. For these indicators, we conduct additional analyses in order to understand what basis there might be for making causal attribution of ESSPIN's impact. This analysis is described in Sections 5 and 6.

2.2 Sample and weights

2.2.1 Sample design

In CS3, all the schools visited in CS2 were visited again with the intention of collecting data that would enable us to draw inferences about what is happening in the population of schools across the six states, and within each state, through the use of sample weights. As in CS2, the sample design in CS3 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 and differences between schools having received different amounts of ESSPIN intervention.

The third round of the Composite Survey was carried out in all six ESSPIN states. The sample consisted of 735 schools across the six states. Most of these were schools that had been visited as part of CS2, although 16 replacements had to be made for schools which no longer exist or were found to be ineligible.[3] Each replacement was made using the same sampling frame, stratification, and approach as in CS2. Replacement schools were selected with probability proportionate to size (measured by number of teachers) from the same LGAs as the schools that were replaced. The number of schools sampled in each of the categories (as defined in CS3, so taking account of the full period of intervention) is shown in Table 11.

Table 11: Sample in CS1, CS2 and CS3 and total number of schools, by state and with intervention groups

	Category for sampling purposes	CS1 sample (2012)	CS2 sample (2014)	CS3 sample (2016)	Population (Total number of schools in each category and state)
Enugu	Minimum	35	35	35	820
	Medium	35	70	70	409
Jigawa	Minimum	32	30	30	1,065
	Medium	71	73	75	872
Kaduna	Minimum	28	61	61	3,361
	Medium	42	42	42	601
	Maximum	35	37	37	165
Kano	Minimum	56	116	121	4,937
	Medium	46	54	54	509
Kwara	Maximum	102	105	105	1,485
Lagos	Medium	69	69	69	903
	Maximum	34	36	36	97

Note: The sample size shown is the actual sample for which data were collected. Intervention groups reflect the number of years of intervention the schools had received by the end of the 2014/2015 school year.

Within each school, the survey team conducted interviews with the head teacher, the SBMC chairperson or deputy, teachers and learners.

[3] Full details of the sampled schools and replacements were submitted to ESSPIN as an annex to the CS3 Training Report and Fieldwork Plan.

As in CS2, we sampled six teachers per school, or all of the eligible teachers in schools with fewer than six teachers. However, for CS3, we attempted to find the six teachers interviewed during CS2, using their photographs and name information, and to interview them again to be able to assess changes over time, as well as rates of change in teacher competence and test results, with more precision. In total, we were able to track 53% of the CS2 teacher sample (Table 12). The main reasons why teachers could not be tracked were that they no longer taught at the same school (31% of CS2 teachers) or were not present on the day (10%).[4]

Table 12: CS2 teachers that were re-sampled during CS3 (%)

Enugu	Jigawa	Kaduna	Kano	Kwara	Lagos	Total
52	53.9	57	47	55.5	54.6	53

In schools where fewer than six teachers from CS2 were re-sampled, the sample was topped up to six teachers with teachers randomly selected from the head teacher's 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 the additional top-up teachers to be sampled. We excluded teachers who only teach religious subjects and those that do not teach in Grades 1–6.

As in CS2, 16 learners were sampled per school in CS3, with a sample of four learners per school for each of the tests (Grade 2 literacy, Grade 2 numeracy, Grade 4 literacy and Grade 4 numeracy).[5] For CS3, we attempted to sample learners who were currently being taught by one of the sampled teachers to allow us to better link teachers with learners in the analysis, and examine, for example, whether learners with better learning outcomes were taught by more competent teachers. To do this, we first sampled teachers in CS3. While sampling them, we gathered information from the head teacher on which teachers teach which arms of Grades 2 and 4, filling in an 'arm eligibility grid'.[6] From this, field teams determined the 'eligible arms' (i.e. arms where at least one of the sampled teachers was teaching) and sampled learners only from these arms. To ensure that the full sample size was maintained, the sample was topped up if there were fewer than eight learners in the eligible arms. In addition, if none of the sampled teachers taught in Grade 2 (or Grade 4), learners were randomly selected from all arms in that grade.

In CS2, learners were sampled from the learner attendance register (if available), but these were found to not always be accurate. For CS3, each learner in the eligible arms was given a 'sampling card', which was either blank or contained a number. Learners who were handed cards with even numbers were sampled for the numeracy test; learners with odd numbers were sampled for the literacy test. This ensured that the sampling was done from the learners actually present in the class, rather than from the potentially incomplete pupil register.

Although it would have been useful to trace the same learners over time, this was not seen as feasible because, for the children sampled in CS1, we only have their names, which was not always sufficient information with which to identify the same children three to four years later. We

[4] In addition, in a few cases teachers no longer taught in Grades 1–6 or no longer taught non-religious subjects, and were therefore ineligible. Teachers could also not be tracked if the school was replaced during CS3.
[5] This was true for all schools except those in Lagos, where our sample of Grade 4 learners needed to be large enough to allow for a comparison with private schools (which were surveyed as part of a separate exercise at the same time). We therefore increased the sample to eight literacy and eight numeracy tests for Grade 4 in Lagos. Therefore a total of 24 learners were sampled in total for Lagos, eight learners in Grade 2 and 16 learners in Grade 4.
[6] Large school grades in Nigeria are often divided into multiple 'arms', usually taught by different teachers in different classrooms. In some cases, however, multiple arms exist in name only and are combined in a single classroom and taught by a single teacher.

therefore collected a random sample within each school in CS3, as in previous rounds of the survey.

In addition to the main sample of 16 learners and six teachers, an additional four learners 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 learners 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 rarely.

A number of schools operate double shifts, with some classes taught in the morning and others in the afternoon. Where a school had double shifts, and the teachers who teach in the morning are different from those who teach in the afternoon, we sampled morning teachers from teachers who were present when we arrived in the morning, and afternoon teachers from the teacher attendance record for the previous day's afternoon. If the sampled teachers did not turn up in the afternoon, we used replacements. We sampled learners from the arms taught by these sampled teachers. If this included arms taught in the afternoon, we sampled from the children who had attended the previous afternoon, according to the pupil register.

In double-shift schools where the same teachers taught both morning and afternoon on the day of the visit, we sampled teachers who were present in the morning. We sampled learners from the arms taught by these sampled teachers. If these included arms taught in the afternoon, we sampled from the children who had attended the previous afternoon, according to the pupil register.

The intended survey sample based on this sampling strategy is presented in Table 13 below.

Table 13: Composite Survey 3 – Intended survey sample

Targeted sample size						
State	**Schools**	**Head teachers**	**SBMC members**	**Teachers**	**P2 learners**	**P4 learners**
Lagos	105	105	105	630	840	1680
Enugu	105	105	105	630	840	840
Kwara	105	105	105	630	840	840
Jigawa	105	105	105	630	840	840
Kano	175	175	175	1,050	1,400	1,400
Kaduna	140	140	140	840	1,120	1,120
Total	**735**	**735**	**735**	**4,410**	**5,880**	**6,720**

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 the profile of schools included in the survey is not identical to the profile of schools in the state as a whole. We address this by applying sample weights, which give greater weight to the results obtained from schools that are relatively under-represented in the survey[7]. Sample weights were calculated for the CS1, CS2 and CS3 schools,

[7] For example, in Kano, three-quarters of schools had one year of intervention, while one-quarter had more than one year of intervention. If we sampled the schools in the same proportion, then there would not be enough schools with more than one year of intervention in the sample to draw firm conclusions about them. We therefore over-sample the schools with more than one year of intervention, and use sample weights to ensure that the different intervention categories are proportionately represented in our estimates.

teachers, and learners. A smoothing technique was also applied to reduce the variability of the weights and to avoid the design effects problem encountered in the CS1 analysis (see Megill, 2014b).

Most of the analysis that follows 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 over time. For this we are limited to the set of schools which were sampled at each of the time points over which the comparison is conducted (e.g. an analysis of change in individual schools between CS2 and CS3 is limited to those schools included in both the CS2 and CS3 rounds of the survey). Additional sets of weights were calculated for use with these 'panels' of schools.

In addition, because we re-sampled teachers from CS2, in some of our analyses we compare how the same teacher performed in CS3 compared to CS2. For this we are limited to the set of teachers who were sampled during both CS2 and CS3. Another set of weights was calculated for use with this panel of teachers.

2.2.3 Sample coverage

Table 14 lists the number of respondents covered by each of the survey instruments, and how this compared to the number of targeted respondents in each category. In line with the targeted sample, the field team successfully visited 735 schools across the six states. Head teachers and SBMC chairs/deputies were interviewed in all the schools visited. The number of teachers interviewed fell short of the targeted sample size in all states. The primary reason for this was that in many schools, fewer than six teachers who teach non-religious subjects in Grades 1–6 were present on the day and were therefore eligible to participate. In all six states, sample coverage of *eligible* teachers was above 99.5%. There were some further minor differences between the numbers of teachers interviewed and those who were tested or whose lessons were observed because some teachers did not give their consent to be tested or observed, and because some teachers became ill and so could not complete the test or lesson. Some teachers were also not able to stay for the teacher tests which were held after lessons had finished.

Similarly, the number of learners assessed fell slightly short of the targeted number because some schools had fewer than eight learners in Primary Grade 2 (P2) or Primary Grade 4 (P4) (fewer than 16 learners in the case of P4 in Lagos). Overall, complete test data was gathered in CS3 for 95.6% of the maximum possible number of students, a similar proportion to CS1 and CS2.

Table 14: Sample coverage in CS3

State	Schools			Teachers				Learner assessments			
	Schools	Head teachers	SBMCs	Inter views	Lit tests	Num tests	Less obs	L2	N2	L4	N4
No. of schools and respondents covered											
Lagos	105	105	105	592	586	584	591	415	413	807	805
Enugu	105	105	105	547	532	533	537	391	377	397	391
Kwara	105	105	105	495	486	488	492	385	373	380	366
Jigawa	105	105	105	462	436	449	462	408	406	411	410
Kano	175	175	175	838	809	812	837	696	691	671	670
Kaduna	140	140	140	653	632	627	653	552	549	542	539
Total	735	735	735	3,587	3,481	3,493	3,572	2,847	2,809	3,208	3,181
Sample coverage (% of targeted sample size)											
Lagos	100	100	100	99.7	99	98	99.5	99	98	96	96
Enugu	100	100	100	99.8	97	97	98	93	90	95	93
Kwara	100	100	100	99.6	98	98	99	92	89	90	87
Jigawa	100	100	100	100	94	97	100	97	97	98	98
Kano	100	100	100	99.8	96	97	99.6	99	99	96	96
Kaduna	100	100	100	100	97	96	100	99	98	97	96
Total	100	100	100	99.8	97	97	99.4	97	96	95	95

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. The 'targeted sample size' for teachers represents six teachers per school, or the number of eligible teachers in schools where this is less than six.

2.3 Training, pilots and fieldwork model

Fieldwork for CS3 was conducted using CAPI during April–June 2016. Following consultation with ESSPIN staff and DFID, several changes were made to the instruments, and some innovations were introduced from other recent Nigerian school surveys (see De *et al*., 2016; Pellens *et al.*, 2016). At the same time, we retained the questionnaire items required for comparability with previous rounds of the Composite Survey. Changes made to the instruments included the following:

- A scale for measuring teacher motivation and the quality of interaction between teachers, which has been used in other recent Nigerian school surveys (Cameron, 2015b), was added.
- Additional items designed to assess school management more broadly, including items on action on increased enrolment, issues related to teacher attendance, action on making sure teachers are present in classrooms during lessons, action to improve school materials or infrastructure. We also visually inspected whether teachers and learners were in class, around the beginning of the day and after the long morning break.
- The overall technical report on CS2 (Cameron, 2015a) noted that there were apparently large increases in enrolment in some states, along with increases in pupil–classroom ratios and PTRs. To better understand enrolment changes we added questions in the head teacher questionnaire to clarify the number of children currently enrolled in the school and the number of teachers. Head teachers were also asked about changes in enrolment that had occurred

compared to the previous year, and where they report increases in enrolment, what issues it might have caused and how they had dealt with the change.

- An expanded version of the wealth index was included in the Grade 4 learner assessments, to better control for children's home backgrounds.

Tests were also altered to include more difficult items and to avoid 'ceiling effects', which occur when a test lacks items targeted towards the top end of the distribution and is consequently not very good at distinguishing among these top students. Respondents who did well in a filter question administered towards the beginning of the test were directed towards the more difficult version, while those who did not answer the filter question correctly were directed to the original version of the test.

As in CS2, learner assessments in CS3 were administered using CAPI. 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 children were to be asked orally, 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 learner. A number of changes were made to the CAPI systems and manuals for the administration of the learner assessments to make them easier to train on and administer. This included a clear manual with consistent instructions across questions of a particular type, automated timers for timed questions, and translations into Hausa, Igbo and Yoruba of text that did not need to be read in English.

The instruments were pre-tested over two days in Abuja during April 2016. State coordinators and monitoring officers collected the data on CAPI after they had been trained on the instruments. Minor revisions were made to the instruments in consultation with state coordinators.

Table 15 lists the instruments used in CS3, 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 learner-level information on socioeconomic status, age, language spoken at home, and gender. The data gathered in general allow 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 an anonymised form for use by ESSPIN and other researchers.[8]

The process of revising instruments for CS3 does leave some possibility of measurement error in comparisons between the previous Composite Surveys and CS3. We tried to ensure consistent and manageable data collection within CS3 by setting clear guidance for data collectors through detailed data collection manuals, applying constant oversight, and providing intensive training for all data collectors, including three pilot field days. Although we avoided large changes in instruments that would compromise comparability with CS1 and CS2, 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, since changes in measurement are consistent across the different intervention categories, they should not affect any within-CS3 comparisons.

[8] The data will be published on the World Bank Microdata Catalog (microdata.worldbank.org) and/or National Bureau of Statistics (nigerianstat.gov.ng) websites.

Table 15: Instruments used in CS3

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 learner 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 has a trained facilitator; action for commonly excluded groups; SBMC raised issue of children's exclusion.
Structured interview with teacher	Knowledge of English and mathematics 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	Teacher test scores in English literacy and numeracy.
Learner assessments	Learner assessment scores in English literacy and numeracy at Grades 2 and 4.
General observation	Length of morning break; number of classes where learners and teachers are in class within half an hour of starting time and long morning break.

3 School management and head teachers

ESSPIN's interventions include leadership training for head teachers on managing the school and its teachers, improving academic standards, planning for the school's development and advocating for more resources, and ensuring the school is inclusive. ESSPIN also encourages the development of SBMCs, and provides support and training, including on how the committees can encourage the participation of women and children. This chapter examines how well schools in the six states are doing in each of these aspects.

A number of indicators relating to school management, inclusiveness, and SBMCs are defined in the ESSPIN logframe. The logframe groups these indicators into a number of 'standards' or composite indicators, each comprised of several discrete criteria. These are as follows:

1. Head teacher effectiveness: Being an effective head teacher is considered to include a number of practices, such as observing teachers' lessons, holding professional development meetings with the teachers to talk about teaching and learning, monitoring teacher attendance, keeping records, and ensuring the school sticks to a regular schedule.

2. School development planning: Schools in the six states are encouraged to undertake a self-review process involving the head teacher, teachers, SBMC, parents and other community members. The aim is to identify the strengths and weaknesses of the school and then list the steps to be taken to improve it in an SDP. Where needed, the plan can also be used to request resources from local government or the community. The relevant logframe standard measures whether a self-evaluation has been undertaken, whether the school has an SDP, and whether it has carried out the activities in its SDP.

3. School inclusiveness: The extent to which the school is inclusive – makes efforts to include all learners, including those from disadvantaged backgrounds – is measured by the action taken, and steps listed in the SDP, for increasing access, and by the extent to which teachers encourage girls and boys from all parts of the classroom to participate during lessons.

4. SBMC: SBMCs are supposed to act as the key link between the school and the community. This standard measures the extent to which they are functioning and active. Additional standards measure how effective they are in ensuring the participation of women and children.

The following sections explain each of these standards and present results on the change over time between 2012 and 2016, and the differences between schools which have had more years of ESSPIN intervention or fewer years of intervention.

3.1 Head teacher effectiveness

The ESSPIN logframe defines head teacher effectiveness in terms of seven criteria (

Box 1: Logframe standard for head teacher effectiveness

). These reflect both activities carried out by the head teacher, observations of the school, and responses from teachers and learners, 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).

Overall, we find 18% of head teachers in the six states are meeting the overall standard for effectiveness, compared to 14% in 2012 and 2014. However, this change is not statistically significant. There are significant improvements in some of the individual criteria. On average, head teachers had carried out one additional lesson observation per fortnight in 2016 compared to in 2014, and half of the head teachers had carried out two or more lesson observations in the last two weeks, compared to under 10% in 2012. Head teachers appear to be getting better at observing their teachers regularly, and in documenting their observations with lesson observation sheets. There has also been rapid progress in the proportion of head teachers who carried out at least one professional development meeting during the last school term, from 11% in 2012 to 40% in 2016.

Schools have worsened since 2012 in terms of our indicator of clear opening time – the extent to which learners and teachers agreed on what the school's opening time is. This appeared to be a result of confusion among learners in particular: on average, only 35% of the learners in each school could agree on its opening time. The reasons for the worsening over time are not clear. However, it can be questioned whether this is a good indicator of school management. Field observations suggested that children were confused over whether to consider the time that they arrived at the school, the time of assembly, or the time when lessons started, as the school opening time.

Fewer schools also conformed to a 35-minute lesson length in 2016 than in 2012. A length of 35 minutes was formerly considered the standard lesson length across the six states, although there is some variation in this and it is not clear whether it is still a useful indicator. Around 11% of lessons observed were longer than 40 minutes, although this varied between 45% in Lagos and 2% in Jigawa. These longer lessons are likely to reflect a move towards one-hour lessons for literacy and numeracy, and should arguably be discounted as an indicator of school quality. However, 55% of the observed lessons were shorter than 30 minutes. This may partly reflect the effect of observation on teachers. For example, they may be teaching components of lesson plans discussed during training, in a bid to impress the observer, but be unable to work these ideas into a full lesson. Even if this is the case, the short lesson times suggest that teachers have difficulty in planning lesson activities that fill a set duration.

Better indicators of whether schools maintain a clear schedule are the proportion of teachers and learners present in their classrooms within 30 minutes of the school opening time, an indicator based on direct observation. This has worsened since 2012, with around 63% of classrooms having both teacher and learners present at the expected time, compared to 74% in 2012. This suggests that schools continue to have serious problems with teacher and learner attendance at the start of the school day.

Box 1: Logframe standard for head teacher effectiveness

A head teacher must ensure that five out of the following seven 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 two or more professional development meetings during the last school term;[9]
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 learners 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; and
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).

Table 16: Head teacher effectiveness in 2012–2016

	2012 (CS1)	2014 (CS2)	2016 (CS3)	Change 2012–16	Change 2014–16
(1) Lesson observations (%)	9.2	20.3	50.0	+40.8*	+29.6*
-- No. lesson observations in past two weeks	1.2	1.4	2.4	+1.2*	+1.0*
(2) Professional development meetings (%)	11.4	20.9	40.2	+28.8*	+19.4*
-- No. professional development meetings last term	1.0	1.5	1.4	+0.4	-0.1
(3) Action on teacher attendance (%)	83.1	53.2	50.7	-32.4*	-2.5
(4) Clear opening time (%)	49.8	28.6	10.4	-39.5*	-18.2*
-- Learners who agree on opening time (%)		53.8	35.0	n/a	-18.8*
-- Teachers who agree on opening time (%)		69.6	73.4	n/a	+3.8
(5) More than 50% in class on time in morning (%)	72.4	62.1	68.9	-3.5	+6.9
-- Classes where learners and teacher are present on time (%)	74.3	67.1	62.9	-11.4*	-4.2
(6) Appropriate morning break (%)	78.9	75.9	76.3	-2.6	+0.4
(7) 35-minute lesson length (%)	30.3	51.3	27.1	-3.2	-24.2*
Number of criteria fulfilled (out of seven)	3.4	3.1	3.2	-0.2	+0.1
Effective head teacher (five out of seven criteria met) (%)	13.6	14.2	17.8	+4.1	+3.6

Note. * indicates that the change over time is statistically significant ($p < .05$).

Are the schools that have had more years of intervention under ESSPIN's Output Stream 3 doing better than those which have received intervention only recently? We present averages in our three intervention groups: those that have only had one year of full intervention up to 2014–15 (minimum); those that have had two to three years (medium); and those that have had four to five years (maximum) (Table 17). As noted in Section 2 above, the differences between these three

[9] In earlier reports, this indicator has been mislabelled as 'Held four or more professional development meetings since the start of the 2011/12 or 2013/14 school year'.

groups cannot necessarily be attributed to the intervention. The pattern of results is sometimes difficult to interpret. For example medium intervention schools are sometimes doing better than both minimum and maximum intervention schools. We also know that the differences between groups may reflect differences between states. For example, a disproportionate share of the medium intervention schools are located in Lagos. To help make sense of this, we estimate the effect of one year of full intervention, controlling for the state, and present this alongside the estimated averages for each group.

Schools that have received more years of intervention do have more effective head teachers. There is a small but statistically significant difference of 0.2 in the number of criteria head teachers fulfil associated with an additional year of intervention. 14% of head teachers in the minimum intervention group meet the overall standard, compared to 26% of head teachers in the medium group and 24% in the maximum group.

Head teachers in schools that have received more years of ESSPIN intervention are more likely to have observed their teachers' lessons, had held more professional development meetings in the last term, and could name more actions that they had taken on teacher attendance. The school was more likely to have a clear opening time – although curiously, this effect is only found among learners and not among teachers.

In CS3, field teams observed how many classes had learners and teachers present both at the start of the school day and at the end of the long morning break. In both cases they found nearly all classrooms had at least some learners present on time,[10] but, on average, only 64% of the classrooms in each school had a teacher present on time at the start of the day. However, teacher presence in the classroom rose by around 5% for each year of ESSPIN intervention. At the end of the long morning break, we again found only 64% of classrooms with teachers present across the six states. Teacher presence after the morning break appeared to be higher in schools that had received more years of ESSPIN intervention, but the effect is not statistically significant once we control for differences between states. We also recorded teacher absenteeism – the proportion of teachers absent each day judging by the school's teacher attendance book (close to 100% of schools kept a teacher attendance book.) On average, only around 80% of the school's teachers were present each day. This goes some way towards explaining why over one-third of classrooms did not have a teacher present during our observations. Some teachers may also be present in school but not teaching, or there may be insufficient teachers for the number of classes.

There is a significant negative effect of ESSPIN's intervention on whether observed lessons are around 35 minutes or not. As noted above, this may reflect the use of 60-minute lessons for literacy and numeracy in some states, and could also involve an observer effect whereby teachers try to adhere to lesson plans from their training materials. (Anecdotally, during pilots and fieldwork monitoring, a small number of lessons were observed being given disproportionately often by teachers. This could reflect the printed lesson plans used in ESSPIN training, resulting in the same lesson being taught in all schools at more or less the same time. Alternatively it could be because they had practised these lessons or received training on them, and hoped to impress the observer by giving them.)

[10] There may be some upward bias here, since for a classroom with neither learners nor teachers present, the head teacher could claim that the room was not in use for teaching.

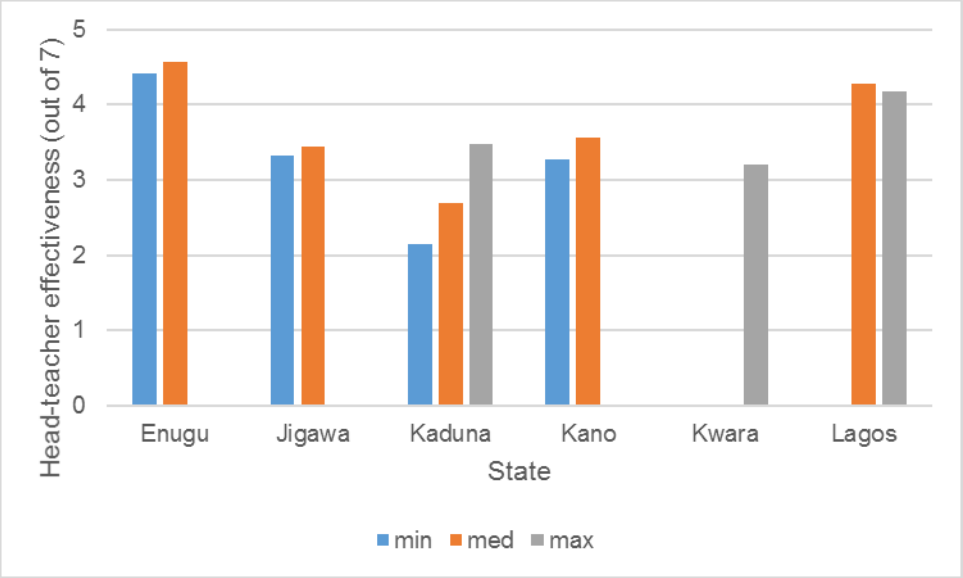
Table 17: Head teacher effectiveness in 2016, by ESSPIN intervention

	Min. (1 year)	Med. (2–3 years)	Max. (4–5 years)	Estimated effect of one year of full intervention
(1) Lesson observations (%)	44.7	71.9	37.9	+7.4*
-- No. lesson observations in past two weeks	2.3	3.3	1.5	+0.4*
(2) Professional development meetings (%)	33.1	60.6	42.2	+5.2
-- No. professional development meetings last term	1.1	2.0	1.7	+0.2*
(3) Action on teacher attendance (%)	47.2	64.8	43.6	+7.8*
-- School has a teacher attendance book (%)	98.8	99.8	100.0	+0.6
(4) Clear opening time (%)	8.8	15.7	9.5	+3.2*
-- Learners who agree on opening time (%)	32.5	40.1	40.1	+3.6*
-- Teachers who agree on opening time (%)	74.1	71.7	72.8	-3.2*
(5) More than 50% in class on time in morning (%)	63.6	75.7	86.7	+7.1
-- Classes where learners present on time (%)	97.9	97.8	99.8	-0.3
-- Classes where teachers present on time (%)	58.4	71.4	79.6	+5.4*
-- Classes where learners and teachers present on time (%)	57.9	69.6	79.6	+5.0*
(6) Appropriate morning break (%)	78.1	70.9	77.6	+0.2
(7) 35-minute lesson length (%)	30.1	11.5	40.2	-9.1*
Number of criteria fulfilled (out of seven)	3.0	3.7	3.3	+0.2*
Effective head teacher (five out of seven criteria met) (%)	14.1	26.0	24.2	+5.2*
Additional indicators				
(A1) More than 50% in class on time after break (%)	68.9	77.1	88.0	+4.2
-- Classes where learners present on time (%)	97.7	99.1	99.2	+0.3
-- Classes where teachers present on time (%)	62.6	74.0	83.6	+3.9
-- Classes where learners and teachers present on time (%)	62.4	73.9	83.6	+4.2
(A2) Enrolment increased since last year (%)	62.3	56.3	58.6	+0.1
(A3) Teacher absenteeism (%)	21.9	16.5	23.3	-0.9

Note. * indicates that the coefficient on years of ESSPIN intervention is statistically significant ($p < .05$) in a regression on the stated indicator, with controls for state.

Looking across the states (Figure 3), there is a consistent pattern of schools that have received more years ESSPIN intervention achieving more of the head teacher effectiveness criteria than those that have received fewer years of intervention. Our analysis here does not seek rigorously to establish a causal link, but the pattern of results is certainly consistent with ESSPIN's intervention having positive effects on head teacher effectiveness. However, average schools in Jigawa, Kaduna, Kano and Kwara are only achieving around three of the seven criteria, and across the states as a whole there has been little change since 2012. Low results among the schools with less intervention in Kaduna suggests that these schools in particular might benefit from further intervention. In Kwara, the relatively low number of criteria met may reflect a reduction in the number of leadership training days.

Figure 3: Head teacher effectiveness in 2016, by state and intervention group



3.2 School development planning

As noted above, ESSPIN's leadership training encourages and supports head teachers to review how the school is doing each year and to put together a plan for the development of the school, which can be used to advocate for more resources from local government or from the community. Ideally, the plan will not just target infrastructure improvement (e.g. a school fence, a toilet block), but will also identify activities relating directly to strengthening teaching and learning, and activities to improve access – particularly for children from disadvantaged backgrounds. SBMC members are also trained on using a cashbook to record the school's expenditures and income.

The definition of effective school development planning is based on five criteria (Box 2). In the CS2 report, we noted there had been some progress in these indicators between 2012 and 2014. In 2016, there appears to have been much more additional progress, with large and statistically significant increases in all five indicators (Table 18). This may reflect the roll-out of the ESSPIN intervention across the six states as a whole; school development planning appears to have become a much more widespread practice. 62% of schools had conducted a self-evaluation, and 75% had a SDP available, compared to only around one-quarter of schools in 2014.

Around one-third of schools had plans which contained three or more activities relating to strengthening teaching and learning, and 18% could show that four or more activities had been carried out. While still low, these proportions were much higher than in 2012 or 2014. Over 80% of schools had a cashbook – compared to under 40% in 2012. Only 27% of schools had actually kept

the cashbook up to date, but this was again a substantial improvement, roughly double the proportion meeting this criterion in 2012.

Overall, the average school met around two of these five criteria, and fewer than 20% met the overall standard for school development planning. Still, this is a massive improvement on 2014, when the average school met only one criterion and only 7% of schools met the standard.

Box 2: Logframe standard for effective school development planning

The school must meet criterion 1 and criterion 2 listed below, and at least two out of the remaining three 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 learning;[11]
4) physical evidence of four or more activities stated in SDP having been carried out;[12] and
5) cashbook is up-to-date (balanced in the last 60 days).

[11] 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 learner 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.

[12] 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 18: SDP effectiveness in 2012–2016

	2012 (CS1)	2014 (CS2)	2016 (CS3)	Change over time	
				2012–16	2014–16
(1) Written evidence of school self-evaluation process (%)	20.4	24.3	61.7	+41.3*	+37.4*
(2) SDP available (%)	20.4	26.2	75.1	+54.6*	+48.9*
(3) SDP contains three or more activities to strengthen teaching and learning (%)	9	13.2	33.5	+24.5*	+20.3*
-- No. activities in SDP to strengthen teaching and learning	0.4	0.8	1.9	+1.5*	+1.1*
(4) Evidence that four or more activities stated in SDP carried out (%)	4.7	5.9	18.4	+13.7*	+12.5*
-- No. activities in SDP carried out	0.3	0.6	1.7	+1.4*	+1.1*
(5) Cashbook up-to-date (%)	13.1	18.3	26.8	+13.7*	+8.5*
-- School has a cashbook (%)	37.1	44.5	81.3	+44.2*	+36.8*
Number of SDP criteria fulfilled (out of five)	0.6	0.9	2.2	+1.5*	+1.3*
School meets effective school development planning standard (four out of five criteria met) (%)	3.8	7.4	18.6	+14.8*	+11.3*

Note. * indicates that the change over time is statistically significant ($p < .05$).

For each of the criteria except the first (school self-evaluation process) there were large and statistically significant differences between schools where there had been more years of ESSPIN intervention and those that had received less intervention, even after controlling for state. We estimate that one year of ESSPIN intervention is associated with an additional 0.4 criteria met. As in the previous section, this is not a direct estimate of the causal impact of ESSPIN, but it is at least consistent with the idea that widespread roll-out of ESSPIN's intervention has had an impact on school development planning.

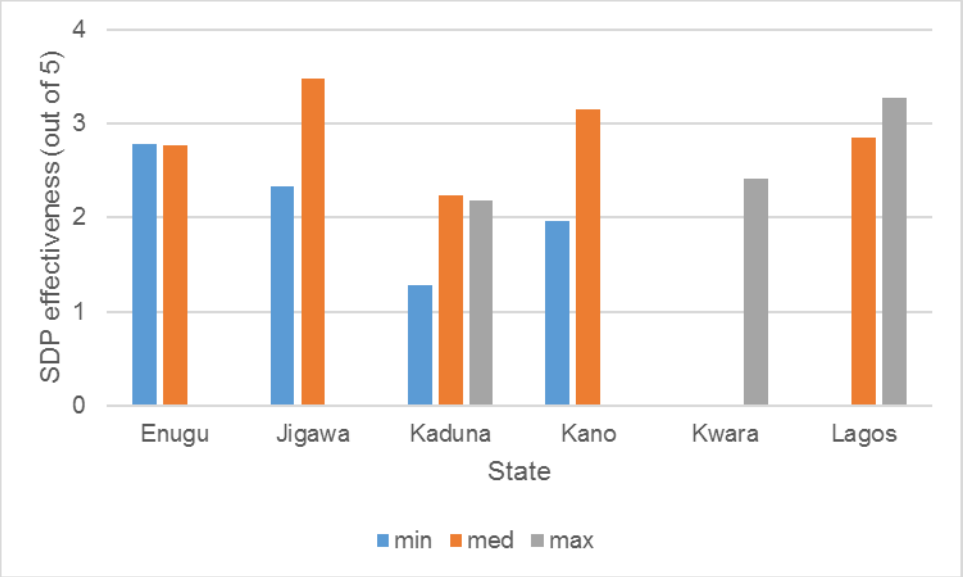
Table 19: SDP effectiveness in 2016, by ESSPIN intervention

	Min.	Med.	Max.	Estimated effect of one year of full intervention
(1) Written evidence of school self-evaluation process (%)	53.4	83.5	67.2	+7.9
(2) SDP available (%)	67.5	93.2	83.7	+13.9*
(3) SDP contains three or more activities to strengthen teaching and learning (%)	25	53.5	44.4	+8.5*
-- No. activities in SDP to strengthen teaching and learning	1.7	2.6	2.3	+0.3*
(4) Evidence that four or more activities stated in SDP carried out (%)	13.9	29.7	22.2	+7.8*
-- No. activities in SDP carried out	1.3	2.6	2	+0.6*
(5) Cashbook up-to-date (%)	24	35.5	26	+5.2*
-- School has a cashbook (%)	80.7	87.2	73.5	+1.3
Number of SDP criteria fulfilled (out of five)	1.8	3	2.4	+0.4*
School meets effective school development planning standard (%)	11.1	36.3	28.3	+8.7*

Note. * indicates that the coefficient on years of ESSPIN intervention is statistically significant ($p < .05$) in a regression on the stated indicator, with controls for state.

Examining differences across the states (Figure 4), high scores for schools that have received more years of ESSPIN intervention are particularly evident in Jigawa and Kano. Schools in Kaduna, and the minimum intervention groups in Jigawa and Kano, are doing the worst, meeting between 1.3 (Kaduna minimum intervention) and 2.3 (Jigawa minimum intervention) criteria on average.

Figure 4: SDP effectiveness in 2016, by state and ESSPIN intervention



3.3 School inclusiveness

The criteria on school inclusiveness measure the extent to which the school makes efforts to include all learners, including those from disadvantaged backgrounds. The overall standard for school inclusiveness in ESSPIN is based on four criteria (Box 3). Further detail on these is provided in the companion Gender and Inclusion Report.

In the CS2 report, we noted that one criterion – the number of activities in the SDP on improving access – had increased significantly between 2012 and 2014. This criterion increased further between 2014 and 2016 by a large and significant amount: 23% of schools now include two or more activities on access, compared to 5% in 2012 and 12% in 2014. This may reflect general improvements in school development planning, as noted in the previous section.

We also noted in the CS2 report that some criteria had declined significantly between 2012 and 2014. These included whether the head teacher had taken action on attendance, and spatially- and gender-inclusive behaviour by teachers in lesson observations. These indicators have neither worsened further nor significantly improved since 2014.

Overall, the proportion of schools reaching the inclusiveness standard has declined since 2012, and has not changed significantly since 2014. On average, schools meet only 1.3 of the four criteria – the same as in 2014. An alternative measure of inclusiveness, however, yields quite different results. We calculate a percentage score based on the number of actions to improve attendance, the number of activities in the SDP on access, the average number of assessment methods used, the average number of zones participating in each lesson observed (observers imagined the classroom as being divided into six zones), and a measure of the extent to which girls and boys participated equally in the class. This measure shows a large and significant *increase* since 2014. This suggests that behind the low proportions of schools meeting each criterion, there has actually been progress in the underlying behaviour.

Box 3: Standard for school inclusiveness (meeting needs of all learners)

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 actions that he/she has taken to improve learner 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); and
4) more than 50% of teachers observed met the spatial inclusion criterion (defined as engaging with at least one learner 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 20: School inclusiveness in 2012–2016

	2012 (CS1)	2014 (CS2)	2016 (CS3)	2012–16	2014–16
(1) Three or more actions on learner attendance (%)	57.9	28.7	26.5	-31.4*	-2.2
-- Number of actions on learner attendance	2.7	2.1	2	-0.8*	-0.1
(2) Two or more activities in SDP on access (%)	5.4	11.9	23	+17.7*	+11.2*
-- Number of activities on access	0.2	0.4	0.9	+0.7*	+0.4*
(3) >50% of teachers use two or more assessment methods (%)	70.7	62.3	55.1	-15.6*	-7.2
(4) >50% of teachers spatially inclusive and >50% are gender inclusive (%)	33.4	23.4	21.5	-11.9*	-1.9
Number of inclusiveness criteria fulfilled (out of four)	1.7	1.3	1.3	-0.4*	+0.0
Inclusiveness score (%)	72.2	63	83.8	+11.5*	+20.8*
School fully met standard (three to four criteria) (%)	18.8	10.5	11.4	-7.4*	+0.9
School partially met standard (two to four criteria) (%)	60.4	43	34.5	-25.9*	-8.5*

Note. The inclusiveness score is a total ranging from 0 to 100 and is calculated as follows: $\mathbf{20}(\frac{s_1}{7} + \mathbf{min}\left(\mathbf{1}, \frac{s_2}{5}\right) + \frac{s_3}{3} + \frac{s_4}{6} + s_5)$, where s1 is the number of actions to improve attendance; s2 is the number of activities in the SDP to improve access for disadvantaged children; s3 is the average number of assessment methods used by sampled teachers; s4 is the average number of classroom zones participating in the lesson during lesson observations, and s5 is the gender equity score (see below). * indicates that the change over time is statistically significant ($p < .05$).

In 2016, we find that schools that have received more years of ESSPIN intervention are more likely to have taken action on learner attendance (

Table 21). In the other criteria, the schools that have received maximum intervention appear to be doing better, but there is no significant effect once we control for state. Overall, there is a small but significant difference among the intervention groups in the number of criteria fulfilled, amounting to 0.1 extra criteria being fulfilled for each year of full ESSPIN Output Stream 3 intervention.

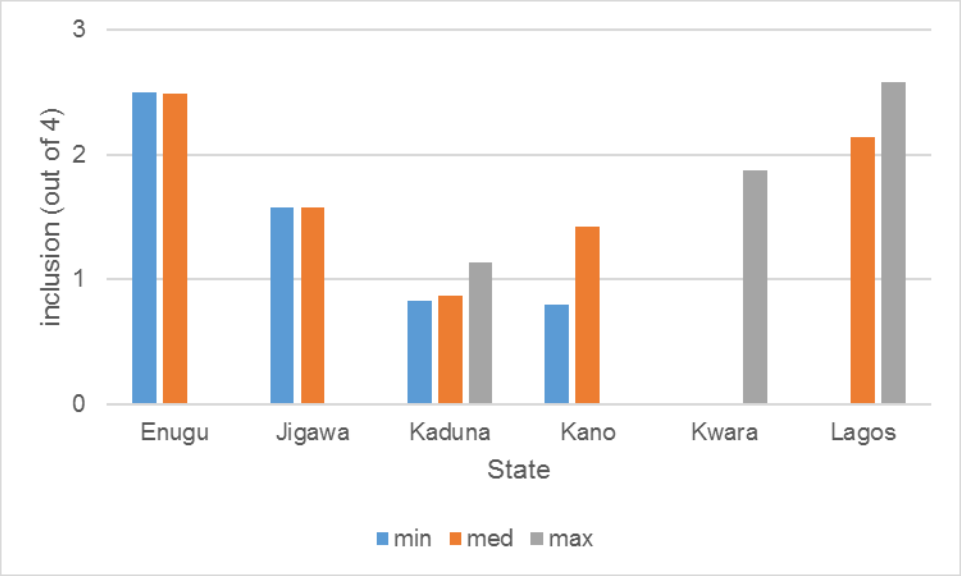
Table 21: School inclusiveness in 2016, by ESSPIN intervention

	Min.	Med.	Max.	Estimated effect of one year of full intervention
(1) Three or more actions on learner attendance (%)	22.8	39.4	22.7	+5.8*
-- Number of actions on learner attendance	1.9	2.2	2	+0.1*
(2) Two or more activities in SDP on access (%)	14	34.6	52.7	+2.5
-- Number of activities on access	0.6	1.3	1.5	+0.1
(3) >50% of teachers use two or more assessment methods (%)	46.9	69.8	74.1	+7.4
(4) >50% of teachers spatially inclusive and >50% are gender inclusive (%)	18.2	24.8	34.3	-0.8
Number of inclusiveness criteria fulfilled (out of four)	1	1.7	1.8	+0.1*
Weighted sum inclusiveness score	81.9	85.5	91	+0.7
School fully met standard (three to four criteria) (%)	7.3	17.4	23.4	+2.3
School partially met standard (two to four criteria) (%)	22.8	53.7	64.9	+6.5*
Additional indicators				
Enrolment increased since last year (%)	62.3	56.3	58.6	+0.1
Change in enrolment since last year	0.1	0.1	0	-0.0

Note. * indicates that the coefficient on years of ESSPIN intervention is statistically significant ($p < .05$) in a regression on the stated indicator, with controls for state.

Comparing differences across states, Kaduna and Kano are doing particularly badly in terms of inclusiveness, with schools meeting on average only around one of the criteria, compared to around 2.5 criteria met in Enugu and Lagos (Figure 5). In Kaduna, Kano, and Lagos, schools that have received more years ESSPIN intervention are markedly more inclusive than schools that have received fewer years of ESSPIN intervention. In Enugu and Jigawa, however, there is little difference in inclusiveness between intervention groups. This suggests that the inclusiveness component of ESSPIN's work may be working better in Kaduna, Kano and Lagos than in Enugu or Jigawa.

Figure 5: Inclusion in 2016, by state and ESSPIN intervention



3.4 SBMCs

To be counted as functioning well, SBMCs are expected to meet regularly, and to work with the community, CBOs, and traditional or religious institutions, to raise awareness about the school and its needs, raise resources, and address exclusion. They are expected to have a women's committee and a children's committee, and to keep financial records, and the chairperson is expected to visit the school regularly. There are nine criteria in the standard for SBMC functionality (Box 4). In most cases, these require evidence to be presented, rather than just accepting the word of the respondent (usually the SBMC chairperson). Thus, they reflect the ability of the SBMC to keep good records of their activities, as well as carrying out the activities themselves. In addition to these, we present statistics on a number of other measures from the Composite Survey SBMC interview (

Table 22). SBMCs are made up of a diverse range of elected community members, with the head teacher of the school acting as SBMC secretary in all cases, and with teacher membership. Most SBMCs had between 12 and 25 members.

In the CS2 report we noted that there were large and significant increases across a number of SBMC-related indicators. In CS3, this trend has continued, with large and significant increases compared to both 2012 and 2014 in seven of the nine criteria. While in 2012 around three-quarters of schools had an SBMC, by 2014 this had reached 95%, and in 2016 all of the schools in our sample have an SBMC. Although nearly all schools had SBMCs in 2014, most of them did not appear to be active. Only 27% had met twice or more in the current school year. This has increased dramatically in 2016, although it is still the case that almost half of the schools have SBMCs that have not met twice or more.

There have been large and significant increases since 2012 and/or 2014 in the proportion of SBMCs that have conducted awareness-raising, networked with CBOs, traditional or religious institutions, or other SBMCs, interacted with LGEAs, and contributed resources for the school. Many more have women's and children's committees, and the proportion of SBMCs whose chairperson has visited the school has increased compared to 2014 (although not compared to 2012). The proportion who could show that they had taken steps to address exclusion (addressed issues which prevent children from attending school or which cause drop-out in the current school year) has risen dramatically, from 27% in 2012 to 68% in 2016.

Overall, schools meet around two additional criteria compared to in 2012, and 56% of schools now meet the overall standard for SBMC functionality, compared to 22% in 2014.

On two key indicators of SBMCs' action on exclusion (see

Box 5: Asking SBMCs about inclusion and exclusion

, below), around 19% have taken action for commonly excluded groups, and 14% have raised issues of children's exclusion. In both cases these represent an increase since 2012, although the increase is only statistically significant for the latter indicator.

Box 4: 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 year[13]:

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); and
9) SBMC chair has visited the school at least three times since the start of the school year (written evidence).

[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 22: SBMC functionality in 2012–2016

	2012 (CS1)	2014 (CS2)	2016 (CS3)	2012–16	2014–16
(1) Two or more meetings this school year (%)	28.7	27.1	52.1	+23.3*	+25.0*
(2) Conducted awareness-raising (%)	35.3	47.5	70.4	+35.1*	+22.9*
(3) Addressed exclusion (%)	26.7	40.1	67.5	+40.8*	+27.4*
(4) Networked with CBOs/institutions/other SBMCs (%)	15	55.6	94	+78.9*	+38.4*
(5) Interacted with LGEA (%)	19.7	21.1	38.4	+18.7*	+17.3*
(6) Has a women's committee (%)	13.1	26.6	41.4	+28.3*	+14.7*
(7) Has a children's committee (%)	19	21	35.8	+16.8*	+14.8*
(8) Contributed resources for school (%)	39	54.5	61.2	+22.2*	+6.6
(9) Chair visited school three or more times (%)	25.2	14.8	24.7	-0.5	+9.9*
Number of SBMC functionality criteria met (/9)	2.3	3.3	4.6	+2.4*	+1.4*
School meets standard for functioning SBMC (%)	21.7	30.9	56.1	+34.4*	+25.1*
Inclusion and drop-out					
(A1) Took action for commonly excluded groups (%)	13.9	23.8	18.7	+4.8	-5.1
(A2) Raised issues of children's exclusion (%)	4.8	19.3	14.2	+9.4*	-5.1
Additional indicators: organising and mobilising resources					
(A7) School has an SBMC (%)	75.7	95.3	100	+24.3*	+4.7*
(A8) Cashbook available (%)	27.2	28.9	64.1	+36.9*	+35.3*
(A9) Requested support from LGEA or SUBEB (%)			65.1	n/a	n/a
(A10) Raised cash to support school improvement (%)	19.7	32.3	22.5	+2.8	-9.8*
(A11) Mobilised non-cash resources (%)	29.1	46.9	54.3	+25.2*	+7.4
(A12) Involved in making SDP (%)		19.6	58.4	n/a	+38.7*
Note. * indicates that the change over time is statistically significant ($p < .05$).					

Schools that have received more years of ESSPIN Output Stream 4 intervention meet significantly more of the SBMC functionality criteria. We estimate that a year of full Output Stream 4 intervention is associated with 0.5 additional criteria being met. Among schools which have had little or no Output Stream 4 intervention, fewer than half meet the overall standard. Schools that have received more years of Output Stream 4 intervention are significantly more likely than those that have received fewer years of intervention to meet regularly, have women's and children's committees, or raise cash to support school improvement. Other differences among the intervention groups are nearly all in the expected direction, but are not statistically significant.

Box 5: Asking SBMCs about inclusion and exclusion

A number of different criteria aim to measure the SBMC's inclusiveness and the actions it has taken on excluded children. These were based on the following questions addressed to the SBMC chairperson. As elsewhere, questions were asked in the local language, with instructions to use a language that the respondent could understand, but not to provide additional explanation or prompts.

Criterion	**Question asked** (with data collector instructions in blue)	Criterion met if…
(2) Conducted awareness-raising	Did the SBMC do anything to raise awareness about the value of education for all boys and girls in the community in the current school year?	Respondent answers yes and can present oral or written evidence
(3) Addressed exclusion	Did the SBMC do anything to address issues which prevent children from attending school or which cause drop-out in the current school year?	Respondent answers yes and can present oral or written evidence
(A1) Took action for commonly excluded groups	Did the SBMC do anything to **support commonly excluded groups** in the **current school year**? You can explain that commonly excluded groups could be orphans, nomadic children, girls, children with disability, ethnic or religious minorities, etc.	Respondent answers yes and can present oral or written evidence
(A2) Raised issues of children's exclusion	Did the SBMC **raise issues of children's exclusion** from school in the community, with the LGEA, or with the state government, in the **current school year**?	Respondent answers yes and can present oral or written evidence
(A3) Raised cash to support vulnerable children	Did the SBMC mobilise any cash to support vulnerable children in the current school year?	Respondent answers yes (no evidence required)
(A4) Monitored drop-out or non-attendance (A5) Communicated with school or community about drop-out	**What actions were taken** to address issues which prevent children from attending school or which cause drop-out in the **current school year**? Do not prompt. This is a multiple response question **– SELECT ALL THAT APPLY** • Monitoring drop-out • Monitoring non-attendance	Respondent answers yes to a previous question (asking whether any action was taken to address these issues) and then provides this information in the follow-up question on what type of action and how many actions were taken. No specific evidence is required

(A6) Number of actions taken to address non-attendance	• Communicating with school about drop-out • Communicating with community about drop-out • Other (specify) • Don't know / refused	

For two key indicators of SBMC action to make the school inclusive – whether the SBMC took action for commonly excluded groups, and whether it raised issues of children's exclusion (see

Box 5: Asking SBMCs about inclusion and exclusion

, above) – around 20% of SBMCs could show evidence that they have met the criteria, and there was a positive but statistically non-significant difference between the schools that have received more years Output Stream 4 intervention and those that have received fewer years of intervention.

Table 23: SBMC functionality in 2016, by ESSPIN intervention

	No intervention	Post-CS1	Pre-CS1	Estimated effect of one year of intervention
(1) Two or more meetings this school year (%)	43.1	72.1	55.9	9.0*
(2) Conducted awareness-raising (%)	67.1	75.9	81.3	4.1
(3) Addressed exclusion (%)	66.3	67.2	82.2	-0.1
(4) Networked with CBOs/institutions/other SBMCs (%)	93.9	93.1	98	1.2
(5) Interacted with LGEA (%)	37.5	42.1	30.2	0.7
(6) Has a women's committee (%)	29.4	65.1	60.9	14.1*
(7) Has a children's committee (%)	22	64.2	53.7	11.3*
(8) Contributed resources for school (%)	60.7	59.5	73.5	4.0
(9) Chair visited school three or more times (%)	21.1	31.7	30.4	0.2
Number of SBMC functionality criteria met (out of nine)	4.2	5.5	5.5	0.5*
School meets standard for functioning SBMC (%)	49.4	70.7	61	5.2
Additional indicators: inclusion and drop-out				
(A1) Took action for commonly excluded groups (%)	**18.6**	**17.4**	**25.1**	**3.3**
(A2) Raised issues of children's exclusion (%)	**13.9**	**13.3**	**21.6**	**2.9**
(A3) Raised cash to support vulnerable children (%)	37.2	30.1	30.6	1.4
(A4) Monitored drop-out or non-attendance (%)	63.9	67.3	68.6	0.5
(A5) Communicated with school or community about drop-out (%)	63.7	66.3	51.1	-3.7
(A6) No. actions taken to address non-attendance	89.7	89.7	91.7	-0.6
Additional indicators: organising and mobilising resources				
(A7) School has an SBMC (%)	100	100	100	n/a
(A8) Cashbook available (%)	63.5	64.5	68.9	6.8*
(A9) Requested support from LGEA or SUBEB (%)	63.9	67.3	68.6	0.5
(A10) Raised cash to support school improvement (%)	20.6	30.7	25	4.1*
(A11) Mobilised non-cash resources (%)	52.5	66	56	4.1
(A12) Involved in making SDP (%)	52.7	72.3	68.6	5.8

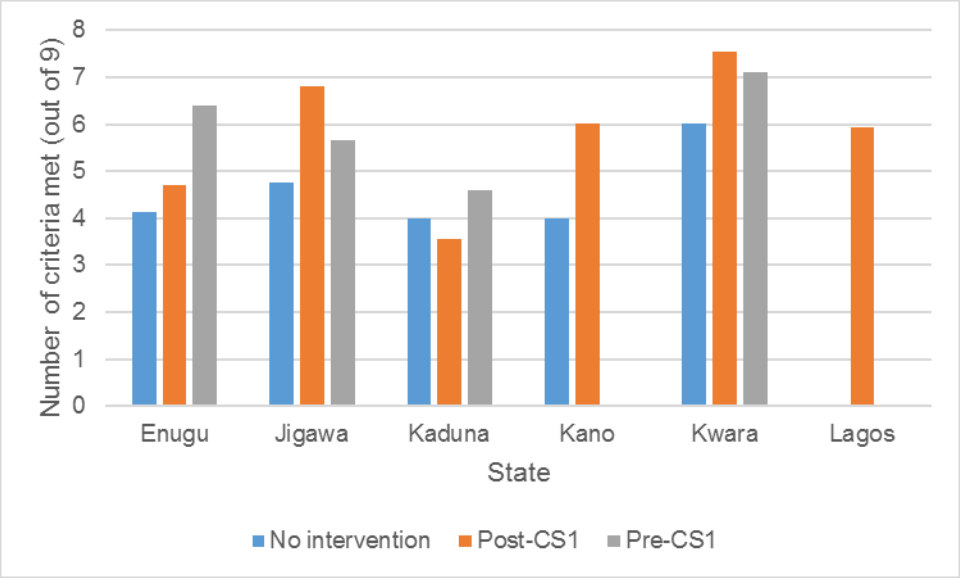
Note. * indicates that the coefficient on years of ESSPIN intervention is statistically significant ($p < .05$) in a regression on the stated indicator, with controls for state.

Comparing the six states (Figure 6), the most functional SBMCs are typically found in Kwara, Lagos, and in the schools that have received more years of intervention in Enugu, Kano, and Jigawa, in each case meeting around 6 of the 9 criteria. Kaduna's schools are lagging behind on SBMC functionality: with even the schools that have received the most years of Output Stream 4 intervention fulfilling fewer than five of the criteria on average. Enugu schools that have received no or relatively little intervention also meet around four to five of the criteria.

The tendency for schools that have received more years of intervention to have better functioning SBMCs is fairly consistent across states, although in some cases 'post-CS1' schools are doing

better than the 'pre-CS1' schools. Pre-CS1 schools started receiving the intervention earlier, and on average have received more intervention, than post-CS1 schools. However, the intervention began with a relatively intense 18-month period of mentoring, which the post-CS1 schools have received more recently. This may explain why post-CS1 schools in some cases have more functional SBMCs in 2016. There may also be differences in the types of schools selected for intervention before and after 2012.

Figure 6: SBMC functionality in 2016, by state and ESSPIN Output Stream 4 intervention



3.4.1 Participation of women and children in SBMCs

As noted above, SBMCs are expected to have women's and children's committees. We also record a number of other measures of the extent to which SBMCs are inclusive of women's and children's concerns. In each case, there are four criteria and an overall standard (

Box 6: Logframe standard for participation of women and children in SBMCs

).

In the CS2 report, we noted that participation of women and children in SBMCs had improved somewhat compared to CS1. This trend appears to have continued, with large and significant increases in several indicators: attendance of women and children in SBMC meetings; whether a female member raised an issue; and whether an issue raised by a female member or child member led to action. In each case, some form of evidence was needed for the SBMC to be counted as meeting the criteria, so these increases may reflect improvements in record-keeping as well as improvements in the underlying action.

Despite these large improvements, on average schools only meet around one of the women's participation criteria and one of the children's participation criteria. Only 23% meet the overall standard for women's participation, and 9% meet the overall standard for children's participation. The results point towards substantial progress compared to both 2012 and 2014, but with severe shortfalls remaining, particularly in regard to the extent to which SBMCs facilitate children's participation.

Box 6: Logframe standard for participation of women and children in SBMCs

The school must meet at least three of the four criteria listed below in order to meet the SBMC **women's participation** standard:

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); and
4. at least one SBMC women's committee meeting took place.[14]

The school must meet at least three of the four criteria listed below in order to meet the SBMC **children's participation** standard:

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); and
4. at least one SBMC children's committee meeting took place and committee has a trained facilitator.[15]

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

[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 24: Participation of women and children in SBMCs in 2012–2016

				Change in average over time	
	2012 (CS1)	2014 (CS2)	2016 (CS3)	2012–16	2014–16
Participation of women in SBMC					
(1) At least one woman attended two or more meetings (%)	19.5	17.4	29.6	+10.1*	+12.2*
(2) Female member raised an issue (%)	26.8	31.9	54.9	+28.1*	+23.1*
(3) Issue raised by female member led to action (%)	28.3	14.5	31.8	+3.5	+17.3*
(4) Women's committee met (%)	7.8	27.1	31.6	+23.8*	+4.5
No. criteria met (out of four)	0.6	0.9	1.3	+0.6*	+0.4*
Meets standard (three out of four criteria) (%)	15.4	15.5	23.2	+7.8	+7.7*
Participation of children in SBMC					
(1) At least one child attended two or more meetings (%)	11.8	8.8	19.6	+7.8*	+10.8*
(2) A child raised an issue (%)	13.6	20.6	22.1	+8.5*	+1.5
(3) Issue raised by child led to action (%)	11.5	7.3	22.4	+10.9*	+15.1*
(4) Children's committee met (%)	2.4	14.3	19.2	+16.9*	+4.9
No. criteria met (out of four)	0.3	0.5	0.8	+0.5*	+0.2*
Meets standard (three out of four criteria) (%)	5.7	6.2	9.7	+4.0	+3.5*

Note. * indicates that the change over time is statistically significant ($p < .05$).

Comparing schools with different levels of ESSPIN Output Stream 4 intervention reveals that in all of our indicators of women's and children's participation schools that have received more years of intervention are doing better. The differences are large and statistically significant. Only 11% of schools that have received no intervention meet the overall standard for women's participation, compared to 43% of schools that have received the most intervention (pre-CS1). For the children's standard, there are similarly large differences across intervention groups, although it is notable that even in the schools that have received the most intervention, only one in four schools meet the overall standard. We estimate that around two to three years of ESSPIN Output Stream 4 intervention are needed for schools to meet an additional criterion for women's and children's participation.

Table 25: Participation of women and children in SBMCs in 2016, by ESSPIN Output 4 intervention

	No intervention	Pre-CS1	Post-CS1	Estimated effect of one year of intervention
Participation of women in SBMC				
(1) At least one woman attended two or more meetings (%)	18.1	42	53.6	8.9*
(2) Female member raised an issue (%)	44.7	71.7	66.1	6.8*
(3) Issue raised by female member led to action (%)	23.7	46.4	47.6	6.4*
(4) Women's committee met (%)	20.6	51.8	53	11.2*
No. criteria met (out of four)	0.8	1.9	2.1	0.4*
Meets standard (three out of four criteria) (%)	11.2	42.5	47.2	8.5*
Participation of children in SBMC				
(1) At least one child attended two or more meetings (%)	16.5	32.7	23.9	4.2*
(2) A child raised an issue (%)	14	39.2	35.8	8.3*
(3) Issue raised by child led to action (%)	19	30.1	28.7	3.4
(4) Children's committee met (%)	9	40.8	38.7	9.0*
No. criteria met (out of four)	0.5	1.3	1.2	0.3*
Meets standard (three out of four criteria) (%)	4.4	25.2	19	5.2*

Note. * indicates that the coefficient on years of ESSPIN intervention is statistically significant ($p < .05$) in a regression on the stated indicator, with controls for state.

3.5 Summary: School management and head teachers

In 2016, 18% of schools meet the standard on head teacher effectiveness; 19% on school development planning; 11% on inclusion; and 44% on functional SBMCs. It is not clear if head teacher effectiveness has improved, overall, since 2012, but school development planning and SBMC functionality have improved markedly. There have been improvements in a number of important criteria such as head teachers carrying out lesson observations and professional development meetings. Fewer schools meet the overall inclusiveness standard in 2016 than in 2012, yet on a more nuanced measure schools have become significantly more inclusive. Women's and children's participation in SBMCs have also improved over time.

On average around 80% of teachers were present on the day of the survey visit, according to the head teacher's own records, and around 70% of classrooms were observed as having both teacher and learners present at the expected time in the morning. There has been little change in the proportion of teachers and learners present on time since 2012. This suggests that schools continue to have serious problems with teacher and learner attendance at the start of the school day.

Schools which have had more years of ESSPIN intervention have more effective head teachers in 2016, are better at school development planning, are more inclusive, and are more likely to have well-functioning and inclusive SBMCs in which women and children participate. With a few exceptions, the differences between intervention groups are consistent across the six states, which

suggests that they cannot wholly be attributed to differences between the states. The estimated effect of a year of full ESSPIN intervention on the proportion of schools that meet each standard ranges from 5 percentage points for head teacher effectiveness to 9 percentage points for school development planning.

4 Teachers

ESSPIN's interventions include teacher training on teaching skills, including the use of teaching aids, participation and giving praise, and techniques for classroom organisation. The training also includes basic literacy and numeracy training. This section examines the changes in teacher competence over time and across different intervention groups. In this chapter, we first look at how teacher competence has changed according to the teacher competence logframe indicator that combines the various aspects on which teachers receive training. We then take a more in-depth look at teachers' performance in the literacy and numeracy content knowledge tests. Finally, we examine teacher motivation, which may be an important factor affecting learning outcomes.

4.1 Teacher competence

The ESSPIN logframe sets four criteria for judging the competence of teachers (Box 7). A teacher who teaches English or mathematics 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 mathematics curriculum) and defined as competent if they meet two of the remaining three criteria.

For CS2 and CS3, a stricter version of the competence indicator was developed. The criterion on using at least one teaching aid during the lesson observation is changed to exclude reading from, writing on, or having learners copy from, the blackboard. The blackboard is still counted as a teaching aid if teachers use it more creatively, bringing learners to the front of the class to carry out activities using the blackboard. The more conventional use of the blackboard for writing out and copying text, while not necessarily a bad practice, is not counted as using a teaching aid. In addition, 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 7: 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 mathematics. 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, writing on, or having learners 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.

Teachers are showing significant improvements on two of the teacher competence indicators over time. In 2016, teachers are more likely to be using more praise than reprimands during their lessons than they were in 2012. Teachers are also significantly more likely to use at least one teaching aid during their lesson, with almost all teachers now doing so (98.6%), compared to 88% of teachers in 2012. However, reading from, writing on, or having learners copy from, the blackboard is the only teaching aid used by many teachers (30%), who fail to make use of more innovative teaching aids.

Teachers' knowledge of the English or mathematics curriculum benchmarks is significantly lower than in 2012, but has improved compared to 2014. As noted in the CS2 report, there were some inconsistencies in how this question was administered in CS1 compared to the other rounds of the survey, which may explain the initial drop in curriculum knowledge.[16]

The proportion of teachers who assign at least two individual or group tasks has decreased dramatically since 2012, from 56% to 29%, and has also decreased since 2014. The high proportion of teachers apparently assigning individual and group tasks in CS1 may include some measurement error. In CS3, for example, it was made clear that the definition of an individual task should not include learners copying from the blackboard – an activity that might be interpreted as an individual task without this clear guidance. Still, the low proportion of teachers who are assigning these tasks in 2016, despite the increasingly widespread availability of lesson plans that promote the use of group work, leaves much room for improvement.

Performance in the content knowledge tests was poor, and worsened significantly between 2014 and 2016. Only 30% of teachers were able to score at least 50% in both the English literacy and numeracy test, compared to 37% in 2014. Further analyses of the content knowledge tests are presented in the section that follows.

Overall, teacher competence has improved significantly compared to 2014, but has not changed compared to 2012. The proportion of teachers meeting the stricter version of the competence standard has not changed over time. However, the total number of teachers has increased during this period, meaning the absolute number reaching the standard (as well as the absolute number falling below the standard) has increased since 2012. About 67% of teachers meet the original version of the competence standard in 2016, but only 21% meet the stricter version of the standard. This is being driven by the very poor performance in the content knowledge tests.

[16] 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. Data were captured digitally and, wherever possible, processed automatically using streamlined, standardised and quality assured processes, applying lessons learned from the experimental and developmental aspects of CS1.

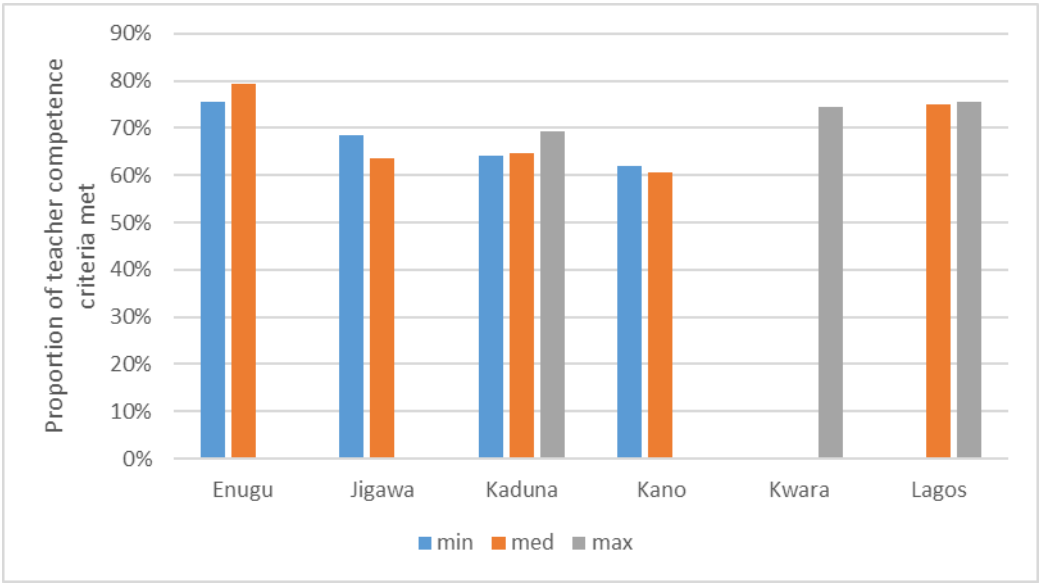
Table 26: Teacher competence in CS1, CS2 and CS3

	2012 (CS1)	2014 (CS2)	2016 (CS3)	Change: 2012 vs. 2016	Change: 2014 vs. 2016
(1) Knowledge of Eng./mathematics curriculum (%)	56.9	34.9	44.1	-12.8*	+9.2*
(2a) Use of one or more teaching aid (%)	88.1	94.8	98.6	+10.5*	+3.8*
(2b) Use of one or more teaching aid, excl. blackboard (%)		67.4	70.3		+2.9
(3) Praise more than reprimand (%)	70.4	80.4	87.5	+17.1*	+7.1*
(4) Assigns two or more ind./group tasks (%)	56.1	34.2	28.7	-27.4*	-5.5*
(5) Passes English and mathematics test		36.7	29.9		-6.7*
Teacher competence score (% of criteria fulfilled)	69.8	63.0	66.3	-3.5*	+3.3*
Teacher competence standard fulfilled (%)	69.7	57.4	66.8	-2.9	+9.4*
Teacher competence score (% of criteria fulfilled, strict version)		55.4	55.9	n/a	+0.5
Teacher competence standard fulfilled (strict) (%)		21.0	20.5	n/a	-0.5

Note. * indicates that the change over time is statistically significant ($p < .05$). 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%.

Comparing differences across states, teacher competence scores are somewhat lower in Jigawa, Kaduna and Kano, compared to the other states (Figure 7). In Enugu and Kaduna, teachers in schools that have received more years of ESSPIN intervention are slightly more competent than teachers in schools that have received fewer years of ESSPIN intervention. In Jigawa, however, teachers in schools that have received more years of ESSPIN intervention seem to be performing slightly worse, while in Kano and Lagos there is little difference in teacher competence between intervention groups.

Figure 7: Teacher competence score in 2016, by state and ESSPIN intervention



Teacher competence also varies by the gender of the teacher, with female teachers performing better on all teacher competence indicators, except the use of a teaching aid. Strikingly, the

proportion of female teachers who pass the content knowledge tests is double that of male teachers.

This may in part be driven by the distribution of teachers across states: there are more female teachers in those states that have higher teacher competence scores (Enugu, Lagos and Kwara). To explore this we examine the proportion of teachers meeting the overall competence standard by state and gender. Although the gap narrows, we still find within each state that a higher proportion of female than male teachers reaches the overall competence standard. However, this difference is only significant in Enugu and Kaduna.

Table 27: Teacher competence in 2016 by gender

	Male	Female	Significant difference
(1) Knowledge of Eng./mathematics curriculum (%)	31.7	55.8	F
(2a) Use of one or more teaching aid (%)	98.3	98.8	
(2b) Use of one or more teaching, aid excl. blackboard (%)	62.4	79.9	F
(3) Praise more than reprimands (%)	84.1	91.6	F
(4) Assigns two or more ind./group tasks (%)	23.8	35.5	F
(5) Passes English and mathematics test	24.8	53.2	F
Teacher competence score (% of criteria fulfilled)	2.5	2.9	F
Teacher competence standard fulfilled (%)	61.2	73.3	F
Teacher competence score (% of criteria fulfilled, strict version)	2.6	3.3	F
Teacher competence standard fulfilled (strict) (%)	23.8	33.7	F
Strict Teacher competence standard (CS2 version) by gender groups and states			
States	Male	Female	Significance
Enugu	12.3	24.6	F
Jigawa	25.5	34.1	
Kaduna	25.3	41.4	F
Kano	23.6	29.9	
Kwara	11.9	15.1	
Lagos	39.7	49.3	

Note. F = significant difference in favour of females; M = significant difference in favour of males ($p < .05$)

We also examine how teachers who reported having received ESSPIN training by CS3 performed compared to those that did not report having received ESSPIN training. For this analysis, it should be noted that there are some pre-existing differences between the teachers selected for training and those who were not selected. The two groups of teachers have similar profiles in terms of age and sex, but the teachers trained by ESSPIN are slightly more likely to have a National Certificate of Education (NCE) or higher qualification. ESSPIN-trained teachers are much more likely to teach English or mathematics than other teachers: 88% of teachers trained by ESSPIN in the CS3 sample taught English and/or mathematics (often alongside other subjects), compared to 67% of non-trained teachers.

Across most indicators, ESSPIN-trained teachers performed slightly better than non-ESSPIN-trained teachers. Encouragingly, 78% of ESSPIN-trained teachers were using teaching aids interactively (not just reading from or writing on the blackboard), compared to only 63% of non-ESSPIN-trained teachers. In addition, a significantly higher proportion of teachers who had received training were assigning at least two individual or group tasks. ESSPIN-trained teachers

also spent significantly more time of their lesson explaining, and less time chanting; and were more likely to test their learners' knowledge. Although we cannot attribute these results to ESSPIN's intervention, ESSPIN-trained teachers appear to have a better grasp of teaching methods that enhance learning.

Overall, there is no difference in the proportion of teachers who meet the teacher competence standard, but ESSPIN-trained teachers score significantly higher on the teacher competence score. They are also more likely to meet the stricter competence standard.

Table 28: Teacher competence in CS3 by intervention group

Intervention group	Non-ESSPIN-trained	ESSPIN-trained	Difference
(1) Knowledge of Eng./mathematics curriculum (%)	44.1	44.0	-0.0
(2a) Use of one or more teaching aid (%)	98.3	98.9	+0.7
(2b) Use of one or more teaching aid, excl. blackboard (%)	62.9	77.6	+14.7*
(3) Praise more than reprimands (%)	86.3	88.7	+2.4
(4) Assigns two or more ind./group tasks (%)	22.1	35.3	+13.2*
(5) Passes English and mathematics test	28.8	31.1	+2.4
Teacher competence score (% of criteria fulfilled)	65.0	67.6	+2.5*
Teacher competence standard fulfilled (%)	68.6	65.1	-3.5
Teacher competence score (% of criteria fulfilled, strict version)	52.4	59.3	+6.8*
Teacher competence standard fulfilled (strict) (%)	18.4	22.6	+4.3*
Additional indicators:			
Proportion of time spent -- explaining (%)	41.4	46.4	+5.0*
-- instructing / presenting / dictating (%)	15.0	17.5	+2.5*
-- chanting (%)	8.5	4.6	-3.9*
-- closed question / response (%)	6.0	5.4	-0.6
-- open question / response (%)	4.8	5.0	+0.2
Proportion of time spent speaking English (%)	32.6	37.3	+4.7
Teacher summarised the lesson (%)	65.2	68.9	+3.7
Teacher revisited the lesson's objectives (%)	34.1	36.1	+2.0
Teacher gave learners homework (%)	22.3	27.4	+5.1
Teacher tested learners' knowledge (%)	51.0	56.9	+5.9*
Teacher marked learners' written work (%)	19.6	20.3	+0.7

Note. * indicates that the difference between ESSPIN-trained and other teachers is statistically significant ($p < .05$). Knowledge of English/mathematics curriculum questions are only asked of teachers who teach English and/or mathematics.

4.2 Findings from the teacher content knowledge tests

The findings above suggest that teachers' content knowledge has worsened between 2014 and 2016, but that it is better among teachers trained through ESSPIN than among those who have not received ESSPIN training. Percentage scores in the teacher content knowledge tests provide a rough indication of teachers' test performance, but analysis using IRT takes account of the difficulty of items to provide an interval scale for more robust comparison of performance levels, and can

also be interpreted more readily in terms of learning benchmarks (see Allen, 2016d). The teachers' results can be divided into four performance bands in literacy and five performance bands in numeracy. Review of the items that teachers in each band can mostly answer correctly then provides descriptors for each band (Table 29). For example, a teacher in Band 2 for literacy is one who shows knowledge of some basic phonics, can write a simple sentence, and can perform basic comprehension of a passage, as well as satisfying the easier items – testing limited comprehension of simple passages, basic nouns and verbs – associated with a teacher in Band 1. The teacher in Band 2 cannot typically correctly answer the harder items associated with Bands 3 or 4, such as identifying simple antonyms.

Table 29: Band descriptors based on IRT analysis

Band	Literacy	Numeracy
5		Understands conversion of fractions to decimals, and place values in decimals
4	Creates several sentences, shows knowledge of phonics, punctuation, formal letter layout, suffixes and alphabetical order	Understands ideas of area, nets, pictograms and rounding
3	Past/present of verbs, completes a sentence, extracts basic information from a passage, identifies simple antonyms, forms plurals	Understands basic sets, use of the number line to represent sums, conversion of units of time and mass, can complete word problems involving division
2	Shows knowledge of some basic phonics, writes a simple sentence, basic comprehension of a passage	Simple division, word problems involving addition, signs for arithmetic operations, integer comparisons and integer place values
1	Limited comprehension of simple passages, basic nouns and verbs	Simple addition with carrying over, simple subtraction, identifying a fraction, counting, simple regular shapes

Within the literacy and numeracy tests, items can be grouped according to specific sub-domains of learning: reading, writing and grammar within literacy, and number concepts and calculation within numeracy.

Teachers' scaled scores in both English and mathematics have declined significantly since 2014, by about 0.3 of a standard deviation in English and 0.2 of a standard deviation in mathematics (

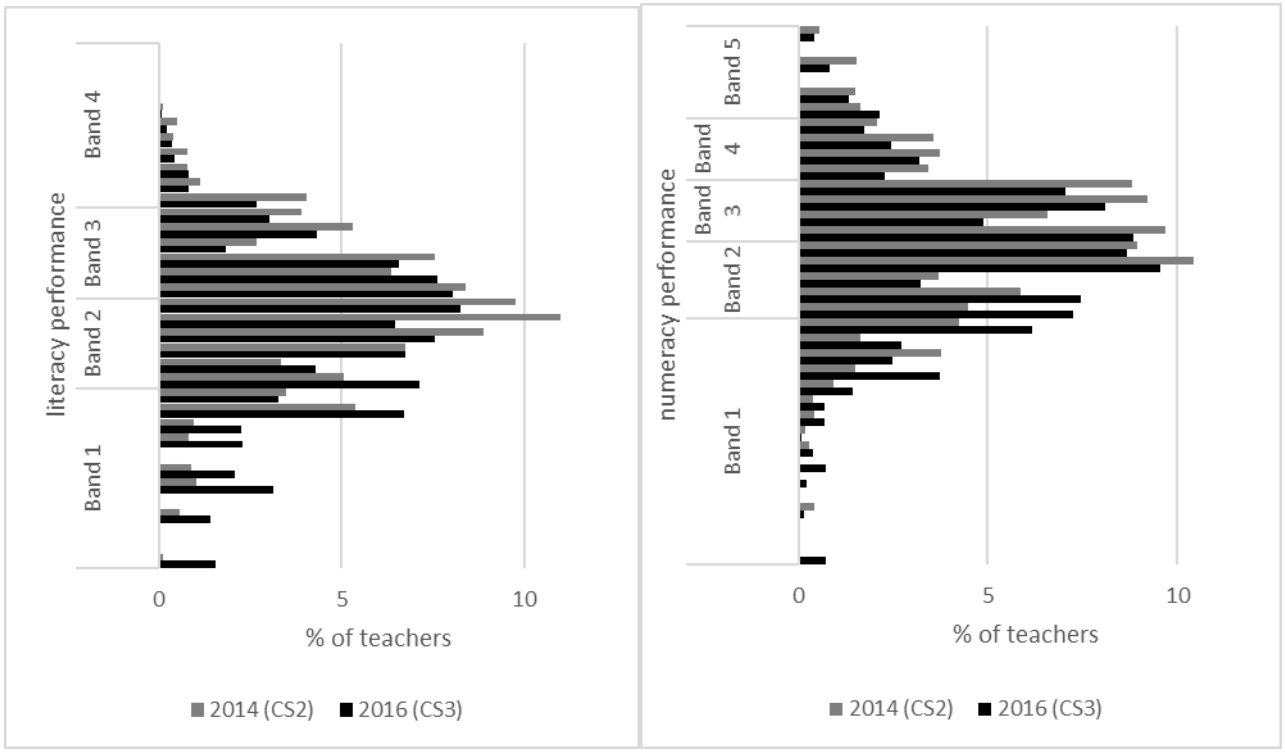
Table 30, Figure 8). The proportion of teachers in the lowest band increased from 13% to 23% in English, and from 12% to 16% in mathematics; these increases were matched by declines in each of the higher performance bands. The declines were similar in magnitude across the different learning sub-domains (reading, writing, grammar, number concepts and calculation). The decline is particularly puzzling because the same teachers tested twice have, on average, worse scores in 2016 than in 2014 – so it cannot be explained in terms of changing composition of the teaching workforce.

Table 30: Teachers' test scores (IRT analysis) in 2014 and 2016

	2014 (CS2)	2016 (CS3)	Change on average, 2014–16
English IRT scale score (mean 500, s.d. 100)	486	457	-29.4*
English Band 1 (%)	13	23	+9.5*
English Band 2 (%)	40	36	-4.0
English Band 3 (%)	37	34	-3.1
English Band 4 (%)	9	7	-2.3*
Reading (English sub-scale, mean 500, s.d. 100)	488	459	-28.6*
Writing (English sub-scale, mean 500, s.d. 100)	487	464	-23.3*
Grammar (maths sub-scale, mean 500, s.d. 100)	486	461	-25.6*
Mathematics IRT scale score (mean 500, s.d. 100)	484	462	-22.7*
Mathematics Band 1 (%)	12	16	+4.7*
Mathematics Band 2 (%)	31	36	+4.5
Mathematics Band 3 (%)	35	30	-5.1*
Mathematics Band 4 (%)	17	13	-3.5*
Mathematics Band 5 (%)	6	5	-0.7
Number concepts (maths sub-scale, mean 500, s.d. 100)	485	464	-20.6*
Calculation (maths sub-scale, mean 500, s.d. 100)	486	463	-23.2*

Note. * indicates that the change over time is statistically significant ($p < .05$). The scaled scores are standardised with mean in the combined CS2 and CS3 sample of 500. The averages shown in the table are below 500 because of the effect of applying sample weights.

Figure 8: Distribution of teacher English and mathematics scores in 2014 and 2016



Teachers who had received ESSPIN training had significantly higher test scores than those who had not, in both English and mathematics (

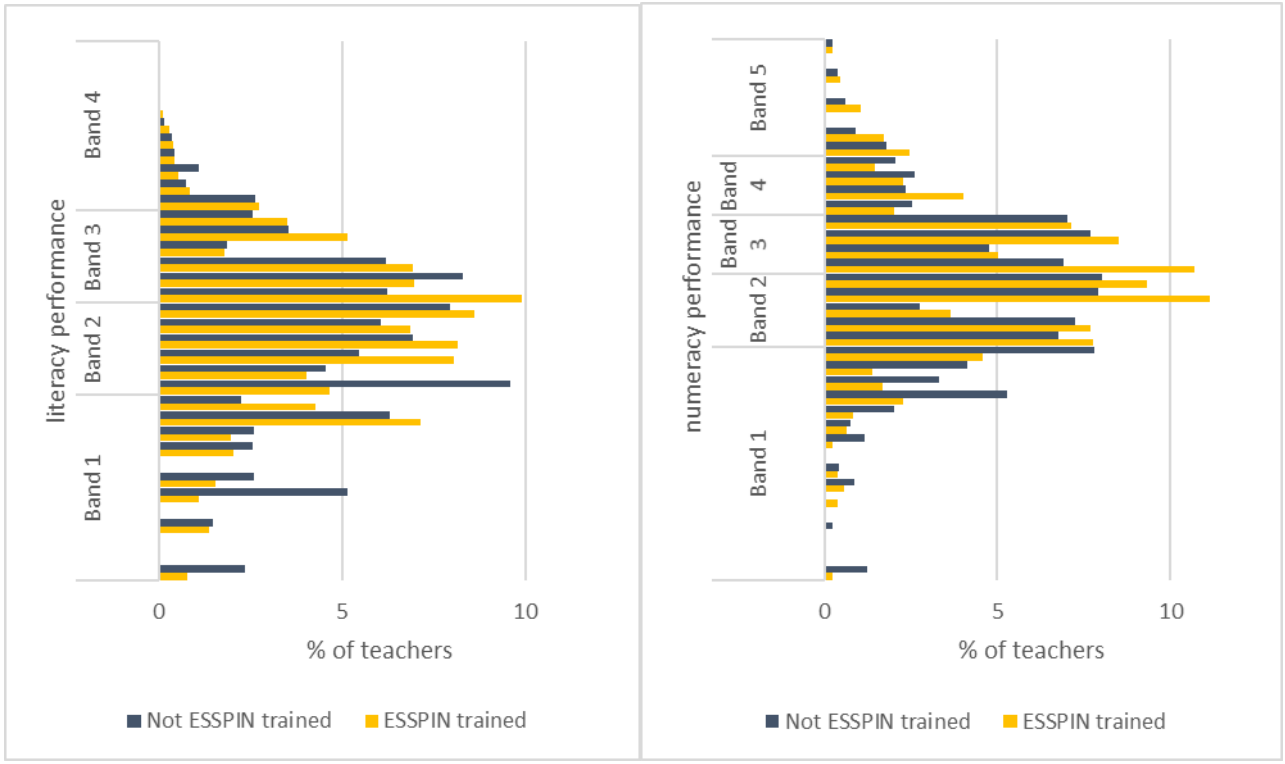
Table 31, Figure 9). The difference was around 0.2 standard deviations in English and 0.3 standard deviations in mathematics, and there were particularly marked differences in reading scores and number concepts. In English, 20% of ESSPIN-trained teachers fell into the lowest performance band, compared to 25% of non-ESSPIN-trained teachers, and in mathematics roughly twice as many non-ESSPIN-trained teachers were in the lowest performance band as ESSPIN-trained teachers (22% vs. 11%). As noted above, ESSPIN-trained teachers are much more likely to be English or mathematics teachers, so this difference in test scores is likely partly to reflect their prior skills and training. However, the difference of 0.2–0.3 standard deviations is large in comparison to typical effect sizes for educational programmes.

Table 31: Teachers' test scores (IRT analysis) by ESSPIN training

	Non-ESSPIN-trained	ESSPIN-trained	Difference in means
English IRT scale score (mean 500, s.d. 100)	446	467	+21.2*
English Band 1 (%)	25.3	20.1	-5.2*
English Band 2 (%)	35.4	36.7	+1.3
English Band 3 (%)	32.4	36.0	+3.6
English Band 4 (%)	6.9	7.2	+0.3
Reading (English sub-scale, mean 500, s.d. 100)	446	472	+26.2*
Writing (English sub-scale, mean 500, s.d. 100)	459	468	+8.9
Grammar (maths sub-scale, mean 500, s.d. 100)	454	467	+13.4*
Mathematics IRT scale score (mean 500, s.d. 100)	448	475	+26.3*
Mathematics Band 1 (%)	21.6	11.3	-10.3*
Mathematics Band 2 (%)	34.4	37.1	+2.7
Mathematics Band 3 (%)	27.1	32.0	+4.9
Mathematics Band 4 (%)	12.9	13.6	+0.7
Mathematics Band 5 (%)	3.9	6.0	+2.0*
Number concepts (maths sub-scale, mean 500, s.d. 100)	451	478	+27.0*
Calculation (maths sub-scale, mean 500, s.d. 100)	453	473	+20.2*

Note. * indicates that the difference between ESSPIN-trained and other teachers is statistically significant ($p < .05$).

Figure 9: Distribution of teacher English and mathematics scores in 2016, by ESSPIN training



There are large differences across the six states in teachers' performances in the tests. In Kano and Jigawa, more than one-third of teachers are in the lowest performance band for English, and around one-quarter are in the lowest performance band for mathematics. Very few teachers reach the highest performance bands in these two states. By contrast, in Lagos, nearly 40% of teachers are in the top performance band, and almost none in the lowest, for both English and mathematics.

Figure 10: % of teachers in each English performance band, by state, 2016

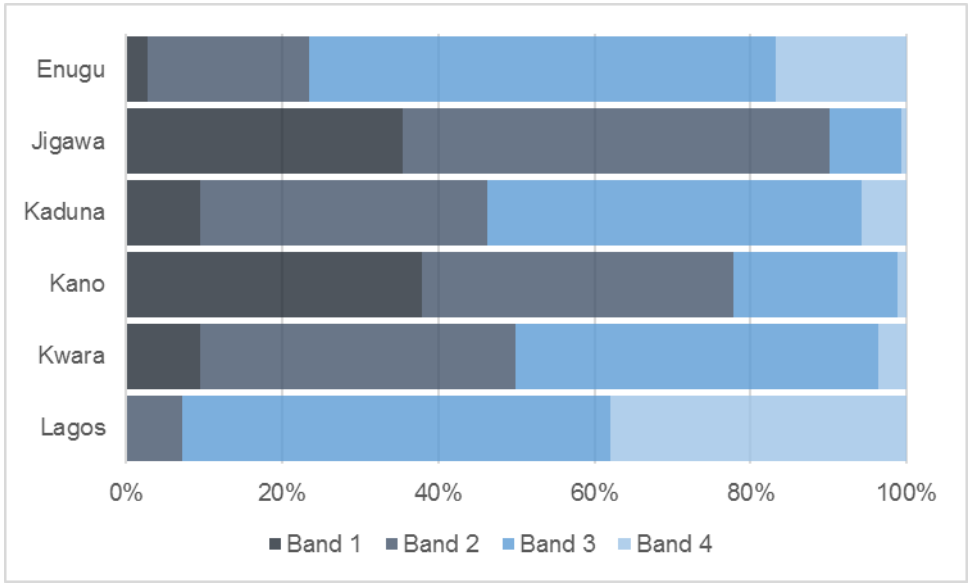
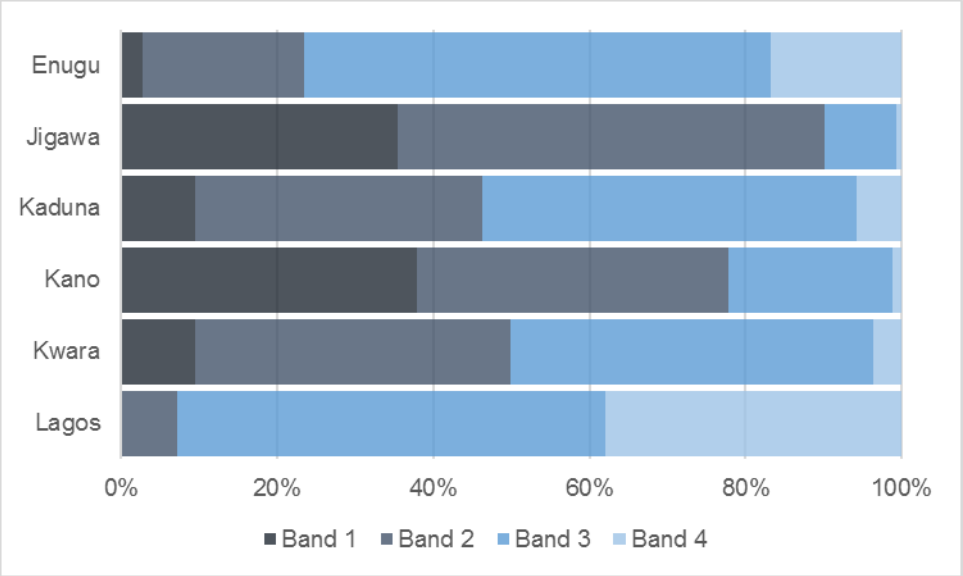


Figure 11: % of teachers in each mathematics performance band, by state, 2016



4.3 Teacher motivation

Teacher motivation is likely to be important in regard to the effort that teachers put into their jobs, absenteeism, and the sustained effectiveness of a training intervention. For example, as teachers acquire new skills through the teacher training, their motivation may increase as they feel more effective.

For this round of the survey (CS3) we included a measure of teacher motivation and teacher interaction using a scale that has been developed for the Nigerian context, and that has been used and tested in two previous school-based surveys. We define teacher motivation as the propensity of teachers to start and maintain behaviours that are directed towards fulfilling their professional goals, and in particular towards achieving better learning outcomes for the school's learners (Cameron, 2015b). Many existing instruments designed to measure teacher motivation focus exclusively on 'efficacy' – the extent to which teachers see themselves as able to influence their pupils' learning outcomes – which can also be seen as the 'can do' aspect of motivation (Bennell and Akyeampong, 2007). We wished to go beyond this to include measures relating more closely to teachers' willingness to work hard, commitment, effort and enjoyment, which might together be labelled as 'will do' aspects of motivation. A motivation scale was therefore developed with items relating to both 'will do' and 'can do' aspects of motivation.

The motivation scale we developed was incorporated into the teacher interviews. Teachers were asked to what extent they agreed ('strongly disagree', 'disagree', 'agree', 'strongly agree') with a series of statements that measure different aspects of motivation. The scale consists of three sub-scales of teacher motivation (satisfaction, skills and engagement) and one scale of teacher–teacher interaction (collegiality) (Table 32). The three sub-scales of teacher motivation were combined into a composite motivation measure by calculating the mean of the three sub-scales[17]. The teacher motivation scale was also analysed using IRT (see Allen, 2016b).

[17] The three sub-scales were also combined into a composite measure using partially non-compensatory methods. These produced composite measures which were very highly correlated with the simple mean composite. Collegiality is not included in the composite motivation measure. We see collegiality as conceptually distinct from, but likely to interact in various ways with, motivation.

Table 32: Teacher motivation and interaction scale and sub-scales

Scale	Description	Example of items
Collegiality	How I see the extent of commitment and collaboration among my colleagues ('teacher–teacher interaction')	• All of the teachers in my school trust each other • All teachers at this school are highly committed to their job
Satisfaction	The value I place on my role as a teacher ('interest and enjoyment')	• I always enjoy teaching very much • I like to spend a lot of energy to make my classes interesting
Skills	The perception I have of my competencies and skills as a teacher ('self-efficacy')	• I believe I know how to teach well • I believe I have the skills needed to encourage my learners to always work hard
Engagement	How engaged and committed I feel I am to my work as a teacher ('pressure/tension')	• It is difficult to manage learners in my classroom • Teaching is very tiring
Composite measure (mean of satisfaction, skills and engagement)		

Table 33 shows the levels of motivation among teachers, as reported during CS3, comparing ESSPIN-trained to non-ESSPIN-trained teachers. ESSPIN-trained teachers are significantly more motivated than non-ESSPIN-trained teachers. ESSPIN-trained teachers report feeling more engaged than non-ESSPIN-trained teachers, as well as slightly more satisfied. Collegiality does not differ by ESSPIN training.

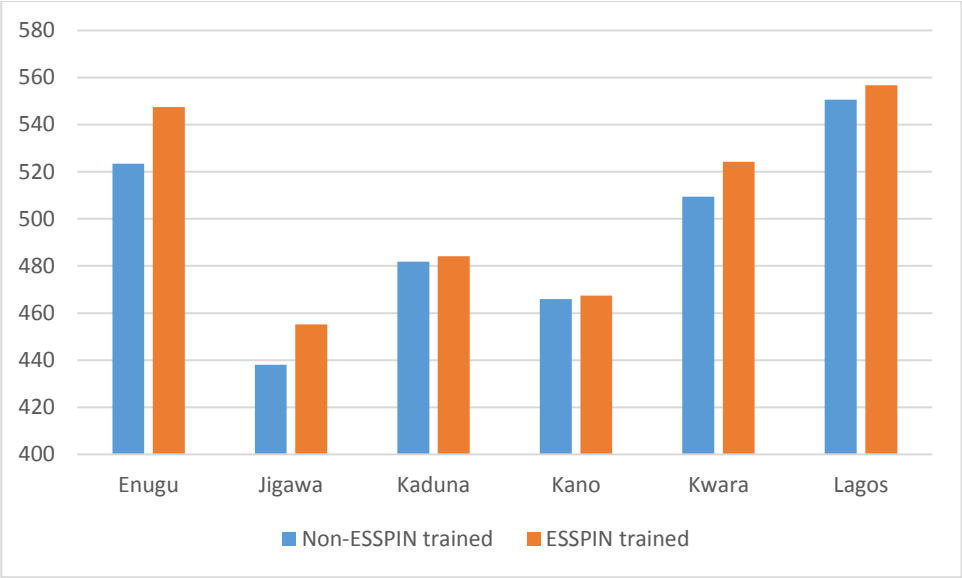
Table 33: Teacher motivation and interaction by ESSPIN training

	Non-ESSPIN-trained	ESSPIN-trained	Difference in means
Collegiality	498.6	497.8	-0.7
Satisfaction	484.1	492.0	+7.9
Skills	490.0	492.9	+2.9
Engagement	469.8	488.5	+18.7*
Composite motivation measure	482.4	491.6	+9.2*

Note: All scores are normalised to have an average (mean) of 500 and a standard deviation of 100. * indicates that the difference between ESSPIN-trained and other teachers is statistically significant ($p < .05$).

There are large differences across the six states in teachers' motivation (Figure 12). Teacher motivation is higher in Enugu, Kwara and Lagos compared to the other states. In Enugu, Jigawa and Kwara, ESSPIN-trained teachers are more motivated than non-ESSPIN-trained teachers. In the other states, levels of motivation do not differ by ESSPIN training.

Figure 12: Teacher motivation composite measure by state and ESSPIN training



4.4 Summary and discussion

Over time, teacher competence has been improving on the original measure, but not on the stricter measure. Teachers have improved on their use of teaching aids and praise during their lessons. On the other hand, the propensity of teachers to assign individual and group tasks remains low, there are persisting gaps in the curriculum knowledge of many English and mathematics teachers, and while most teachers use teaching aids, far fewer teachers use them interactively.

These measures of teacher competence are likely to reflect the large gaps in teachers' content knowledge. Performance in the teacher content knowledge tests has been worsening over time. In 2016, only 7% of teachers perform in the highest performance band (Band 4) in a primary grade-level literacy test, while 19% perform in the two highest performance bands (Bands 4 and 5) in the numeracy test. At these levels of performance, teachers are likely to lack the skills and understanding to effectively perform some of the competency tasks, such as assigning individual or group tasks during their lessons or demonstrating an understanding of the curriculum benchmarks. The deterioration in test performance is a matter that warrants further investigation. One possibility is that some states have been recruiting less qualified or less competent teachers to cope with enrolment increases. It is also possible that teachers are not motivated to complete the test, or that the administration of the test in schools (as opposed to in testing centres in 2014) has in some way affected the results. These factors deserve exploration as part of piloting and validity testing for future teacher tests.

ESSPIN-trained teachers are more motivated, perform better in the content knowledge tests, and are more competent than non-ESSPIN-trained teachers. In particular, we observe ESSPIN-trained teachers performing better on aspects of teaching that require a better understanding of how to enhance learning: for example using teaching aids more interactively and assigning more individual or group tasks during their lessons. While there may be alternative explanations for these findings, this set of results suggests that ESSPIN training may be having a positive effect on teachers' content knowledge and competence.

5 Trends in school quality

For a broader sense of school quality, how it differs between schools with different levels of ESSPIN intervention, and how it has changed over time, it is useful to define an overall measure of school quality. We do this using the standard developed as part of ESSPIN's logframe. This is a combination of the standards discussed above 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 8). 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. A school that meets all of the criteria under all of the standards will get 100%, while a school that meets none of the criteria will get 0%. The original version of these indicators, used in CS1, did not take into account teachers' content knowledge. For CS2 and CS3, we also present a 'strict' version of the standard, which takes into account results in the teacher content knowledge tests (see Section 4.2 above).

In this section, we first examine changes over time in our measures of overall school quality, and then examine the differences between schools with different levels of ESSPIN intervention. We then use a difference in differences approach and regression analysis to help establish whether the better quality found in schools that have had ESSPIN intervention can be causally linked to the intervention, or might instead be due to other confounding factors such as differences between the states.

Box 8: 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; and
4) SBMC functionality standard.

The version of this standard used in CS1 did not include teacher content knowledge tests. For CS2, we introduced a second, stricter version of the standard, in which teachers must get above 50% in English and mathematics tests to be classed as competent (see Section 4.1 and Box 7 above).

School quality has increased significantly between 2014 and 2016, with a rise of about 11 percentage points in the school quality score (

Table 34). The proportion of schools meeting the overall quality standard as used in CS1 has increased dramatically, from around 4% in 2012 to 20% in 2016. In terms of the ‘stricter’ standard introduced in CS2, there has also been a large increase in the score between 2014 and 2016. However, even in 2016 only 5.6% of schools meet the strict standard, because in most cases teachers score below 50% in the English and mathematics tests.

Table 34: School quality in 2012–2016

	2012 (CS1)	2014 (CS2)	2016 (CS3)	Change 2012–16	Change 2014–16
Quality score (%)	38.3	40.4	51.4	+13.1*	+11.0*
School meets quality standard (%)	3.9	8.3	20.2	+16.2*	+11.8*
Quality score (strict version) (%)		39.1	48.8	n/a	+9.7*
School meets quality standard (CS2 version) (%)		4.6	5.6	n/a	+1.0

Note. * indicates that the change over time is statistically significant ($p < .05$).

By any measure, school quality is significantly better in schools that have received more years of ESSPIN intervention than in those that have received little or no ESSPIN intervention (Table 35). One year of full Output Stream 3 intervention is associated with a 9 percentage point increase in the chance of meeting the quality standard. Using the stricter standard, this effect is more muted – even in medium and maximum intervention schools, many teachers do not get above 50% in the English and mathematics tests – but still statistically significant. 17% of maximum intervention schools meet the standard, compared to only 1% of schools that have received the minimum intervention.

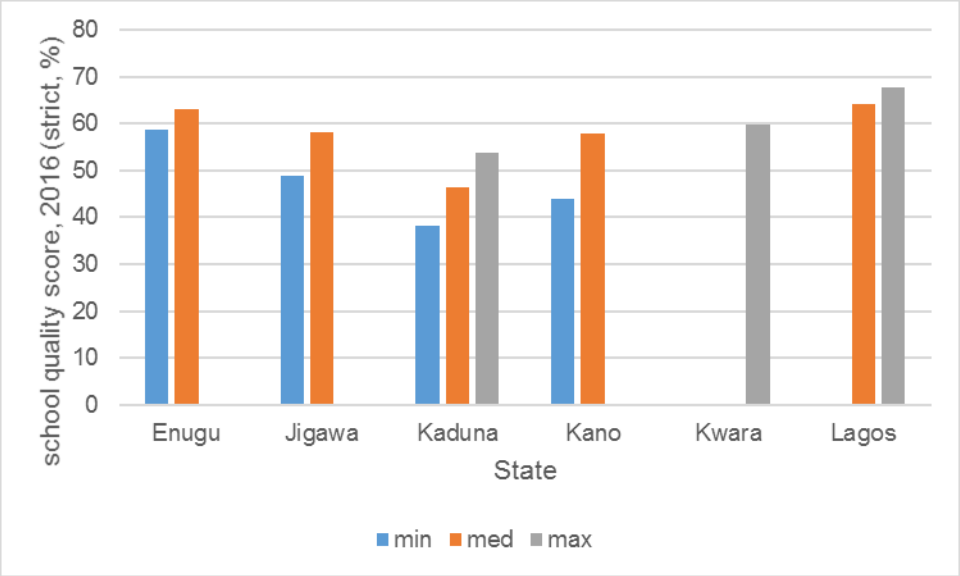
Table 35: School quality by ESSPIN intervention group in 2016

	Min.	Med.	Max.	Estimated effect of one year of full intervention
Quality score (%)	46.7	60.3	61.5	+5.1*
School meets quality standard (%)	12.1	35.6	36.7	+9.4*
Quality score (strict version) (%)	43.7	58.2	59.7	+5.5*
School meets quality standard (CS2 version) (%)	1.1	13.2	17.3	+3.6*

Note. The estimated effect of one year of full intervention is the coefficient of the number of years of full Output Stream 3 intervention in a regression on the quality score or likelihood of meeting the quality standard, controlling for state. * indicates that the coefficient on years of ESSPIN intervention is statistically significant ($p < .05$) in a regression on the stated indicator, with controls for state.

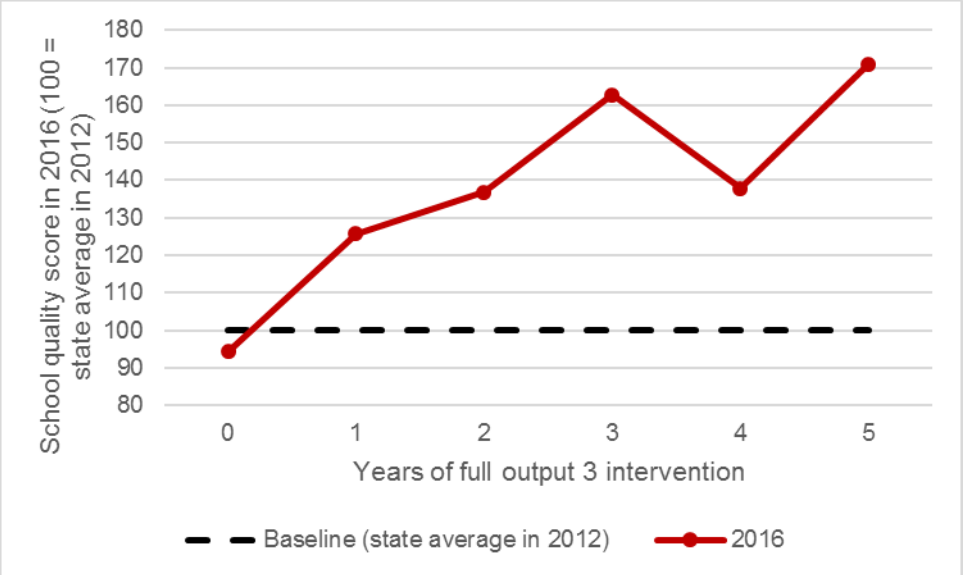
The difference between intervention groups is consistent across the states, and is most pronounced in Jigawa, Kaduna and Kano (Figure 13) – the states starting from the lowest base, and containing the largest and most disadvantaged public primary school populations in the ESSPIN programme.

Figure 13: School quality by ESSPIN intervention group and state, 2016



Can the differences in quality between the intervention groups be attributed to the intervention, or are they associated with differences in the schools at baseline? One way of answering this is to focus on the change over time in the different intervention groups. Have the schools which received more intervention between 2012 and 2016 also improved more?[18] In Figure 14 we use the average quality score in each state in 2012 as a baseline and calculate how much higher the score is in 2016.[19] The relative score in 2016 rises steadily with the number of years of full Output Stream 3 intervention under ESSPIN. The exception is a 'dip' at four years – the quality is higher in schools that have received three or five years of intervention. The dip reflects differential progress in school quality across the states; the schools that have received four years of intervention include all of the schools in Kwara and the smaller group of pilot schools in Kaduna, while a majority of the schools that have received three or five years of intervention are in Lagos.

Figure 14: School quality in 2016 relative to 2012, by years of intervention



[18] This analysis relies on the 'parallel trends' assumption: the assumption that, while schools may have differed in terms of quality as baseline, there were no systematic differences in their trends of change over time.

[19] The relative score for a school *i* in state *j* is $100 \times \frac{S_{i,2016}}{\bar{S}_{j,2012}}$ where $\bar{S}_{j,2012}$ is the state average quality score in 2012 (CS1) and $S_{i,2016}$ is the school's quality score in 2016 (CS3).

Examining the difference between intervention groups in the change over time between 2012 and 2016 confirms that there is a significant difference: in schools that received two to three years of intervention during the relevant time period, the average quality score increased by 19 percentage points, while in schools with no intervention or one year of intervention, the score increased by 10 percentage points (Table 36).

Table 36: Difference between intervention groups in change over time (2012–16)

	Intervention during 2011/12–2014/15		
	0–1 years	**2–3 years**	**Difference**
2012 (CS1)	36.7	42.0	5.3
2014 (CS2)	34.2	52.3	18.1
2016 (CS3)	47.1	60.8	13.7
Difference (2012–2016)	10.4	18.8	8.4 *

Looking at quality changes across the states, there are at least three potential effects at work, which are difficult to disentangle. First, ESSPIN may be having an effect on the pace of improvement in school quality. Second, states may be following different trends when it comes to school improvement, regardless of ESSPIN. Third, the amount of ESSPIN intervention – measured here as the years of full Output Stream 3 intervention – varies a lot between the states.

We use linear regression analysis to try to disentangle these effects. Results are presented in Annex C. Using the original quality score (which does not take into account teachers' test scores), each year of full intervention during the period between CS1 and CS3 was associated with a statistically significant rise of around 4–6 points on the school quality score, accounting for the school's quality at baseline (CS1).

In order to look at the differences between states, we have to limit the analysis to Enugu, Jigawa, Kaduna and Kano, which are the states where there is variation in the amount of ESSPIN intervention between CS1 and CS3. This reduces the sample size and statistical power, and the effect loses its statistical significance. Adding additional controls to take account of the state in which the school is located, the estimated effect remains positive, with a similar magnitude, but is not statistically significant.[20] We estimate the effect at around 5 percentage points for each year of intervention, but the 95% confidence intervals indicate a range of possible values from -0.7 to +11.3 percentage points. These results suggest that part of the school improvement effect may be due to pre-existing differences in trends between the states that have received more or less ESSPIN intervention, and make it harder to be confident in our estimate of the size of the true ESSPIN effect. However, this could also reflect some combination of reduced sample size, and the general difficulty in disentangling state differences from ESSPIN effects. An additional confounder is that states vary in the way they have implemented ESSPIN, and so part of the effect attributed to states in the regression analysis could be due to unmeasured variation in the intervention.

Further linear models use the stricter quality score introduced in CS2, which takes into account teachers' English and mathematics test scores. In this case the effect remains significant even after controlling for state. The estimated effect is around 6 percentage points for each year of intervention. These provide greater confidence that there is a genuine ESSPIN effect on school quality that cannot be explained away in terms of differences between the states.

[20] 'Effect' is used here in the statistical sense, referring to an association between differences in one variable and differences in another variable. It does not imply a causal relationship of the intervention with school quality, although the aim is to support causal inference by eliminating some other possible explanations for the association.

In summary, the results with regard to overall school quality are encouraging. School quality has been improving over time across the six states, and is much higher in schools that have received more years of ESSPIN intervention. The estimated ESSPIN effect is robust to controlling for differences in school quality at baseline and controlling for state. The estimated effect is statistically non-significant in some specifications, but this may reflect reduced sample size and unmeasured variation in the intervention between states.

6 Learning outcomes

The ultimate aim of ESSPIN is to improve learning outcomes in government schools in the six states. In this chapter, we examine the trends in learning outcomes over time, differences in learning outcomes between schools that have received more or less ESSPIN intervention, and evaluate whether effects on learning achievement can be attributed to ESSPIN.

6.1 Pupil learning achievement in English literacy and numeracy

Learning outcomes were measured in literacy and numeracy at Grades 2 and 4, and analysed using IRT (see Allen, 2016a and Allen, 2016c). The analysis for each test produces a scale score which, by design, has an average (mean) of 500 and standard deviation of 100. This scale is also divided into bands, indicating the level of proficiency of the learner. For the Composite Surveys, bands have been designed to correspond to the levels of proficiency expected at each grade in the Nigerian curriculum. For example, a learner in Band 2 for literacy is one who is able to demonstrate knowledge and skills in at least some of the tasks that are considered to be within the range of Grade 2 proficiency. Table 37 and Table 38 list some examples of the tasks within each band. In addition to being able to assess which band each learner is in, we can analyse their performance in different learning sub-scores or learning domains within literacy and numeracy (see

Box 9: Developing scores for learning domains within literacy and numeracy

Using IRT, it is possible to analyse the way that different types of question cluster together in the data. Together with a review of the content of each item, this information can be used to develop sub-scores relating to different learning domains. In the Composite Survey data, there was evidence for three domains in the literacy tests and three in numeracy.

Learning domain	Description	Example
Literacy: receptive	Responds appropriately to material provided in a prompt	Where is the book?
Literacy: fluency	Ready and fluent production of writing or speech	Read the passage aloud as carefully as you can: 'Good morning, my name is Fatima…'
Literacy: productive	Responds to more open-ended prompts, drawing on their own knowledge, imagination etc.	*Please make one complete sentence* [about three things the learner has identified on the table]
Numeracy: calculations	Responds to items requiring the recall and application of ways of doing arithmetic taught in schools	I want you to do some addition sums: '2 + 3 = ____'
Numeracy: everyday maths	Feeling for how numbers work	I want you to measure the distance from your chair to the door using your feet. Tell me how many steps it takes.
Numeracy: word problems	Convert verbal statements into sums	What profit will I make if I buy a pineapple at 100 naira and sell it at 120 naira?
Other (Grade 2 only)	Grade 2 numeracy is relatively short and so the remaining items form a mixed collection	*What time is it?* [Pointing to a picture of a clock showing three o' clock]

, below).

Table 37: Examples of knowledge and skills that learners in each literacy band can demonstrate

Band 4: Grade 4 and above	Read and understand the grammatical structure of a sentence and complete a missing word using 'where', 'which', 'what' and 'who' Follow the conventions of letter-writing to complete a letter template. Completing grammatically accurate sentences, with correct spelling, and a greeting and sign-off Read for meaning a short, simple text with a range of sentence structures independently
Band 3: Grade 3 literacy	Read phonically decodable two-syllable and three-syllable words that include common diagraphs and adjacent consonants Independently plan and write a grammatically correct simple sentence Read a simple sentence for meaning and complete a missing word using correct spelling
Band 2: Grade 2 literacy	Use phonic knowledge to utter initial sounds of the names of familiar animals Use knowledge of common inflections in spellings, plurals, to write the answer to a question Spell simple high frequency words accurately
Band 1: Emerging literacy	Verbally compose a short grammatically correct sentence in the continuous present tense in response to a question about a picture Listen to a short passage and remember specific details to respond verbally to a question Clearly shaped and correctly orientated copying of words, with an understanding of space and full stops
Band 0: Pre-literacy	Understand and respond verbally with a grammatically correct sentence to a simple question about their age Understand and respond verbally with a grammatically correct sentence to a simple question about their name Use phonic knowledge to utter initial sounds of the names of familiar objects and animals

Table 38: Examples of knowledge and skills that learners in each numeracy band can demonstrate

Band 5: Grade 5 and above	Solve a word problem involving differences in time Determine which number rule was used to make one number into another Solve a simple algebra problem
Band 4: Grade 4 numeracy	Being able to gather information by interpreting simple graphs Calculate the area of a rectangle, multiplying a decimal number, to one decimal place, by a one-digit number, and record the answer in m2 Choose the most appropriate strategy to subtract a decimal number, to two decimal places and a two-digit number, involving measure
Band 3: Grade 3 numeracy	Multiply a two-digit number by a one-digit number Use short division; subtract a two-digit number from a two-digit number crossing the tens boundary Choose a strategy to add a three-digit number and a two-digit number crossing the tens boundary, involving money
Band 2: Grade 2 numeracy	Use non-standard units of measure to compare the capacity of three containers Subtract a two-digit number from a two-digit number Name common 2D shapes Extend counting past 800 and count in tens
Band 1: Emerging numeracy	Recognise and complete a sequence of three two-digit numbers that are multiples of five Subtract a one-digit number from a two-digit number 1–19 Read analogue clock to the hour
Band 0: Pre-numeracy	Compare the length of two straight lines Use non-standard units of measure to compare the capacity of three containers Count to 10

Box 9: Developing scores for learning domains within literacy and numeracy

Using IRT, it is possible to analyse the way that different types of question cluster together in the data. Together with a review of the content of each item, this information can be used to develop sub-scores relating to different learning domains. In the Composite Survey data, there was evidence for three domains in the literacy tests and three in numeracy.

Learning domain	Description	Example
Literacy: receptive	Responds appropriately to material provided in a prompt	Where is the book?
Literacy: fluency	Ready and fluent production of writing or speech	Read the passage aloud as carefully as you can: 'Good morning, my name is Fatima…'
Literacy: productive	Responds to more open-ended prompts, drawing on their own knowledge, imagination etc.	*Please make one complete sentence* [about three things the learner has identified on the table]
Numeracy: calculations	Responds to items requiring the recall and application of ways of doing arithmetic taught in schools	I want you to do some addition sums: '2 + 3 = ____'
Numeracy: everyday maths	Feeling for how numbers work	I want you to measure the distance from your chair to the door using your feet. Tell me how many steps it takes.
Numeracy: word problems	Convert verbal statements into sums	What profit will I make if I buy a pineapple at 100 naira and sell it at 120 naira?
Other (Grade 2 only)	Grade 2 numeracy is relatively short and so the remaining items form a mixed collection	*What time is it?* [Pointing to a picture of a clock showing three o' clock]

There have been significant improvements over time in learners' numeracy in Grade 4, but overall the trend in learning outcomes between 2012 and 2016 has been mixed (Table 39). Grade 2 literacy and numeracy scores have worsened significantly since 2014, while Grade 4 literacy has not changed significantly. For Grade 2 literacy, the change reflects a larger proportion of students in Band 0 (pre-literacy) as opposed to Bands 1 or 2. For Grade 2 numeracy, the pattern is more complicated: there are fewer children in the lowest band (pre-numeracy) and in the highest band (Grade 2 numeracy), and more children in the middle band. This suggests that the situation may be improving in some schools (where a higher proportion of learners would formerly have been in the lowest achievement band) but worsening in others (with fewer learners now reaching the highest band). For Grade 4 numeracy, there are more learners in the highest bands compared to 2014, although the change is only statistically significant for the top band (Band 5).

Looking at the sub-scores for our different learning domains (see

Box 9: Developing scores for learning domains within literacy and numeracy

Using IRT, it is possible to analyse the way that different types of question cluster together in the data. Together with a review of the content of each item, this information can be used to develop sub-scores relating to different learning domains. In the Composite Survey data, there was evidence for three domains in the literacy tests and three in numeracy.

Learning domain	Description	Example
Literacy: receptive	Responds appropriately to material provided in a prompt	Where is the book?
Literacy: fluency	Ready and fluent production of writing or speech	Read the passage aloud as carefully as you can: 'Good morning, my name is Fatima…'
Literacy: productive	Responds to more open-ended prompts, drawing on their own knowledge, imagination etc.	*Please make one complete sentence* [about three things the learner has identified on the table]
Numeracy: calculations	Responds to items requiring the recall and application of ways of doing arithmetic taught in schools	I want you to do some addition sums: '2 + 3 = ____'
Numeracy: everyday maths	Feeling for how numbers work	I want you to measure the distance from your chair to the door using your feet. Tell me how many steps it takes.
Numeracy: word problems	Convert verbal statements into sums	What profit will I make if I buy a pineapple at 100 naira and sell it at 120 naira?
Other (Grade 2 only)	Grade 2 numeracy is relatively short and so the remaining items form a mixed collection	*What time is it?* [Pointing to a picture of a clock showing three o' clock]

above), in English literacy there are negative changes across each of the different domains. Among Grade 2 learners, the change is particularly pronounced in the 'receptive' domain, which includes questions where learners have to respond correctly to a particular prompt, such as by using a full sentence in English to say where an object has been placed. Among Grade 4 learners, the change is most pronounced in the 'fluency' domain – testing their ability to read, write or speak with fluency. In Grade 4, learners' 'productive' literacy – responding to more open questions from their own knowledge or imagination – has also worsened. This result would be consistent with an increase in the number and proportion of children from disadvantaged or traumatised backgrounds joining school at Band 0 level, irrespective of age and grade of entry.

For numeracy, it is harder to interpret the pattern of change in the sub-scores: in the Grade 2 test, both sub-scores worsened over time. In the Grade 4 test, the three sub-scores show different trends over time: the calculations score worsened between 2012 and 2014, and worsened further (although the change was not statistically significant) between 2014 and 2016. The everyday mathematics score improved during 2012–14 but then worsened during 2014–16; and the word problems score worsened during 2012–14 but improved during 2014-16. Thus, while the overall Grade 4 numeracy score did not change significantly between 2012 and 2016, there was a diverse pattern of change in the different learning domains.

Table 39: Learning outcomes in 2012–16

	2012 (CS1)	2014 (CS2)	2016 (CS3)	Change 2012–16	Change 2014–16

Grade 2 literacy	460.7	465.6	447.2	-13.5	-18.5*
– Band 0: Pre-school (%)	62.9	69.8	83.4	+20.6*	+13.6*
– Band 1: Grade 1 (%)	12.1	12.1	5.7	-6.4*	-6.4*
– Band 2: Grade 2 (%)	25	18	10.8	-14.2*	-7.2*
– Receptive score	480.4	465.6	429.6	-50.8*	-36.0*
– Fluency score	456.8	475.6	458.7	+1.9	-16.9*
Grade 4 literacy	458.1	449.1	447.5	-10.7	-1.6
– Band 1: Grade 1 (%)	61.8	71.9	71.6	+9.8*	-0.2
– Band 2: Grade 2 (%)	13.3	10.7	11.3	-1.9	+0.7
– Band 3: Grade 3 (%)	5.9	5.9	4.9	-1.0	-1.0
– Band 4: Grade 4 (%)	19	11.6	12.2	-6.8*	+0.6
– Receptive score	473.9	443.3	436.2	-37.7*	-7.1
– Fluency score	465.5	472.2	436	-29.5*	-36.2*
– Productive score	485.1	464.5	443.3	-41.8*	-21.2*
Grade 2 numeracy	498.6	457.1	444.7	-53.9*	-12.4*
– Band 0: Pre-school (%)	11.6	15.7	8.3	-3.3	-7.4*
– Band 1: Grade 1 (%)	63.1	68	79.4	+16.2*	+11.3*
– Band 2: Grade 2 (%)	25.2	16.3	12.3	-12.9*	-4.0
– Calculations score	512.2	460.3	443.9	-68.4*	-16.5*
– 'Other' score	502.3	484.2	415	-87.3*	-69.2*
Grade 4 numeracy	463.1	448.7	460.9	-2.2	+12.2*
– Band 0: Pre-school (%)	27.3	37.3	34	+6.7	-3.3
– Band 1: Grade 1 (%)	34	31.1	30.2	-3.8	-0.9
– Band 2: Grade 2 (%)	21.1	19	19.6	-1.6	+0.6
– Band 3: Grade 3 (%)	10.8	6.7	7.4	-3.4*	+0.7
– Band 4: Grade 4 (%)	6.7	5.7	8.1	+1.4	+2.3
– Band 5: Grade 5 (%)	0	0	0.7	+0.7*	+0.6*
– Calculations score	490.3	448.8	441.3	-49.0*	-7.5
– Everyday mathematics score	433.9	481.2	460.2	+26.3*	-20.9*
– Word problems score	494.5	451	472.5	-22.1*	+21.4*
Note. * indicates that the change over time is statistically significant ($p < .05$).					

Learning outcomes appear to be better for learners whose schools have received more years of ESSPIN intervention (Table 40). For all four tests, the estimated effect of a year of full intervention is positive, but it is only statistically significant for the literacy tests. The estimated effect is modest in magnitude: it is in the range of 0.04 to 0.12 standard deviations. In schools that have received more years of ESSPIN intervention, there appear to be fewer learners in the lowest achievement bands and more learners in the higher achievement bands.

Focusing on the sub-scores for different learning domains in literacy, learners in ESSPIN schools are doing better in all of the domains in both Grades 2 and 4, although in Grade 4 the largest difference is in terms of 'fluency', while the difference in 'productive' scores is small and not statistically significant. Students in ESSPIN schools are showing marked differences in their ability to read set passages, but less so in their ability to respond to verbal prompts and less still in their ability to produce answers to open questions from their own knowledge or imagination. In

numeracy, as for the overall scores, the difference in sub-scores between schools with different levels of ESSPIN intervention is small and not statistically significant.

Table 40: Learning outcomes by ESSPIN intervention group in 2016

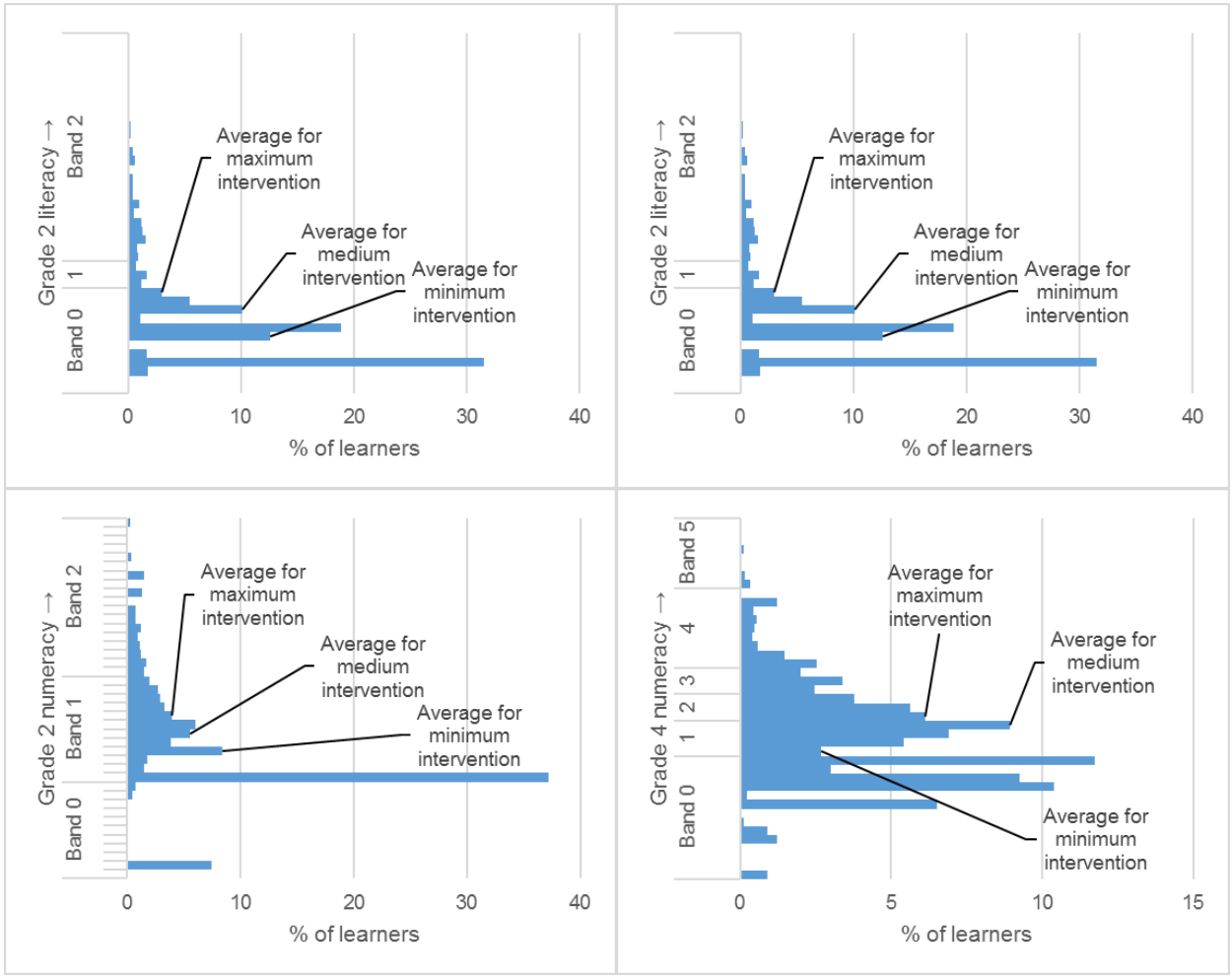
	Min.	Med.	Max.	Estimated effect of one year of full intervention
Grade 2 literacy	431.6	470.9	496	+9.2*
– Band 0: Pre-school (%)	89	77	55.3	-2.2
– Band 1: Grade 1 (%)	4.8	5.2	18.9	+1.3
– Band 2: Grade 2 (%)	6.2	17.8	25.7	+1.0
– Receptive score	416	450	473.3	+6.4*
– Fluency score	447.7	475.9	490.7	+5.0*
Grade 4 literacy	428.5	473.4	494.1	+11.5*
– Band 1: Grade 1 (%)	82.2	55.8	53.5	-7.9*
– Band 2: Grade 2 (%)	8.9	14.9	15.5	+4.3*
– Band 3: Grade 3 (%)	2.6	8.7	7.2	+1.9
– Band 4: Grade 4 (%)	6.3	20.7	23.7	+1.6
– Receptive score	415.9	463.5	487	+10.5*
– Fluency score	415.6	465.2	478.6	+14.3*
– Productive score	431.5	459.6	471.5	+2.8
Grade 2 numeracy	433.5	459.2	493	+4.4
– Band 0: Pre-school (%)	9.5	6.5	4	-2.3
– Band 1: Grade 1 (%)	82.4	75	69.1	+1.0
– Band 2: Grade 2 (%)	8.1	18.5	26.9	+1.1
– Calculations score	434.8	455.4	484.3	-0.1
– Other score	402.9	430	471.5	+3.5
Grade 4 numeracy	442.1	485.2	512.2	+9.2
– Band 0: Pre-school (%)	41.5	24.1	14	-6.1
– Band 1: Grade 1 (%)	29.5	31.3	31	+2.8
– Band 2: Grade 2 (%)	17.8	21.9	24.5	+3.0
– Band 3: Grade 3 (%)	6.5	8.4	11.5	-0.3
– Band 4: Grade 4 (%)	4.7	12.8	16.1	+0.4
– Band 5: Grade 5 (%)	0	1.4	2.8	+0.6
– Calculations score	426.8	457.6	493.9	+3.0
– Everyday mathematics score	440.5	486.3	512.2	+10.4
– Word problems score	457.2	493	509.6	+8.7

Note. * indicates that the coefficient on years of ESSPIN intervention is statistically significant ($p < .05$) in a regression on the stated indicator, with controls for state.

Learners in schools that have received minimum intervention from ESSPIN are disproportionately concentrated in the lowest performance bands (blue bars in Figure 15, below) while those in

schools with medium or maximum concentration (yellow bars) are more likely to be found in the medium and higher bands.

Figure 15: Distribution of test scores by Output Stream 3 intervention group in 2016

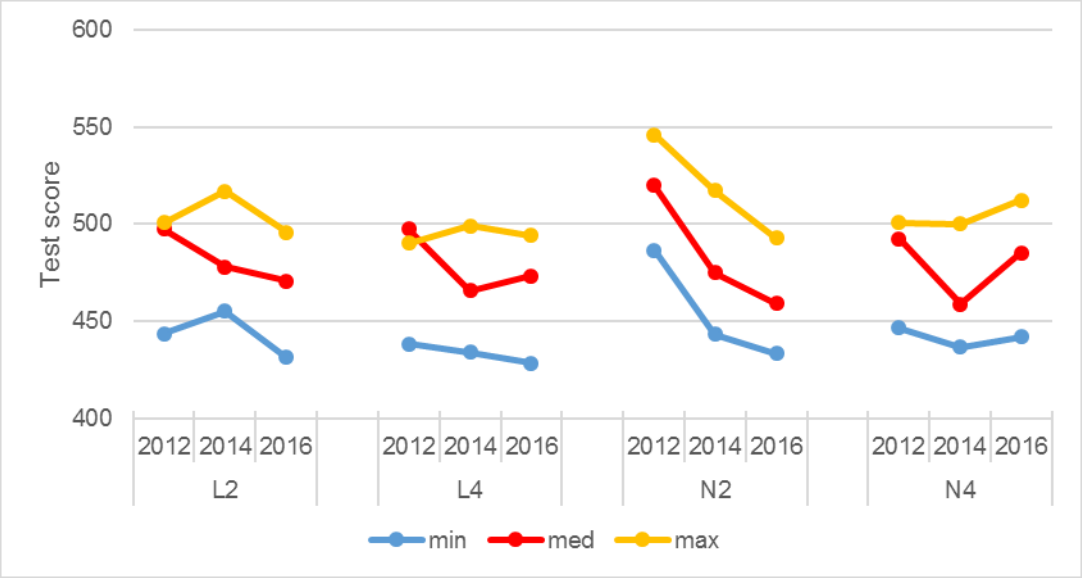


In the CS2 report we noted that, although test scores were generally worsening over time, there was some evidence of a less severe deterioration in schools that had received ESSPIN's intervention compared to control schools. As noted above, and shown in Figure 16, below, by 2016, there has been some recovery in Grade 4 numeracy results, so that these are now higher than in 2012, while Grade 2 literacy and Grade 2 numeracy have continued to worsen. When we disaggregate the change over time according to ESSPIN intervention, a number of points emerge. First, schools that received more years of ESSPIN intervention already had better learning outcomes in 2012. The schools in the minimum and medium intervention groups show varying patterns of change over time – in particular, Grade 4 test results dipped in 2014 in the medium intervention group, before partially recovering in 2016. Taking the overall trend between 2012 and 2016, however, the medium and minimum intervention groups have followed roughly parallel tracks, with learning outcomes worsening by about the same amount.

Comparing the medium and maximum intervention groups, it is notable that these had roughly similar learning outcomes in 2012 in the literacy tests and in Grade 4 numeracy, but have since diverged, with the maximum intervention schools improving slightly, while the medium intervention schools worsened. However this pattern does not apply for Grade 2 numeracy, where the maximum intervention schools were already doing better in 2012 than the medium intervention schools, and all three intervention groups have worsened at roughly the same pace.

This rather mixed pattern of results makes it difficult to draw clear conclusions about change over time. These patterns are explored more rigorously in Section 6.2.4 below, using regression and matching analysis to examine how change over time varies with ESSPIN intervention, and controlling for possible confounding variables, such as school characteristics and unobserved state differences.

Figure 16: Learning outcomes by test, year, and ESSPIN intervention group



6.2 Controlling for school and learner characteristics

6.2.1 Differences in background characteristics

As was noted in the CS2 overall technical report, schools in the different intervention groups – minimum, medium or maximum – do have somewhat different background characteristics (Table 41). Although the pattern varies across states, schools that have received more years of ESSPIN intervention tend to be older, larger, and more urban. The schools with medium or maximum intervention were more likely to have had a parent–teacher association and an SBMC in 2009/10 than the schools which later received less ESSPIN intervention. Urban schools, in particular, tend to have better learning outcomes than rural schools, and, if left uncorrected, this difference between the intervention groups could bias our estimates of ESSPIN intervention effects upwards.

In Jigawa, Kaduna and Kano, the EMIS data suggest that there have been large increases in enrolment during 2009/10–2014/15, of around one-third in Jigawa and around 50% in Kaduna and Kano. In the latter two states, the schools that have received more years of ESSPIN intervention show significantly larger increases. In Kano's maximum intervention schools there were massive increases, of over 80%, in enrolments, while in Kaduna's medium intervention schools enrolments increased by more than 50%.

These figures may reflect policies specifically designed to raise enrolments in states where low enrolment is an issue. For example, the increase in Kaduna may relate to the introduction of 'free education' policies with school feeding. Schools may have difficulty coping with rapid enrolment increases, especially if the new learners are from more disadvantaged backgrounds[21], and

[21] Little data are available on the profile of new learners entering schools. However, there have in the past been strong correlations between school attendance and socioeconomic status (Mezger, 2016). This means that further increases in enrolment are likely to come disproportionately from households that are poorest and where the learner's parents have relatively low levels of education.

increases in the PTR would tend to reduce teachers' ability to ensure all students achieve good learning outcomes. Therefore, this difference between intervention groups in Kaduna and Kano could bias our estimates of ESSPIN intervention effects downwards.

In summary, there are a number of differences between the groups of schools that have had more ESSPIN intervention and those that have had less, and taken together these could bias our estimates of ESSPIN's effect in either direction. We use a number of statistical methods to control for these differences in the following sections.

6.2.2 Timing of ESSPIN intervention and learning outcomes in 2016

Analysing the effects of ESSPIN's intervention is made complicated by the diversity of ESSPIN's intervention between and within states. Among the schools that have received ESSPIN's intervention, the timing and duration of the intervention varies considerably. States have rolled out the programme to increasingly large proportions of schools, but have followed different schedules in doing so. Some groups of schools were part of an initial pilot but then had no additional training – or only part of the full ESSPIN training package – for one to three years before being included again as part of state-wide scale-up. Others have had consistent intervention throughout the period between 2008/09 and 2015/16. Still others were not included in the pilot or early phase of the programme and have been included only since 2014/15.

The timing of the intervention could affect learning outcomes in at least two ways. First, it may take time for the effects of leadership training, teacher training and other interventions to filter through to measurable gains in learning outcomes. This is especially so given that children are tested in the second and fourth grades. From this perspective, for example, a Grade 4 learner whose teachers benefited from training during the current school year is unlikely to benefit as much as one whose teachers were trained four years ago so that they have had consistent exposure to better teaching. This would suggest that training initiated a longer time ago would have larger effects on learning outcomes. Second, however, it is possible that effects of training fade out over time, particularly if it is common for teachers to transfer between schools or to retire or leave the profession. More recent training would then have a larger effect on learning outcomes.

We explore this using a regression analysis that compares every combination of timing and duration of intervention (with the average across these intervention patterns serving as a baseline for comparison). We try two versions of the regression: the first controls for differences between the states, while the second controls for both state and for differences in school characteristics (Table 42). Looking at the results when we control for state and other characteristics, learning outcomes are highest in the schools which have received the most intervention (three to six years) continuing up to 2014/15 or 2015/16. Those schools that had three to four years of intervention overall, but that have not had any intervention since 2013/14, have lower learning outcomes.

Table 41: Background characteristics of schools according to ESSPIN intervention

	Enugu			Jigawa			Kaduna				Kano			Lagos		
	Min.	Med.		Min.	Med.		Min.	Med.	Max.		Min.	Med.		Med.	Max.	
Distance from LGA HQ (km)	11	10		18	14	*	30	20	18	*	9	9		5	3	*
Age of school	51	56	*	31	36	*	22	28	28	*	22	33	*	46	46	
Urban (%)	3	0		7	18	*	7	24	10	*	34	53	*	78	75	
Nomadic (%)	0.0	0.0		8.7	3.3	*	5.8	3.9	2.5	*	3.6	1.3	*	0.0	0.0	
Islamic (%)	0.0	0.0		4.8	3.3		1.3	0.0	0.0	-	59.6	35.1	*	0.0	0.0	
Double-shift (%)	0.1	0.5		0.0	0.0		1.5	6.3	2.6	*	8.5	21.5	*	0.1	0.0	
Has parent–teachers association (%)†	96.1	98.0	*	95.1	95.4		97.5	98.5	98.7		97.8	96.5	*	99.1	100.0	
Has SBMC (%)†	83.1	96.0	*	95.1	97.6	*	95.8	98.6	99.4	*	91.6	95.0	*	99.1	100.0	
PTR in 2014/15	17.6	16.2	*	60.8	47.9	*	51.7	39.5	38.3	*	70.0	63.2	*	26.3	29.9	*
No. classrooms†	5.1	6.3	*	3.8	5.9	*	4.6	6.7	7.3	*	5.3	10.4	*	11.3	13.4	*
No. teachers†	9	10	*	5	8	*	7	14	14	*	8	17	*	16	18	*
Enrolment 2009/10	198	204	*	183	438	*	225	365	285	*	348	833	*	384	496	*
Enrolment 2013/14	150	164		218	506	*	250	410	386	*	383	1037	*	390	483	*
Enrolment 2014/15	140	158		215	493	*	257	415	396	*	397	1024	*	391	470	*
% change in enrolment 2009/10 – 2014/15	2.4	-10.1		31.0	33.5		46.2	40.3	83.8	*	40.5	51.8	*	11.1	1.0	
% change in enrolment 2013/14 -– 2014/15	2.1	1.3		5.8	0.7	*	9.8	8.3	7.5		9.5	0.8	*	3.3	-1.0	
% of teachers with academic diploma/degree	62	60		46	58	*	60	60	65	*	61	66	*	41	41	*
-- % with PGDE, BEd or MEd	31	32	*	4	4	*	5	5	5	*	8	9	*	18	19	
-- % with NCE, Grade II or equivalent	64	63		75	75	*	81	79	80		54	61	*	39	41	
Number of schools	820	409		1,065	872		3,361	601	163		4,973	509		903	97	

Notes. † 2014–15. * indicates a statistically significant difference between the different intervention groups ($p < .05$)

Schools that started to receive the intervention recently, and have therefore had a shorter duration of intervention (one to two years), have lower learning outcomes compared to the schools that have had three or more years of intervention. However, comparing the different groups of schools that have had two years of intervention, those that received the intervention most recently had the highest learning outcomes in 2016.

Overall, this suggests that the effect of training in a school may fade out over time. Possible explanations for this include teachers and head teachers transferring to other schools; enrolment increases that offset any improvements in teaching methods; and teachers forgetting their skills over time or becoming less motivated to use them. Learning outcomes are best in the schools that have had relatively consistent intervention over three or more years up to the present. There is no evidence of a lag between intervention and improvements in learning outcomes: if such a lag exists, it appears to be fully offset by a negative effect of having stopped the intervention two or more years ago.

Table 42: Difference in test scores in 2016 by timing of ESSPIN intervention

Years of intervention	Total years	L2		L4		N2		N4	
Controlling for state									
2013/14, 14/15	2	-79	*	-55	*	-49		-37	
2014/15, 15/16	2	-46	*	-32	*	-4		-31	*
2013/14, 14/15, 15/16	3	-17	*	-4		-10		-18	*
2014/15 only	1	17		-11		-36		-20	
2012/13, 14/15, 15/16	4	1		-10		0		-12	
2013/14 only	1	9		-2		-15		1	
2012/13, 13/14, 14/15	3	-35	*	0		15		25	
2011/12, 12/13, 14/15, 15/16	4	17	*	6		14	*	-10	
2012/13, 13/14	2	21		8		-3		6	
2009/10, 11/12, 12/13, 14/15, 15/16	6	23		7		14		-10	
2009/10, 11/12, 12/13	4	28		11		-4		16	
2009/10, 10/11, 13/14	3	27		25		12		22	
2009/10, 10/11, 14/15	3	-14		25	*	45	*	46	*
2011/12, 12/13	2	49	*	32	*	21	*	22	*
Controlling for state and other characteristics									
2012/13, 13/14	2	-92	*	-107	*	-78	*	-78	*
2009/10, 10/11, 13/14	3	-91	*	-95	*	-60	*	-79	*
2014/15 only	1	-75	*	-94	*	-80	*	-73	*
2013/14 only	1	-90	*	-92	*	-70	*	-67	*
2009/10, 11/12, 12/13	4	-65	*	-72	*	-57	*	-52	*
2011/12, 12/13	2	-51	*	-65	*	-39	*	-45	*
2013/14, 14/15	2	5		32		1		40	
2014/15, 15/16	2	32	*	46	*	44	*	26	*
2013/14, 14/15, 15/16	3	51	*	62	*	34	*	41	*
2012/13, 13/14, 14/15	3	19		64	*	43	*	69	*
2009/10, 10/11, 14/15	3	42	*	63	*	76	*	51	*
2011/12, 12/13, 14/15, 15/16	4	102	*	82	*	62	*	47	*
2012/13, 13/14, 14/15, 15/16	4	100	*	87	*	58	*	63	*
2009/10, 11/12, 12/13, 14/15, 15/16	6	113	*	89	*	66	*	57	*

Note. * indicates a statistically significant difference in test scores compared to the average for each state ($p < .05$).

6.2.3 Are learning outcomes better in schools that have received more years of intervention?

As explained in Section 6.1 above, learners' test results are better in schools that have received more years of ESSPIN intervention, controlling for the state in which they are located. However, there are also some differences in the pre-existing characteristics of the schools that received more intervention (see Section 6.2.1 above). If urban schools have received more intervention than rural schools, for example, then better learning outcomes in the intervention schools may reflect better conditions for learning, or the fact that learners come from wealthier and more educated backgrounds. In this section, we add statistical controls for these 'confounding variables' –

characteristics of schools or learners that might affect learning outcomes and make it harder to tell whether the intervention is having an effect or not.

We use a combination of regression analysis and matching techniques. Regression analysis estimates the correlation of learning outcomes with ESSPIN's intervention, conditional on school characteristics. Matching techniques work by identifying pairs of schools in the sample that have received different levels of ESSPIN intervention, but are similar in other respects, and so can be used to make a fair comparison.[22] Combining the two techniques provides 'doubly robust' estimates of effects, helping to eliminate sources of bias.

We find significant differences between schools with more or less ESSPIN intervention (Table 43). Matching schools across intervention groups and adding regression controls for their characteristics actually increases the size of the estimated effect (rows 4 and 6 in the table) and makes it statistically significant across all four tests. Adding controls for state (rows 5 and 7) reduce the estimated effects, suggesting that part of the apparent intervention effect may be attributed to unobserved state-level variables. In the case of Grade 2 numeracy, the coefficient becomes negative, although not significantly so. In the preferred model, which matches and controls for state and school characteristics (row 7), the effect remains positive and significant for the Grade 4 tests in both literacy and numeracy; it is positive but non-significant for Grade 2 literacy; and it is negative and non-significant for Grade 2 numeracy. (Grade 4 students were asked questions about their parents' assets, which enable us to add a wealth index to our control variables and potentially provides more precise estimates for Grade 4 than Grade 2 students.) Among the Grade 4 students, the effect of spending at least one year in a school that has had more than one year of full ESSPIN intervention is estimated at around 0.1 standard deviations.

In short, the regression and matching analyses support the claim that ESSPIN's intervention leads to better learning outcomes. The effects appear to be somewhat larger for Grade 4 than Grade 2 students. Although the effects are not found consistently in every model, there is enough consistency to be reasonably confident that the better learning outcomes in ESSPIN schools are not just an effect of pre-existing differences in school or learner characteristics.

[22] We use *propensity score matching*, which involves using a statistical model to estimate the relationship between school characteristics and the likelihood of being in the medium or maximum intervention group (known as the school's propensity score). Schools with similar propensity scores are matched together for comparisons. Low intervention schools that cannot be matched with a medium or maximum intervention school – because their characteristics are too different – are either ignored, or assigned a low weight in the analysis. This allows us to ensure that we are comparing like for like when we compare across the different intervention groups.

Table 43: Estimates of the effect of ESSPIN's intervention on learning outcomes in 2016

Model	L2		L4		N2		N4	
(1) Control for state	19.30	*	22.88	*	12.29	*	28.04	*
(2) Control for state and school characteristics	2.20		11.26	*	-5.92		10.79	
(3) Matched (near-neighbour matching)	1.13		13.48	*	-6.02		18.75	*
(4) Matched (propensity score matching)	24.68	*	35.95	*	27.87	*	48.44	*
(5) Matched and controlled for state	-1.04		4.34		-5.73		4.88	
(6) Matched and controlled for characteristics	23.96	*	37.49	*	26.3	*	49.36	*
(7) Matched, controlled for state and characteristics	4.94		12.5	*	-6.04		9.5	*

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 ($p < .05$). The coefficients shown are the estimated effect of being in a school with an additional year of full intervention, for one year.

6.2.4 Has learning improved faster in schools that have received more years of intervention?

Another approach to determining causal effects of ESSPIN's intervention is to examine the change over time in learning outcomes between 2012 (CS1) and 2016 (CS3). In schools that received more years of ESSPIN intervention between these two years, learning outcomes should have improved faster if the intervention has worked. This can be tested with a 'difference in differences' regression analysis using the panel of schools that were included in all three rounds of the survey, and testing the extent to which change over rounds is greater for schools that have received more ESSPIN intervention (row 1 in Table 44). An alternative approach is to examine learners' test results in 2016, controlling for the school's average learning outcomes in 2012 (row 4 in Table 44).

With both approaches, we find significant positive effects of the intervention on learners' numeracy scores (rows 1 and 4, Table 44). Using the second approach, we also find significant positive effects on literacy scores. This suggests that schools which had more ESSPIN intervention between 2012 and 2016 showed more improvement in learning outcomes than those which had less intervention.

Difference in differences analysis relies on an assumption of parallel trends: that is, that schools with more intervention and less intervention during 2012–2016 were not already moving on different trends with regard to learning outcomes in 2012. This assumption may not hold, for example, if urban schools tend to improve learning outcomes faster than rural schools. We therefore add controls for the state the school is in (rows 2, 6, and 9) and for the schools' characteristics (rows 3, 5, 8 and 9). We also use propensity score matching, as in Section 6.2.3 above, to ensure we are comparing like for like when we compare intervention groups.

The results are somewhat dependent on the method of analysis used. Using the first difference in differences approach, adding controls for state and school characteristics makes the estimated ESSPIN effects non-significant and in some cases negative (but still non-significant). Using the lagged outcome variable approach, the effects remain significant and positive when school characteristics are added. Where significant, the effects of exposure to the intervention are mostly between 0.03 and 0.14 standard deviations, although for Grade 2 numeracy we estimate effects as high as 0.4 standard deviations. However, controlling for state again makes the effect non-significant, and in some cases negative.

Table 44: Estimates of the effect of ESSPIN's intervention on learning outcomes in 2016 compared to 2012

Model	L2		L4		N2		N4	
(1) Difference in difference	0.67		0.67		3.96	*	3.03	*
(2) – with controls for state	0.86		-0.61		0.13		0.48	
(3) – with controls for state and school characteristics	-0.46		-1.26		-1.66		-0.31	
(4) Cross-section in 2016 controlling for learning outcomes in 2012	13.61	*	4.36	*	19.14	*	8.44	*
(5) – with school characteristics	7.69	*	2.58	*	8.74	*	4.54	*
(6) – with controls for state and school characteristics	2.15		0.91		-1.72		0.36	
(7) Matched cross-section controlling for learning outcomes in 2012	12.2		0.34		40.44	*	-3.61	
(8) – with school characteristics	12.7	*	9.04	*	34.2	*	10.78	*
(9) – with controls for state and school characteristics	-1.48		3.98		-10.68		-3.45	

Note. * indicates a positive coefficient in the regression ($p < .05$). Analysis is at the individual learner level but learning outcomes in 2012 are the average for each school. A unit of 'exposure' to ESSPIN is conceptualised as being in a school which has had one year of intervention for one year. For models 1–6, the coefficient is the estimated effect of an additional unit of exposure. For models 7–9, we use a binary exposure variable: for Grade 2 learners, this means they have four or more units of exposure, and for Grade 4 learners it means they have had seven or more units of exposure. Models 1–3 test an interaction term between exposure to ESSPIN intervention during 2012–16 and the round of the survey.

6.2.5 Changes in enrolment and PTRs

Have increases in enrolment and in PTRs made it difficult for schools to maintain quality? In this section we briefly examine trends in enrolments, teacher numbers and PTRs, and use further regression analysis to examine whether rapid increases in PTRs have a negative effect on learning outcomes.

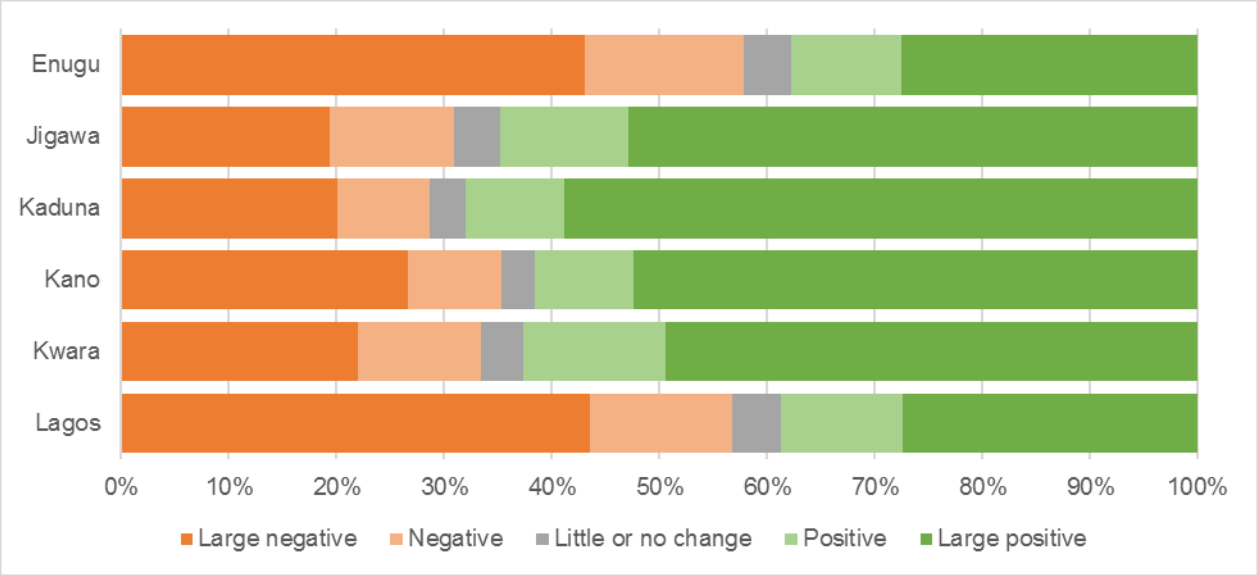
Looking at change over time within the schools for which data are available in the Education Management Information System (EMIS), there have been large positive changes in enrolment between 2009 and 2014 in 40% of schools, and smaller positive changes in a further 13% of schools. The growth has been particularly large in Jigawa, Kaduna and Kano, while most schools in Enugu, Kwara and Lagos have shrunk. Looking at the more recent period – between 2013/14 and 2014/15 – there have been positive changes in 47% of schools but negative changes in 37% of them. There was a continued high rate of growing schools in the three northern states, and also in Kwara and Lagos during 2013/14 to 2014/15.

Head teachers' responses in the survey, based on their own registers, suggested that many schools, particularly in Kaduna, continued to grow rapidly in 2014/15 and 2015/16, with large positive changes in enrolment. Overall (using survey weights) 28% of head teachers reported changes in enrolment above 20%, and 24% of head teachers reported smaller increases in enrolment, although one-in-three schools also reported declines in enrolment.

Teacher numbers also grew in one-third of schools during 2009/10–2014/15, but these schools with more teachers were concentrated in Lagos. The result is that in most schools in Jigawa, Kaduna, Kano, and Kwara, there were large increases in the PTR during 2009/10–2014/15 (Figure 17). In Enugu and Lagos, by contrast, the PTR declined in most schools. During 2013/14 to 2014/15, most schools in Kaduna and Kano, and many schools in Jigawa and Kwara, continued to

see rising PTRs, while in Lagos most schools had declining PTRs. In Kaduna and Kano there were particularly large proportions of schools with increasing PTRs in rural areas, while in Jigawa both rural and urban schools had increasing PTRs. These are the states and locales with the teachers who struggle the most to meet curriculum knowledge standards, that contain the most out-of-school children, internally displaced persons, poverty, and illiterate and marginalised households.

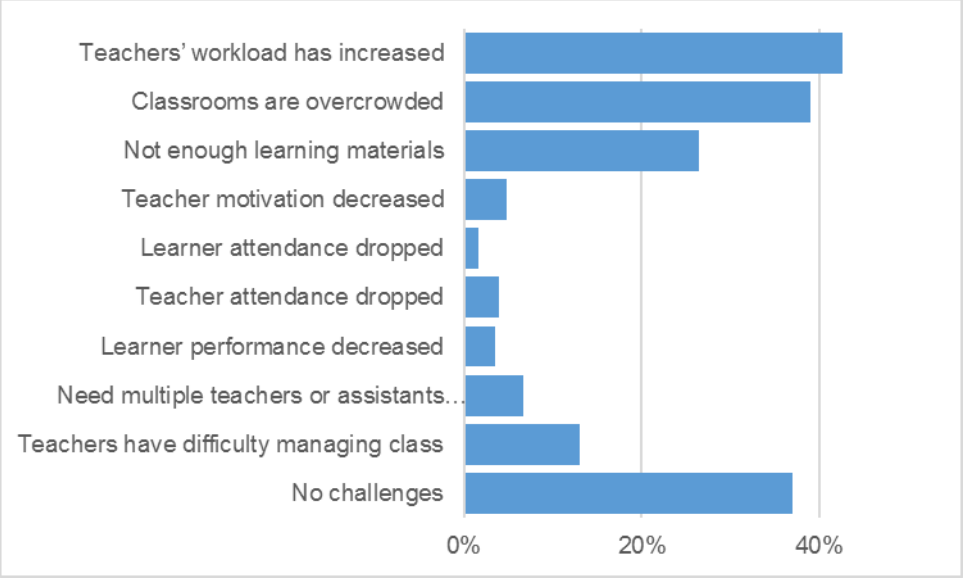
Figure 17: Change in PTRs during 2009/10–2014/15



Note: 'Large' negative or positive change refers to a change of more than 20%. 'Little or no change' refers to change of less than 3%.

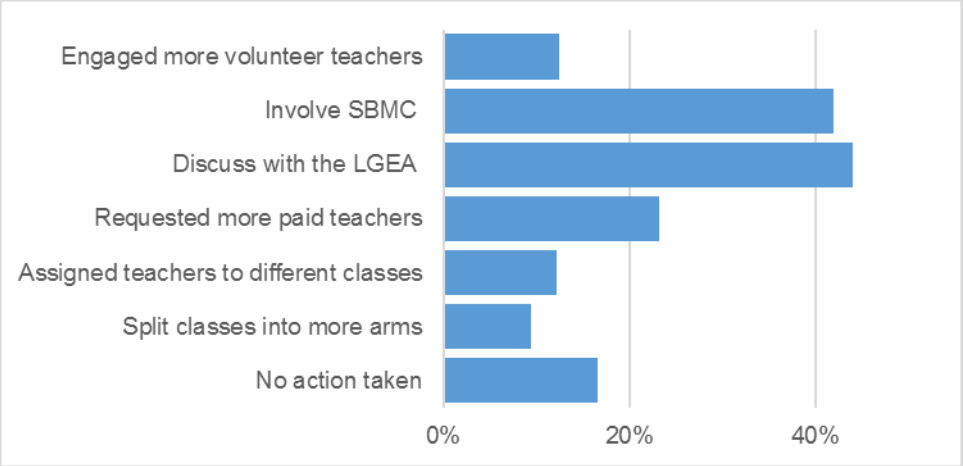
Some 63% of head teachers who reported increasing enrolments in their school reported that the increase had caused some form of challenge. Most common among these were increases in teachers' workload, overcrowding of classrooms, and insufficient learning materials (Figure 18). Most head teachers in schools with increasing enrolments claimed to have taken some form of action to deal with the increase, most commonly reporting that they had involved the SBMC or the LGEA, and/or requested more paid teachers (Figure 19).

Figure 18: Main challenges related to increasing numbers of learners



Note: The number who gave each response is given as a proportion of those who said enrolment had increased. Multiple categories were allowed (unless they said there were 'no challenges').

Figure 19: Action taken by head teachers to address increasing enrolments



Note: The number who gave each response is given as a proportion of those who said enrolment had increased. Multiple categories were allowed (unless they said no action had been taken).

In our impact regression analyses, the coefficient on the PTR in 2009/10 is consistently negative, and in most cases statistically significant, although very small in magnitude once other school characteristics have been controlled for. The estimated coefficients are around -0.2, meaning that an increase in enrolment of 10 pupils per teacher would lead to a decline in test scores of two-hundredths of a standard deviation. This result is driven partly by a small proportion of outliers – the approximately 3% of schools in the sample that have PTRs of over 100.

To investigate this further, we model a school's average test score in 2016 as a function of test scores in 2012 and the change in PTR between 2013/14 and 2014/15. We hypothesise that schools with more rapidly increasing PTR will see less progress in test scores. The coefficient on PTR is consistently negative, supporting this hypothesis, but is not statistically significant. Based on these findings, we are not able to say with confidence whether growing PTRs are responsible for worse learning outcomes, against an alternative explanation in which growing PTRs and worsening learning outcomes are symptomatic of wider problems in the education system.

Table 45: Estimates of the effect of changing pupil-teacher ratios on learning outcomes

Model	Coefficient on change in PTR 2013/14–2014/15	N
1. Simple model controlling for learning outcomes in 2014	-14.9	536
2. Controls for states	-4.5	536
3. Controls for states and ESSPIN intervention	-4.7	536
4. Controls for states, ESSPIN intervention, and school characteristics	-10.2	490

Note. Analysis is at the school level. Learning outcomes are averaged across the four tests. The coefficients shown are the estimated effect of a 100% increase in the PTR during 2013/14 to 2014/15. A coefficient of -10 would mean that a 10% increase in the PTR is associated with a worsening of one standard deviation in test scores.

6.3 Summary and discussion

We have applied IRT to measure learners' performance in literacy and numeracy, in Grades 2 and 4, across the six states. We find that Grade 2 literacy and numeracy scores have worsened significantly since 2014, while Grade 4 literacy has not changed significantly and Grade 4

numeracy has improved significantly. Learning outcomes appear to be better for learners whose schools have received more years of ESSPIN intervention. For all four tests, the estimated effect of spending time in a school which has had two or more years of ESSPIN intervention is positive, but it is only statistically significant for the literacy tests. The estimated effect is modest in magnitude: it is in the range of 0.04 to 0.12 standard deviations. This effect is, very roughly, one-tenth of the difference between a learner performing at Grade 1 level and a learner performing at Grade 2 level in the literacy or numeracy tests.

There are some significant differences between the pre-existing schools that have received more years of ESSPIN intervention and those that received less. The pattern varies by state, but schools that have received more years of ESSPIN intervention tend to be older, larger, and more urban. In Kaduna and Kano there appear to be particularly rapid enrolment increases in schools that have received more intervention. We use a number of statistical methods to control for these differences and reduce bias in our estimates of the effect of ESSPIN's intervention.

First, we examine how the timing of the intervention alters the effect on learning outcomes. There is a wide variety of patterns across and within the states in terms of the timing of the pilot and roll-out of ESSPIN's intervention. The effect appears to be weakest for those schools that had a short duration of intervention and for those where the intervention stopped more than a year ago, suggesting that the effect of training on a school may fade out over time. Possible explanations for this include teachers and head teachers transferring to other schools; enrolment increases that offset any improvements in teaching methods; and teachers forgetting their skills over time or becoming less motivated to use them. Learning outcomes are best in the schools that have had relatively consistent intervention over three or more years up to the present.

Secondly, we add statistical controls for state, school characteristics, and learner socioeconomic background in the results from 2016. Significant differences associated with ESSPIN's intervention remain. Third, we examine change over time, controlling for the school's learning outcomes in 2012, and again, still estimate significant and positive ESSPIN effects. The results are not entirely consistent, and in particular tend to lose significance when state controls are added, but they remain significant (and in some cases, increase in magnitude) when we add controls for school characteristics and learners' socioeconomic background. Where significant, the estimated effect is generally around 0.1 standard deviations for exposure to ESSPIN's intervention, although in some models it rises to around 0.4 standard deviations.

We note that PTRs have increased dramatically in many schools during the past seven years, although there are also a number of schools in which the PTR has decreased. We find evidence that high PTRs have a negative effect on learning outcomes. However, focusing specifically on the growth in PTRs between 2013/14 and 2014/15, there is limited evidence that this growth is itself responsible for worsening learning outcomes, as opposed to being a symptom of wider problems in the education system in some states.

The results are broadly in line with findings on similar programmes in other countries. A review of 21 'structured pedagogy' programmes – programmes based on changes to curricula or instructional approaches along with lesson plans and training for teachers – finds, on average, that hey improve language scores by 0.23 standard deviations, and mathematics scores by 0.14 standard deviations (Snilstveit *et al*., 2016). These programmes involve a wide range of contexts and implementation methods.

The ESSPIN results partly reflect the inherent difficulty in controlling properly for confounding differences between states without also controlling away some of the differences in the pattern of intervention. We are not completely able analytically to separate ESSPIN's intervention from other unmeasured differences between states – such as, for example, the policy environment and

functioning of the education system at the state level. However, at least controlling for measured school characteristics and learners' socioeconomic status, there remain positive and statistically significant effects on pupils' learning as a result of ESSPIN's intervention. This is the case using both cross-sectional and time-series analysis, and using both matching and conventional regression techniques.

7 Conclusion and implications

This report has examined trends over time for all schools in the six states, and evaluated the impact of ESSPIN's intervention, using a range of indicators relating to school management, inclusion, teaching, and learning outcomes. The analysis of ESSPIN's impact finds that the programme has had a positive impact on school management, inclusion, teacher competence, overall school quality and learning outcomes. The analysis of trends over time finds marked improvements across the states as a whole in many aspects of school management and inclusion, in a measure of overall school quality, and in the numeracy of learners in Grade 4. However, teachers' competence and content knowledge, and Grade 2 test scores, have worsened over time.

High proportions of schools across the six states still do not meet ESSPIN's standards for a good school. In 2016, 18% of schools meet the standard on head teacher effectiveness; 19% on school development planning; 11% on inclusion; 44% on functional SBMCs; and 18% or 4% on overall school quality (depending on which indicator we use).

In terms of school management and overall quality, there have been marked improvements over time since 2012, particularly in terms of school development planning, SBMC functionality and inclusiveness, and overall school quality. Moreover, schools which have received more years of ESSPIN intervention have more effective head teachers in 2016, are better at school development planning, are more inclusive, and are much more likely to have well-functioning SBMCs in which women and children participate. The estimated effect of a year of full ESSPIN intervention on the proportion of schools that meet each standard ranges from 5 percentage points for head teacher effectiveness to 19 percentage points for SBMC functionality. ESSPIN's intervention has a significant positive effect on our overall measure of school quality. A year of intervention is associated with an improvement of 4–6 percentage points in the school quality score.

Teachers' competence appears to have worsened between 2012 and 2014, and then recovered in 2016, but the net result is no significant improvement between 2012 and 2016. Teaching behaviours such as the use of teaching aids and praising students have increased over time, but teachers' scores in content knowledge tests in English and mathematics have worsened since 2014. Most teachers are able to complete simple primary school-level tasks, such as writing a simple sentence in English or word problems for addition in mathematics, but stumble with more advanced tasks such as extracting basic information from a passage or using a number line to represent sums. Teachers trained through ESSPIN are significantly more competent, and less likely to be in the lowest performance bands, but even among these teachers, only a small proportion make it to the highest performance bands.

As well as having issues with content knowledge, teachers are often not in attendance, even though the schools that participated in this survey had prior notice of our visit. On average, around 80% of teachers were present on the day of the survey visit, according to the head teacher's own records, and around 70% of classrooms were observed as having both teacher and learners present at the expected time in the morning. There has been little change in the proportion of teachers and learners present on time since 2012. This suggests that schools continue to have serious problems with teacher and learner attendance at the start of the school day.

We find mixed results in learners' test scores: Grade 4 numeracy has improved significantly since 2014, while Grade 4 literacy has not changed significantly, and Grade 2 literacy and numeracy have worsened significantly. Learning outcomes, particularly in literacy, are better in schools that have received more years of ESSPIN intervention. We estimate that ESSPIN's intervention has an effect on learning outcomes of around 0.1 standard deviations. In some cases effects become non-significant when we add controls for the state that the school is located in. It is difficult completely

to disentangle the effects of ESSPIN's intervention from unmeasured state factors, such as the policy environment – and probably variation between states in how well the programme has been implemented. However, the effects are robust to controlling for school characteristics such as urban location and facilities, students' socioeconomic background, and learning outcomes at baseline (in 2012). Despite some ambiguity around the state effects, the evidence overall points towards a genuine casual impact of ESSPIN on learning outcomes. The magnitude of the impact is roughly in line with typical effects found for comparable programmes in other countries.

If ESSPIN's intervention has been rolled out to all schools, and the intervention works in improving learning outcomes, then why have learning outcomes remained flat or fallen (except in Grade 4 numeracy) across the states as a whole? The answer appears to lie partly in the duration and timing of intervention, as measured by the years in which each school received a full package of leadership training, teacher training, and school visits. The political and financial context has reportedly constrained the effective implementation of ESSPIN programming, with many states struggling to finance salary and development expenditure, and limited state funding of the SIP in several of the states. Rapid rises in enrolment and PTRs may also present challenges for schools, although we were not able to find strong evidence of a causal link between PTR increases and worsening average learning outcomes. The absorption of over 800,000 additional children in the school system within five years (assuming census results are accurate), in resource-poor and conflict-affected school systems, itself represents a significant achievement in terms of equitable access to education, as well as potentially explaining some of the profiles of learning outcome distributions being stubbornly crowded into the lower bands.

Effects on learning outcomes are weakest for those schools that had a short duration of intervention and for those where the intervention stopped more than a year ago, suggesting that the effect of training on a school may fade out over time. Many schools in the six states continue to fall into these categories, and have seen limited improvements in learning outcomes. Learning outcomes are best in the schools that have had relatively consistent intervention over three or more years up to the present. This has important implications for the continuity and sustainability of ESSPIN programming, suggesting that gains made in learning outcomes may be lost if there is a gap in provision.

The gains in average learning outcomes from ESSPIN's intervention, though statistically significant, are modest in magnitude. This may reflect some of the constraints described above in terms of teachers' content knowledge and attendance in the classroom. Again, ESSPIN has a significant but modest effect on these constraints, and they are likely to remain key barriers to larger improvements in learning outcomes.

References

Allen, R. (2016a) CACSA report ESSPIN2.3: ESSPIN scaled scores and subscores – Literacy. Unpublished note. Curriculum, Assessment and Certification Systems Architects.

Allen, R. (2016b) CACSA report ESSPIN3.1: ESSPIN teacher survey – Notes on subscales (updated). Unpublished note. Curriculum, Assessment and Certification Systems Architects.

Allen, R. (2016c) CACSA report ESSPIN3.5: ESSPIN scaled scores and subscores – Numeracy. Unpublished note. Curriculum, Assessment and Certification Systems Architects.

Allen, R. (2016d) CACSA report ESSPIN4: ESSPIN teacher literacy and numeracy – Notes. Unpublished note. Curriculum, Assessment and Certification Systems Architects.

Bennell, P., and Akyeampong, K. (2007). Teacher Motivation in Sub-Saharan Africa and South Asia (Department for International Development: Educational Papers). DFID.

Cameron, S. (2015a) Composite Survey 2: Overall technical report. Oxford Policy Management.

Cameron, S. (2015b) Teacher motivation in the TDP baseline survey. Unpublished technical note. Oxford Policy Management.

Calder, J. (2016) Education and Child Protection. ESSPIN report 439.

Coinco, E. (2014) Education, Conflict and Violence Research. ESSPIN report 428.

De, S. and Cameron, S. (2015) ESSPIN Composite Survey 2: Gender and Inclusion Report. Oxford Policy Management.

De, S., Pettersson, G., Morris, R., and Cameron, S. (2016) Teacher Development Programme (TDP) Impact Evaluation of Output 1: In-Service Training. Final Baseline Technical Report, Volume I. Results and Discussion. Abuja: Education Data, Research and Evaluation in Nigeria (EDOREN).

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.

Mezger, C. (2016) Trends of inequality in access to schooling in Nigeria. Evidence from three DHS surveys (2003, 2008, 2013). Abuja: Education Data, Research and Evaluation in Nigeria (EDOREN).

Pellens, T., Outhred, R., Majeed, Z., Kveder, A., Binci, M., Wallin, J., Kelleher, F., Beavis, A. and Rai, S. (2016) Evaluation of the Girl's Education Project 3. Baseline technical report prepared by EDOREN on behalf of UNICEF GEP3. Abuja: Education Data, Research and Evaluation in Nigeria (EDOREN).

Raleigh, Clionadh, Andrew Linke, Håvard Hegre and Joakim Karlsen (2010) Introducing ACLED-Armed Conflict Location and Event Data. *Journal of Peace Research* 47(5) 651-660.

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).

Sanni, K, 2015, Taking School Improvement to Scale: The Education Sector Support Programme in Nigeria, Cambridge Education.

Snilstveit, B., Stevenson, J., Menon, R., Phillips, D., Gallagher, E., Geleen, M., Jobse, H., Schmidt, T., and Jimenez, E. (2016). The impact of education programmes on learning and school

participation in low- and middle-income countries (Systematic Review Summary No. 7). 3ie. Retrieved from http://www.3ieimpact.org/media/filer_public/2016/09/20/srs7-education-report.pdf.

Steenbergen, V., Nnodu, I., Nwachukwu, T., and Nyager, S. (2016) Public financing of basic education in Nigeria. An analysis of government budgeting and expenditures on basic school improvement in ESSPIN programme states. Oxford Policy Management.

Annex A ESSPIN Output Stream 3 interventions

The table below shows the ESSPIN Output Stream 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 full intervention received before 2015/2016. Full intervention means the school received some leadership training, some teacher training, and some school visits during the year, though the amount of each may vary. The schools have been grouped as follows: minimum (zero to one years), medium (two to three years), maximum (four to five years).

	Category (years of intervention)	2009/2010			2010/2011			2011/2012			CS1	2012/2013			2013/2014			CS2	2014/2015			2015/2016			CS3
		L	T	SV	L	T	SV	L	T	SV		L	T	SV	L	T	SV		L	T	SV	L	T	SV	
Enugu	Minimum (1)																		3	2	9	3	2	9	
	Medium (2)														6	3	9		3	2	9	3	2	9	
	Medium (3)							6	3	9		6	3	9	3		9		3	2	9	3	2	9	
Jigawa	Minimum (1)																		8	3	16	8	3	14	
	Medium (2)														6	3	9		20	6	14	3		14	
	Medium (3)											6	3	9	6	3	9		4	6	14	3		14	
	Medium (3)	5*	5*	9*	10*	5*	9*						3	9		3	9		4	6	14	3		14	
Kaduna	Minimum (1)																		8	2	3				
	Minimum (1)														6	3	9				9				
	Medium (2)											6	3	9	6	3	9				9				
	Medium (2)							6	3	9		6	3	9	3		9				9				
	Maximum (4)	5*	5*	9*	10*	5*	9*	6	3	9		6	3	9	3		9				9				
Kano	Minimum (1)														9	9	9			4	6		20	8	
	Medium (2)											6	3	9	9	9	9			4	6		20	8	
	Medium (3)	5*	5*	9*	10*	5*	9*								9	9	9			4	6		20	8	
Kwara	Maximum (4)	6	3	30	6	3	30			30				30	6	3	30		6	3	30	2	4	15	
Lagos	Medium (3)											6	3	9	6	3	9		6	6	30	4	6	30	
	Medium (3)							6	3	9		6	3	9	3		9		6	6	30	4	6	30	
	Maximum (5)	5*	5*	9*	10*	5*	9*	6	3	9		6	3	9	3		9		6	6	30	4	6	30	

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

Annex B ESSPIN Output Stream 4 interventions

The table below shows the days of Output stream 4 intervention under different headings: SBMC training; women and children participation training; and mentoring visits. The schools have been grouped as follows: 'no intervention' (5 days or less of intervention received so far), 'post-CS1' (started receiving intervention in 2012/13 or after), 'pre-CS1' (started receiving intervention in 2011/12 or earlier)

	Category (years of intervention)	2010/2011			2011/2012			CS1	2012/2013			2013/2014			CS2	2014/2015			2015/2016			CS3
		S	P	M	S	P	M		S	P	M	S	P	M		S	P	M	S	P	M	
Enugu	No intervention																		1			
	Post-CS1											7		2		5	4	5				
	Pre-CS1				7		4		1				6	4		3	2	8				
Jigawa	No intervention																	5				
	Post-CS1 (a)											2				2		2	2	2	1	
	Post-CS1 (b)											7		2		2	2	4	2	3	4	
	Pre-CS1	7		4	1		4		6	4			4	6		2	2	3	1	2	4	
Kaduna	No intervention																					
	Post-CS1											7									5	
	Pre-CS1 (a)				7		4		1		4		6	4		5	2	8	1	2	10	
	Pre-CS1 (b)	7		4	1		4			6	4			4		5	2	15	1	2	17	
Kano	No intervention															2						
	Post-CS1 (a)											7		4		1	2	6	1	2	4	
	Post-CS1 (b)								1	6	4			4		3	2	8	1	2	4	
Kwara	No intervention																					
	Pre-CS1				7		4		1		4		6	4		5	5	16	1	2	17	
	Post-CS1 (a)											4		2		5	5	6	1	2	7	
	Post-CS1 (b)											7		1		5	5	5	1	2	6	
Lagos	Post-CS1 (a)								7	6		1	6	4		1	2	4	1	2	4	
	Post-CS1 (b)											7		4		1	2	4	1	3	4	

S = SBMC training; P = women's and children's participation training; M = mentoring visits