OECD Working Papers on Fiscal Federalism No. 22

Improving the Performance
of Sub-national
Governments through
Benchmarking
and Performance Reporting

Leah Phillips

<https://dx.doi.org/10.1787/ffff92c6-en>

OECD NETWORK ON FISCAL RELATIONS ACROSS LEVELS OF GOVERNMENT (THE “FISCAL NETWORK”)

FISCAL FEDERALISM WORKING PAPER SERIES

This series is designed to make available to a wider readership selected studies drawing on the work of the OECD’s Network on Fiscal Relations across Levels of Government. Authorship is usually collective, but principal writers are named. The papers are generally available only in their original language (English or French) with a short summary available in the other.

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed are those of the author(s).

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcomed, and may be sent to the OECD Network on Fiscal Relations, 2 rue André-Pascal, 75775 Paris Cedex 16, France. This working paper has been authorised for release by the Director of the Policy Studies Branch of the Economics Department, Luiz de Mello.

Comments on the series are welcome, and should be sent to either fiscalnetwork@oecd.org or the OECD Network on Fiscal Relations, 2 rue André Pascal, 75775 PARIS CEDEX 16, France.

Applications for permission to reproduce or translate all, or part of, this material should be sent to OECD Publishing, rights@oecd.org.

Copyright OECD 2018.

ABSTRACT

Improving the performance of sub-national governments through benchmarking and performance reporting

Performance systems are one tool available to central governments to improve the performance of sub-national service delivery. This paper provides a preliminary review of suitable metrics and mechanisms to reliably measure and monitor the efficiency and quality of public services that are provided by sub-national governments. This review aims to conceptualise the challenges associated with measuring public sector inputs, outputs and outcomes and implementing effective performance systems. Robust performance systems aim to measure both the efficiency and effectiveness of public services, as well measure cost efficiency, to better understand and remedy cost discrepancies across regions. Implementing qualitative mechanisms such as external inspections and user surveys are also useful in providing insights into consumer experience and well-being.

Keywords: benchmarking systems, public sector productivity, sub-national government performance

JEL Codes: H44, H77, O43

* * * * *

RÉSUMÉ

Améliorer les performances des administrations infranationales grâce aux analyses comparatives et à la remontée d'information sur les performances

Les systèmes de gestion des performances font partie des outils sur lesquels les administrations centrales peuvent s'appuyer pour améliorer la prestation des services publics à l'échelon des administrations infranationales. Le présent rapport propose un examen préliminaire des éléments de mesure et des dispositifs pouvant permettre de mesurer et de suivre de façon fiable l'efficacité et la qualité des services publics assurés par les administrations infranationales. Le présent examen a pour objet de conceptualiser les difficultés liées à la mesure des intrants, des produits et des réalisations du secteur public et à la mise en place de systèmes efficaces de gestion des performances. Un système robuste de gestion des performances vise à mesurer à la fois l'efficacité et l'efficacité des services publics, mais aussi l'optimisation des ressources, de façon à mieux comprendre et traiter les disparités de coût selon les régions. Il est également utile de mettre en œuvre des dispositifs qualitatifs tels que des inspections externes et des enquêtes auprès des usagers, pour obtenir des indications sur le vécu des usagers et leur bien-être.

Mots-clés : systèmes d'analyse comparative, productivité du secteur public, performances des administrations infranationales

Classification JEL : H44, H77, O43

TABLE OF CONTENTS

IMPROVING THE PERFORMANCE OF SUB-NATIONAL GOVERNMENTS THROUGH BENCHMARKING AND PERFORMANCE REPORTING	4
1. Introduction and main findings	4
2. The effects of benchmarking in the sub-national government context	6
2.1. Benchmarking can be a tool to improve the performance of public services	6
2.2. Limitations of benchmarking	8
2.3. The effect of decentralisation on benchmarking	8
2.4. Explicit monetary rewards	11
3. Approaches to benchmarking	13
3.1. Challenges with measuring outcomes	13
3.2. Cost approach to benchmarking	17
3.3. The use of composite indicators to benchmark multiple metrics	19
3.4. Quality assurance systems	20
3.5. Consumer experience and satisfaction surveys	22
4. Avenues for further research	24
REFERENCES	25
ANNEX A. AUSTRALIA'S REPORT ON GOVERNMENT SERVICES	30
ANNEX B. DENMARK'S HEALTH AND EDUCATION PERFORMANCE SYSTEMS	34
ANNEX C. INDIA'S HEALTH PERFORMANCE INDEX	39
ANNEX D. NORWAY'S KOSTRA SYSTEM	42
ANNEX E. THE USE OF PUBLIC PERFORMANCE DATA IN THE UNITED KINGDOM'S SCHOOL SYSTEM	45
ANNEX F. THE UNITED STATES' NO CHILD LEFT BEHIND ACT, 2001-2015	48

Figures

1. The relationship between performance, efficiency and effectiveness	5
2. Measurement of public sector productivity at different levels	14
A1. Generalised performance indicator framework	31
A2. RoGS School education performance indicator framework	32

Boxes

Box 1. The concepts of productivity, efficiency and effectiveness	5
Box 2. Allocative efficiency gains due to decentralisation	7
Box 3. Quality measures for school education in England	15

IMPROVING THE PERFORMANCE OF SUB-NATIONAL GOVERNMENTS THROUGH BENCHMARKING AND PERFORMANCE REPORTING

By Leah Phillips¹

1. Introduction and main findings

1. Productivity is widely considered to be the most important long-term driver of economic growth and a main contributor to increasing well-being (OECD, 2015). Slowing productivity growth in most countries has raised widespread alarm. While the focus has been primarily on the private sector, the public sector is a large contributor to GDP in most OECD countries (OECD, 2017a). Given the public sector's central role in the provision of many services, a better understanding on how to make improvements in the performance of public services is essential. This paper characterises performance improvements as increases in both the efficiency and effectiveness of public service delivery, including improving the mix and use of inputs, as well as enhancing the quality of outputs to achieve better outcomes for the whole population (Box 1). Given the general absence of market prices in the public sector, the measurement of efficiency and effectiveness is fraught with challenges and require the government to deploy a range of information and accountability levers to overcome them.

2. The trend towards decentralisation of government and the ensuing dispersion of power often results in sub-national governments being responsible for the delivery of key government services, including education, health, law and order, and social services. However most central governments still see it as their role to ensure public services are delivered efficiently and equitably, for a range of economic, social and financial reasons. In this context, national governments implement systems to measure and influence the performance of sub-national service delivery, including frameworks based on financial rewards and reputational effects. A previous paper (Mizell, 2008) produced for the *OECD Fiscal Network* explored this topic. Mizell (2008) used information collected from a 2007 Fiscal Network questionnaire to investigate how OECD countries monitor sub-national governments through indicator systems. This paper aims in part to update the findings presented in Mizell's paper, and also to provide further insights into how performance benchmarking can be achieved and better understood.

3. For benchmarking, good quality data are required, to allow for the calculation of reliable indicators – metrics aimed at determining the equity, efficiency and effectiveness of public sector services. These metrics usually require data on the inputs, outputs and quality of public services. A general framework for measuring all aspects of performance is shown in Figure 1. In addition to clear metrics, transparent mechanisms are needed so that information on public sector performance can be disseminated to service providers, government officials and the public.

1. This paper was prepared while the author was a consultant to the OECD Network on Fiscal Relations Across Levels of Government, on leave from the Australian Treasury. It incorporates feedback from delegates at the 2017 meeting of the Fiscal Network. Comments from Sean Dougherty, Peter Hoeller, Hansjörg Blöchliger, El Iza Mohamedou, and Douglas Sutherland were valuable in revising the paper. Technical assistance from Celia Rutkoski was most appreciated.

Box 1. The concepts of productivity, efficiency and effectiveness

This paper characterises performance improvements as increases in both the efficiency (or productivity) and the effectiveness of public service delivery. Service providers that lie at the production possibility frontier are seen as achieving 'best practice', which alludes to the concept of efficiency (Mandl et al., 2008). When measuring efficiency, a distinction can be made between technical efficiency ('doing more with less') and allocative efficiency ('doing the right thing, at the right place').

Technical efficiency characterises a production process where the maximum possible output has been achieved, given a fixed set of inputs and given a certain technology. Allocative efficiency refers to the allocation of resources (either in financial terms or in labour or physical/capital terms) and occurs when the input-output combination is cost-minimising and/or profit-maximising (Lafortune, 2015; OECD, 2001).

The concept of technical efficiency is similar to the notion of productivity. Productivity is commonly defined as a ratio of a volume measure of output to a volume measure of input use. Productivity is much easier to compute if the unit of production that is being analysed uses a single input to produce a single output. If the production unit uses several inputs to produce several outputs, then the inputs and outputs must be aggregated (Daraio and Simar, 2007).

Efficiency is difficult to measure as it requires knowledge on the sector's production possibility frontier – that is, which providers are performing under 'best practice' conditions. In practice, most indicators actually measure productivity, since the best possible result that is achievable is not taken into account (Mandl et al., 2008).

When conceptualising the measurement of public service productivity, the integral role of the consumer to the production process needs to be considered. In contrast to the classical concept of productivity, which generally occurs in a closed system, customer satisfaction should ideally be integrated and accounted for in the concept of services productivity (Balci et al., 2011). In contrast to efficiency which is the ratio of outputs and inputs, effectiveness is the ratio of defined outcomes to defined inputs (see Figure 1), and conditional on the quality of service provision. Improving service quality and outcomes across all population groups is a key policy priority in all OECD countries (OECD, 2017b). In this paper, access to public services is seen as a desired outcome of public service delivery, and is therefore classified as one aspect of effectiveness. Effectiveness is viewed here as a critical component of performance of service delivery.

Figure 1. The relationship between performance, efficiency and effectiveness

Source: OECD Secretariat.

4. Several key findings emerge from this paper:

- Performance systems are one tool for central governments to improve the efficiency and effectiveness of and access to sub-national services. Performance systems accomplish this by reducing information asymmetries between different levels of government, identifying providers that are over or under-performing, or by stimulating competition between sub-national governments. When designing performance systems, collaboration across levels of governments is necessary to construct relevant metrics, which may also reduce the incentives of sub-national providers to game the system.
- Performance systems that aim to create competition between sub-national governments (through transparency of performance information) may be more applicable for countries with strong, centralised governments. A more collegiate or collaborative form of benchmarking which is less likely to rate or rank participants will be more amenable to sub-national governments with greater revenue power and administrative responsibilities.
- Measuring the output and quality of public services presents many challenges. However, capturing the quality of services is an integral aspect that requires further statistical work. Although they have weaknesses, composite indicators can help simplify vast amounts of information into an easily digestible framework. Qualitative mechanisms in the form of external inspections and user surveys are useful in providing insights into consumer experience and well-being. Governments should ensure that performance systems are implemented that measure both the efficiency and effectiveness of public services.
- Performance systems that aim to measure and compare costs across jurisdictions are helpful to ensure that services are cost-efficient and to better understand cost discrepancies across regions. However, cost adjustments should generally be made to remove the effect of external or geographical differences, which can better ensure the accurate portrayal of cost differences.

5. This paper looks at a range of benchmarking practices, including those focused on costs, outputs and outcomes, in order to capture different aspects of performance and productivity. The second section of this paper provides an overall summary of the advantages and limitations of benchmarking systems and analyses the effects of financial and non-financial rewards for good performance. The third section reviews the various practices used to measure performance and productivity: measuring the costs of a service, and quantifying the outputs or outcomes of a service. It also analyses benchmarking *mechanisms*, including composite indicators (i.e. league tables and star ratings), survey-based approaches and external inspections of service providers. The paper concludes with areas where further research would be warranted. The annexes provide more detailed case studies of benchmarking systems in selected OECD countries, namely for Australia, Denmark, India, Norway, the United Kingdom and the United States.

2. The effects of benchmarking in the sub-national government context

2.1. Benchmarking can be a tool to improve the performance of public services

6. Benchmarking systems involve the comparison of one government or service provider's performance on a set of measurable indicators, with a strategic benchmark. Developing benchmarking systems is an iterative process. The steps involved in benchmarking are to: capture information; generate meaningful comparisons or case studies; and distribute information throughout a network of actors, often across multiple levels of government.

7. Although there are various reasons to benchmark, the main aim is to identify weak points in order to improve services and to gain knowledge on best-practice providers (Kouzman et al., 1999). Benchmarking can be used to:

- Assess public sector performance objectively;
- Reveal areas where improvement is needed;
- Identify regions or providers with better performance and improve the dissemination and adoption of ‘best practice’;
- Test whether pilot projects and certain technologies have been successful;
- Help in the drawing up of government budgets at the central and sub-central levels; and
- Make governments and service providers more accountable to citizens and underpin yard-stick competition.

8. Benchmarking can also correct co-ordination failures that are likely to be exacerbated by decentralisation. With decentralisation comes a larger number of stakeholders and greater complications in intergovernmental fiscal relations, which can intensify information gaps between levels of government. The creation of multiple public agencies can also make co-ordination more complex, hindering the central government’s ability to steer public policy in the desired direction (Sørensen, 2014). Benchmarking can help eliminate information and coordination gaps, create a common frame of reference and enhance the services delivered by regional governments.

Box 2. Allocative efficiency gains due to decentralisation

Since the early 1990s there has been a substantial transfer of spending responsibilities, particularly in education and health care, to sub-national governments in many OECD countries. On average, around 40 per cent of overall public expenditure in OECD countries occurs at the state and local level (OECD, 2017a). There are benefits from decentralisation. From a theoretical perspective, decentralisation of spending and political responsibilities to sub-central governments can result in allocative efficiency gains. This rests on the idea of subsidiarity, where responsibility should lie with the lowest level of government that is capable of performing the function effectively. Adam et al. (2014) investigated the effect of fiscal decentralisation on (overall) public sector efficiency, using data envelopment analysis for 21 OECD countries over 1970-2000. They concluded that fiscal decentralisation is beneficial for efficiency but only up to a certain point - this creates an inverted U-shape relationship between public sector efficiency and fiscal decentralisation.

Sub-national governments may find it easier to engage with their communities, and be more flexible and targeted in their approaches to improving outcomes for citizens, as decentralised decision-making produces non-uniform “tailored” policies that reflect local preferences (Koethenbueger, 2008). Although empirical findings are mixed, it is generally found that there is a positive relationship between spending decentralisation and economic activity (Blöchliger and Égert, 2013). Many of the economic benefits of decentralisation stem from competition among sub-central governments. Revenue and expenditure decentralisation creates competition among regional governments for mobile, tax-paying citizens, which have an incentive to move to regions that offer services and taxes that match their preferences (Tiebout, 1956). Regional governments benefit from their greater knowledge of local preferences and cost conditions, which a central agency is unlikely to possess (Oates, 1999).

The quality of many public services run by regional governments also directly affects citizen well-being and long-term economic growth. For example, the quality of the education system affects the acquisition of cognitive skills and the degree of human capital accumulation, and thus long-term economic growth (Dumciuviene, 2015). In this context, many central governments find it imperative to monitor the level and effects of sub-national government spending. Benchmarking is one tool that is useful for increasing the transparency and accountability of government spending, and monitoring resource allocation.

9. If benchmarking is accompanied with public transparency, it can also create competition between regional governments and service providers and thus lead to efficiency gains (Box 2). Outsourcing and private sector provision exist in some markets traditionally controlled by public agencies. However, in the absence of competition at the local level market forces cannot drive out inefficient providers. Moreover, the lack of incentives that are present in competitive markets may reduce innovation and cause higher costs in public markets. Benchmarking can imitate competitive pressures among public organisations and governments, especially across providers that exhibit similar demographic or geographical characteristics. If this information is public, performance systems can help voters to judge performance. This can put pressure on politicians to ensure that local services are cost-efficient and of high quality (Sørensen, 2014).

2.2. Limitations of benchmarking

10. Benchmarking should lead to superior economic performance and service delivery. However, implementing a benchmarking system with appropriate metrics and incentives is difficult, with improper implementation resulting in the deficiencies discussed below.

11. The effectiveness of a benchmarking system is influenced by the selection of indicators that are used to benchmark service providers, which if selected carelessly, can distort the incentives and policy objectives of participants. If a performance system focuses on one specific output, input, process or area of service delivery, attention to monitor and measure that activity will likely increase and so will that activity. This may be beneficial in itself, but it will create issues if this occurs at the expense of some other activity which is also important but which is not included in the benchmarking system. Further, the greater number of stakeholders affected by the benchmarking system, the more difficult it will be to reduce the amount of perverse incentives created by a uniform benchmarking system. It will also be even more difficult to decide which metrics should be used (Curristine et al., 2007).

12. Information asymmetries can occur between those levels of government that are being monitored and the central government responsible for oversight. These asymmetries can create inefficiencies in the performance system or make them ineffective. Information asymmetries are especially prevalent in decentralised systems when sub-national governments are not consulted about the form of national government oversight.

13. Information asymmetries make it more likely that monitored public service providers will game the system. “Gaming” refers to strategic behaviour intended to ensure positive performance results (OECD, 2009) resulting in a distortion of policy decisions or resource allocation, and is more likely to occur under a competitive benchmarking approach, since there are stronger incentives in terms of a reputational or monetary reward. Gaming could be in the form of outright deception, through providing false information to central governments. Alternatively, it could result in public service organisations meeting or even exceeding the specified targets but at the expense of other less measurable aspects of service delivery, like safety or customer satisfaction. Historically, there is evidence that gaming or sub-optimal results have taken place including: meeting waiting-list targets at the expense of clinical priorities (NAO, 2001) and trains being more ‘punctual’ by missing stations (SSRA, 2000). Unfortunately, due to information asymmetries, it is difficult to identify gaming without the benefit of hindsight. However, ensuring a performance system has a sufficient number of indicators, verifying data (if possible) and revising indicators on a regular basis can help reduce gaming amongst participants.

2.3. The effect of decentralisation on benchmarking

14. The type of decentralisation (that is, administrative, fiscal or political decentralisation) as well as the extent of decentralisation, can affect what type of benchmarking system will be most effective. Unitary and federal governments have different opportunities for fiscal decentralisation. Federal governments have

constitutionally protected sub-national governments and share public power across governments. Given this, the possibility of autonomous decision-making at the sub-central level is obviously stronger. Quasi-subordinate levels in unitary countries have no constitutional powers or responsibilities, and can only exercise the powers that the central government delegates, leaving greater scope for intervention by central governments. That said, decision-making power and responsibilities for public services by sub-national governments vary widely across countries, and does not necessarily depend on a country's constitutional background. Political decentralisation generally requires constitutional or statutory reforms, the development of local political parties and systems as well as the strengthening of legislatures and regulations in order to categorise the formal responsibilities of a decentralised system (World Bank, 2001). Political decentralisation aims to devolve power to lower level governments in order to give citizens more influence over government decision-making. Administrative decentralisation aims to increase the efficiency of the delivery of public goods and services by redistributing authority, responsibility and financial resources from the central government to regional or local governments, or semi-autonomous public sector organisations.

15. Collegiate (or collaborative) benchmarking is based on learning from best practices, as opposed to using 'naming and shaming' techniques. It generally involves consultation and collaboration between levels of government. Collegiate benchmarking is unlikely to involve the public dissemination of 'high-stakes' performance information meaning that sub-national governments may be more likely to participate in these forms of benchmarking and share their experiences. Collegiate benchmarking will more likely be implemented if all levels of government perceive that it will lead to new or better information channels, improve policy effectiveness, or if they can share the additional resources and political leverage (OECD, 2009). Central governments can play an important role in facilitating knowledge and data sharing across regional governments and consolidating data sources.

16. A performance system based on collegiality is more likely to be amenable to regional governments in a federal, decentralised system. In Australia for instance, the central and regional governments undertake a collaborative exercise to produce an annual report on the performance of sub-national service delivery. The national government has not mandated formal benchmarks, and the Australian States and Territories (States) have resisted the publication of any summary information that ranks or rates them, which may have facilitated the participation by all States. More information on Australia's performance system can be found in Annex A. Benchmarking may also be voluntary rather than imposed by the central government, which may be more prevalent if sub-national governments are not heavily reliant on the central government for funding. For example, performance comparisons across local authorities in Germany are largely voluntary and rely on self-management. The Association of Local Government Management (KGSt), for instance, encourages voluntary participation from districts and municipalities, with the KGSt establishing voluntary benchmarking networks and a Common Assessment Framework to assess performance across governments (Kuhlmann and Jäkel, 2013).

17. Competitive benchmarking aims to generate competition amongst regional governments by disseminating information and thus facilitating comparisons by citizens and voters. In fact, the performance monitoring of public services in many countries seems to have moved from using performance information to improve organisational processes towards transparency mechanisms to enhance accountability. While transparency of public services has increased, the availability of information is only useful if it actually achieves the long-term objective of improving the efficiency and effectiveness of or access to public services.

18. Some research suggests that this public dissemination of information is likely to improve service delivery. Braadbaart (2007) looked at the effect of benchmarking on transparency and economic performance in the Netherlands' water supply industry. The study found that benchmarking immediately enhanced transparency, but only affected economic performance after benchmarking information entered the public domain. Bevan and Hamblin (2009) studied the effects of an English reform that targeted the

transit time of ambulance services. This target was used as part of a ranking system of ambulance services which inflicted reputational damage on services that failed to hit targets. When compared to other United Kingdom countries that did not implement these targets in their performance systems, Bevan and Hamblin (2009) showed that only in England was this target met, suggesting that systems that are designed to inflict reputational damage on those that have performed poorly are more likely to improve performance. Annex E provides more information on the public dissemination of information in the United Kingdom, namely with regards to education.

19. Countries with relatively centralised government spending and taxation can more easily impose a system of competitive benchmarking to monitor sub-national service delivery (Kuhlmann and Jäkel, 2013). Competitive benchmarking generally involves centralised measurement of performance and the use of public funds. However, a top-down approach is likely to result in the benchmarking indicators being less reflective of regional or local specificities, due to knowledge asymmetries. A benchmarking system implemented without genuine consultation with the other levels of government will not be desirable. Regional service providers may comply with the reporting requirements or national standards but fail to use the framework to raise performance or quality. Thus, even competitive benchmarking frameworks require genuine buy in from all levels of government. A case study of the United States' *No Child Left Behind Act* is provided in Annex F, which identifies the challenges with implementing a competitive benchmarking framework in a decentralised federal system.

20. Competitive benchmarking schemes generate explicit incentives as they exert reputational effects that generate external pressure for accountability and reform (OECD, 2009). Competition aims to improve government behaviour, as providers aim to improve their performance to avoid being labelled as poor or failing organisations (Wilson, 2013). Benchmarking systems can be used as a political tool. This is beneficial if these frameworks are used to support evidence-based policy making, to achieve or improve public policy outcomes. However, under adversarial political settings (including across levels of government), the results from benchmarking systems may be interpreted and communicated differently across opposing political parties, which can hamper performance improvements. In countries with political decentralisation and intense political competition, the publication of negative performance information about sub-governments could result in significant political backlash (Kuhlmann and Jäkel, 2013). At the same time, local elections may impede the publication of performance data, at least when poor results are anticipated (Bogumil, 2003). Political polarisation and citizens' distrust in public institutions make the success of benchmarking system more unpredictable (OECD, 2017b). Given the long-term nature of most public service sector reforms, robust political commitment across all levels of government and across party lines is required to ensure benchmarking systems fulfil their long-term objective of increasing productivity.

21. Benchmarking systems can also vary widely in their objectives, which can affect how the framework should be designed and implemented. Certain objectives, such as monitoring the achievement of national standards or targets, or determining budget allocations, may be well aligned with benchmarking frameworks in which the central government plays the dominant role. Objectives that emphasise achieving regional goals, inter-governmental learning or capacity building, may be better achieved by engaging sub-national actors in the design, implementation, and use of indicators and reporting frameworks (OECD, 2009).

22. Countries with large local authorities are more capable of coping with the transaction and opportunity costs that underlie any benchmarking system. They are more likely to benefit from economies of scale, with a more extensive human resource pool from which to draw and enough organisational freedom to use comparative performance information effectively (Kuhlmann and Wollmann, 2006; Kuhlmann and Jäkel, 2013). For example, in Sweden, there are 21 county councils; but studies show that reducing the amount of counties to six would create conditions for a more effective pooling of resources, especially for health care services (Blomqvist and Bergman, 2010). Denmark successfully merged municipalities in 2007, reducing the total number to around 100, from 300, and the number of councils to

five, from 14. One of the main drivers of this reform was to reach a more adequate size for health care service provision (OECD, 2015). Further information can be found in Annex B.

2.4. Explicit monetary rewards

23. Central governments sometimes utilise rewards and sanctions on sub-central governments, to incentivise service providers to improve performance, by linking access to and the size of transfers with performance in predetermined service areas (FFC, 2013). This is a type of performance budgeting which directly links performance results to resource allocation (Curristine et al., 2007). The literature provides plenty of evidence that performance-based transfers improve service delivery and accountability (Goddard and Mannion, 2004).

24. There are two main types of explicit reward systems for sub-national governments: financial and administrative. The aim of both is to stimulate effort by local and regional governments to meet certain objectives or standards. Financial incentives refer to the availability of funds based on meeting specified performance targets. Administrative incentives are changes to rules and regulations that affect regional governments, such as a relaxation of budgetary rules or reduced central government oversight (OECD, 2009). The information below focuses on the former, but the benefits of administrative rewards are likely to be similar.

25. The funding of rewards is expensive for central governments, and as such, should only be implemented if other more cost-effective mechanisms do not result in the desired performance gains. If intrinsic motivation in a certain policy or service area is high, incentives through reputational sanctions and rewards may be more effective. Providing financial rewards can create doubt as to the true motive for which good deeds are performed, and this ‘over-justification effect’ can result in a crowding out of pro-social behaviour by intrinsic incentives, reducing the total contribution provided by service staff (Bénabou and Tirole, 2006). In the presence of explicit rewards, agents may conclude that their previous willingness to put in additional effort is not needed as an explicit reward scheme ensures that what needs to be done will actually get done. By comparison, the public dissemination of information may act to strengthen intrinsic motivation, where behaviour is driven by internal rewards. In the education sphere for instance, many argue that most teachers are primarily motivated by an intrinsic desire to teach effectively and recognition of good performance can strengthen this desire. The United Kingdom, for instance, organises events for head school teachers that are judged to be outstanding, which supports morale and allows for peer recognition of good performance (NAO, 2008).

26. Ashraf et al. (2014) conducted a field experiment to evaluate the effects of extrinsic rewards on the performance of staff recruited by a public health organisation to promote HIV prevention. They concluded that non-financial rewards are effective at improving performance, and are stronger for pro-socially motivated staff. However, both financial and non-financial rewards can improve the performance of agents engaged in public service delivery, and non-financial rewards can be effective in settings where the scope for financial incentives is limited.

27. Monetary incentive schemes need to be designed and implemented with considerable focus on the specific cultural and administrative environment. However, some general lessons concerning sound monetary reward systems exist. Firstly, explicit monetary incentives have the advantage of being able to focus service providers on areas of low performance or output that would have been neglected if intrinsic motivation alone had been relied upon (Prentice et al., 2007). However, linking performance directly to budgetary decisions should be done with caution. The greater the pressure an organisation is under to meet a certain benchmark, the greater the temptation for gaming. Because of this, a direct or tight link between funding and performance results should not be implemented on a government-wide scale. Such automatic linkages can distort incentives and require a high level of robust performance data, which is rarely

available across all public services. Direct monetary linkages should be decided on a case-by-case basis rather than by establishing a government-wide system (Curristine et al., 2007).

28. Secondly, reward/sanction schemes should ensure that every sub-national government is awarded funding by meeting minimum standards. If reward schemes only reward the best performers and sanction the worst, then the majority of stakeholders may be content to maintain a mediocre level of performance with no real motivation to improve (NAO, 2008). If a region does not meet the minimum conditions, then sanctions can be brought into effect. Sanctions should be adjusted to match the extent of the performance shortfall. If the problem is less severe, then the municipality should be provided with assistance (e.g. building institutional, organisational and individual capacity) to enable it to meet the basic conditions. If the problem is severe, sanctions may include suspending or withholding a payment. Sub-national governments that exceed expectations should receive an additional funding allocation.

29. The standards or targets should also be realistic, which can be judged by using historical benchmarks. This involves comparing the change in the metric with the evolution of the metric over time. If the new target establishes a rate of change that substantially outpaces previous experience it may be over-ambitious in the absence of a clear justification (Curristine et al., 2007). There has been considerable interest in the use of performance-related pay for public servants, including teachers and public sector doctors. Burgess et al. (2017) analysed the effects of a pilot scheme that applied a team-based performance pay scheme – workers were rewarded bonus payments on the basis of team rather than individual production. They found that the impact of the incentive scheme was positive, substantial and significant in small teams and negligible in large teams. Thus, while some mechanism such as peer monitoring does overcome free-riding problems in small teams, it appears not to do so in large teams. They also showed that less precise metrics of quality and poorer monitoring of quality relative to quantity in the pay-for-performance scheme meant that the scheme did not raise quality. This is just one example of where further analysis into labour incentives, including how public sector employees may differ from those in the private sector, is warranted.

30. Further research is needed into the effects of explicit reward schemes, in terms of their cost-effectiveness and how they improve quality. Unfortunately, evaluating pay-for-performance schemes is made difficult by the lack of comparable data before and after the introduction of a scheme, or of control areas that are not exposed to the scheme (Meacock et al., 2014). When designing the implementation of a performance system, governments should consider how the scheme will be evaluated, which would be helped by piloting schemes across specific jurisdictions or regions. Consideration should also be given to the timelines for pay for performance schemes and ways to withdraw payments. The effects of such schemes are generally short term. The potential gains from reward funding fall over time as providers approach a natural maximum level of achievable performance, yet the size of bonus payments is often kept constant. Analysis by Lester et al. (2010) suggests that the subsequent performance declines may be very large, and even fall below levels observed before the introduction of the policy. Lester et al. (2010) evaluated the effects of financial incentives on clinical quality at 35 medical facilities in California. Incentives for two indicators—screening for diabetic retinopathy and for cervical cancer—were removed during the study period. During the five consecutive years when financial incentives were attached to screening for diabetic retinopathy, the rate rose from 84.9 per cent to 88.1 per cent. This was followed by four years without incentives when the rate fell to 80.5 per cent. A similar outcome was seen for cervical cancer screening. Although this study had some limitations, it generally showed that the removal of direct financial incentives was associated with declines in performance.

31. Meacock et al. (2014) aimed to calculate the cost-effectiveness of a reward scheme in the United Kingdom's health system. This scheme was only implemented in the North West of England, allowing for comparison of changes in the outcomes of patients across England. The outcomes studied were risk-adjusted mortality rates, emergency readmissions and length of stay. They found a significant reduction in mortality and length of stay due to the reward scheme, whilst readmission rates remained unchanged. The

study concluded that the scheme represented a cost-effective use of resources. The potential cost-savings resulting from avoided hospital readmissions and the quality-adjusted life years gained across a large patient population outweighed the costs of implementing the funding programme.

32. As suggested previously, performance schemes that allow for consultation with all stakeholders are more likely to result in performance gains, but national frameworks need to be available. The Commission for Quality and Innovation (CQUIN) framework was introduced across England in 2009 with the aim of improving the provision of acute, ambulance, community, mental health and learning disability services. The framework required local providers to negotiate and implement an annual pay for performance scheme. The indicators included in CQUIN were agreed locally, with the intention of offering flexibility to include local priorities and generating local enthusiasm, as well as safeguarding against gaming and any unintended distortions. National guidelines were published outlining the dimensions of care that locally designed CQUIN schemes should cover, including guidance on the choice of indicators, but ultimately the schemes were to be agreed locally. The local development of indicators allowed for variations in local needs, but studies suggest that the involvement of clinical staff was insufficient to generate local enthusiasm around the scheme. Kristensen et al. (2013) analysed 337 locally negotiated CQUIN schemes. They concluded that while local strategic input is important when designing funding schemes, this should be kept separate from the technical design process of defining indicators, agreeing thresholds and setting rewards. These tasks require expertise that is unlikely to exist in each locality. The schemes did not in general live up to the requirements set by the Department of Health to ensure that local schemes addressed the original objectives of the CQUIN framework. Balancing the policy goal of localism with the objective of improving patient outcomes led Kristensen et al. (2013) to conclude that a firmer national framework would have been preferable.

3. Approaches to benchmarking

33. This section looks at various methodologies to measure the productivity of public services in the national accounts as well as more qualitative mechanisms to measure performance. There has been considerable improvement with regards to conceptualising the challenges associated with measuring public sector productivity. However, the implementation of measurement techniques has not been widespread. It is very challenging to establish a clear link between outputs and outcomes, and to ensure that the policy outcomes measured are solely the result of public sector activities. A further challenge with regard to output measurement for productivity purposes is how to incorporate changes in the quality of outputs. Beyond the pure quantitative mechanisms to measure productivity and outcomes, mechanisms like consumer surveys and inspection systems can help governments measure the quality of public services.

3.1. Challenges with measuring outcomes

34. Public organisations across many OECD countries are responsible for the delivery of key services, with the expectation that they will be provided efficiently, effectively, and equitably. OECD countries are facing ageing populations and increasing dependency ratios that will increase the demand for certain public services and put pressure on public spending, making measuring cost-efficiency and quality of services even more a priority. Improvements in the quality of private goods and services are normally taken into account by hedonic pricing or other techniques. However, given that many public services are funded through taxes and provided for free, it is typically not possible to use this approach for government services.

35. Historically, public sector outputs have been quantified in the national accounts by equalling inputs, which assumes that public sector productivity is zero. This simplification is no longer seen as acceptable, but the implementation of alternative methods is challenging in the public sector, given the absence of market prices and multiple objectives that are difficult to measure. Education and health are the most common examples of services provided by governments free of charge or at prices which are not

economically significant and thus constitute non-market output (Schreyer, 2010). Despite this, across OECD and partner countries, only twelve measure public sector productivity for the health care sector and eleven for the education system, as shown in Figure 2 below (OECD, 2017a).

Figure 2. Measurement of public sector productivity at different levels

Source: OECD Survey on measures of productivity in OECD and partner countries (OECD, 2017a).

36. It is important to distinguish between outputs and outcomes. Outputs are the goods and services that the public sector supplies, and are useful for measuring service efficiency. Outcomes are the effect or consequence of service consumption (or lack of consumption), and are used when measuring the effectiveness or quality of a service. Outcomes can be broken down into direct and indirect outcomes, the distinction being that direct outcomes are closer to the act of service provision than indirect outcomes (Schreyer, 2010). Output measurement is critical for pure productivity or efficiency measurement. However, outcomes are integral to measuring the effectiveness or quality of services. Implementing a benchmarking framework based on outcomes is important for public sector services, as ultimately, the purpose of the public sector is to achieve a desirable social or community outcome. An outcomes-based approach requires an organisation to define their desired end result in detailed terms prior to implementing the measurement strategy. In this way, outcome benchmarking will guide an organisation to focus on improvements and their ultimate impact on society.

37. Measuring service effectiveness requires understanding the relationship between inputs and outcomes, as well as separating outcomes from outputs. Measuring effectiveness is especially difficult as most service processes require several inputs with the aim of achieving several social and economic outcomes. There are often delays between the implementation of policies and their impact, made more difficult by large, sweeping reform packages that are difficult to analyse. While the measurement of policy outcomes is not particularly hard in many cases (e.g. citizens' health status, education level of pupils), it is much more challenging to dissect public policy action (outputs) from policy outcomes. In other words, efficiency and effectiveness are easily confused and can be difficult to separate. This is known as the attribution problem (Lau et al., 2017). Similarly, in most OECD countries many public services including health and education are delivered by a mix of public and private providers, often with funding for one

service provider originating from both public and private sources. This can make it difficult to separate the effects of public and private inputs on outcomes, especially given data limitations in many countries.

38. Outcomes can be influenced by a range of external variables, sometimes beyond the control of governments which makes it challenging to identify outcomes that are solely the result of public sector activities. Benchmarking based on ‘gross outcomes’ is often used, which measures outcomes at a designated date but do not aim to measure the actual output of a service itself (Propper and Wilson, 2003). For example, a gross benchmark for education may be the number of pupils passing exams at a certain grade. However, this makes no adjustment for students that would have passed exams in the absence of government spending on schools. Gross outcome benchmarks are much easier to measure and collect but may not be fully attributable to policy. Net or valued-added outcomes show how a service improves performance over time. They are a better indicator of the impact of a government service, but require establishing a counterfactual of what would have happened in the absence of the government service or statistically controlling for these external factors. Box 3 examines value-added measures used in the English education system.

39. A further challenge with regard to output measurement is how to incorporate changes in quality. Unfortunately, measuring quality is much more difficult and does not constitute a widespread practice among OECD countries. Only five of the 16 countries that measure public sector productivity make use of quality adjustments – Hungary, Ireland, the Slovak Republic and the United Kingdom apply quality adjustments to output data. It should also be noted that quality adjustments can be made to inputs, for example by taking into account the education level of the teaching workforce when making quality adjustments to education services, and New Zealand reports using quality adjustments for input measures (OECD, 2017a).

Box 3. Quality measures for school education in England

A measure of educational outcome should take into account the quantity of teaching provided by the producers of education, the quality of the education provided and the level and field of education for which education is provided. For secondary education, an important quality component is how much a school can be expected to contribute to the attainment of knowledge and skills, and data on exam scores may provide a conceptually correct and empirically feasible option for explicit quality adjustments.

“Value-added” performance measures in the education system relate to quantifying the gain in the knowledge and skills of students, and their ability to progress towards education objectives, over a given time period (Sanders, 2006). The concept of ‘value-added’ aims to isolate the impact of the school environment on pupil progress by incorporating prior attainment to control for factors outside the school’s control. Contributions by socio-economic background, inherited skills or private tutoring should not be included (Schreyer, 2010). This is with the desire to recognize service delivery improvements in all schools equally and fairly, including those with a greater proportion of students from low socio-economic backgrounds.

Since 2003, a value-added measure of schools was added to league tables in England. The measure is calculated by comparing the A-level results of students at a school with the A-level results of students in schools across England who started with similar results for a particular grade. A score above zero means students made more progress, on average, than students across England who got similar results. Wilson (2003) examined the impact of the value-added indicator on the rankings of English schools. The paper suggests that the value-added indicator provides a more accurate measure of school performance and hence should help parental choice. However there was evidence that a single value-added indicator may not be sufficiently informative.

It should be noted that the English value-added measure cannot be linked to efficiency as it does not account for the resources used by a school. This is not a disadvantage if the measure is simply to improve parental choice. However, this type of measure should not be used by governments aiming to raise standards from the same resource base, which would require a measure net of peer effects (Wilson, 2003).

For most governments, benchmarking indicators will only be one tool with the aim to improve academic performance. Although benchmarking indicators are not perfect, they go some way towards reducing information gaps and supporting parental choice. It is imperative to ensure that a performance measurement system promotes and tracks the types of skills and knowledge demanded by a society.

40. In order to account for changes in quality, one of the main approaches is to differentiate or stratify products, so that only products of the same specification are compared over time or in space. Such grouping ensures that only prices or quantities of products of very similar quality are compared. The idea is that products of different quality are treated as different products. The more detailed the stratification, the more similar the activities that are compared over time, and the better the implicit quality adjustment (Schreyer, 2010). Explicit quality adjustment techniques like hedonic price indices² can be used to account for quality change, but only when information on prices is available, which is often the case for hospitals but not many other public services. When this is the case, production functions can be constructed to model outcomes (or outputs) as a function of inputs, institutional variables and external characteristics.

3.1.1. Measuring education services

41. In the education sector, the delivery of services can be accomplished either by a market or a non-market activity. The output for education services suggested by the OECD (Schreyer, 2010) and the European Commission (European Commission, 2016) is “the amount of teaching received by students for each type of education”. The measurement of output should be performed at the individual level and focus on individual benefits for the students (OECD, 2017a). Output indicators for education include: the number of pupils and the number of pupil-hours, attainment rates or the percentage of pupils that graduate from specific education levels.

42. There are different ways of accounting for changing quality in education services. One way is through stratification, where different types of education require different types of teaching services. Given that hours taught at different levels of education are not the same type of service, and cannot therefore simply be added up, stratification by different levels of education is recommended. In addition, it may be necessary to distinguish across different education services, for example, for disabled students.

43. An explicit quality adjustment can be used when stratification is insufficient. Explicit adjustments require identification of quality characteristics that are not present in the stratification process and the calculation of an adjustment factor that is applied to unadjusted measures of teaching services. For instance, the number of exam scores could be taken as the output of the education sector, provided that it is possible to capture only those movements in exam scores that are due to teaching services (Schreyer, 2010). The OECD PISA results are based on test scores of 15-year-old pupils, and represent a well-known example of outcome measurement of the performance of the educational system.

3.1.2. Measuring public hospital services through case mix adjustments

44. Health care services can be organised either as a market or a non-market activity. Health sector performance can be measured at the system-wide level, sub-sector level or the disease-based level. Efficiency by sector can be measured by relating certain measures of inputs (either in terms of the availability of human resources like GPs or in terms of spending) to outputs (e.g. the number of consultations per doctor) or outcomes (which can be measured indirectly through avoidable hospital admissions for conditions that should normally be treated outside hospital) (Lafortune, 2015). The present considerations suggest that it is best to treat the measurement of output of medical services by type of health care provider. The most important drawback of this approach is that it is not able to capture substitution effects between providers (Lau et al., 2017).

2. A hedonic price approach redefines goods in terms of their characteristics so that modified or new models do not open up a new product category but simply represent a new combination of characteristics (OECD, 2001).

45. This section focuses on the hospital sector because this sector still accounts for a large part of total health spending, and data on inputs and outputs tend to be more widely available (Lafortune, 2015). To measure hospital treatments, there are two common patient classification systems both of which attempt to deal with the heterogeneity of hospital output while making comparisons between hospitals possible. The first system classifies diseases, commonly through the development of diagnosis-related groups (DRGs). Hospital discharge registers constitute an alternative data source for the construction of a volume index of hospital output. Discharges correspond to diagnoses for particular diseases and can be grouped in a meaningful way so as to represent similar types of episodes of treatment (Schreyer, 2010). Other techniques exist for other health care providers.

46. The method of differentiation is gaining significant traction in the healthcare sector, through case mix adjustments. Patients are classified into groups known as DRGs based on their diagnosis and the severity of their condition. According to the latest data, the majority of countries use direct volume measures (19 countries) for non-market outputs, most of them in the form of DRGs together with DRG cost measures, while a few of them use “days of hospital care” (OECD, 2017a). In September 2014, Denmark adopted an output-based method for measuring the volume of government production, rather than the traditional input method, which involved the construction of DRGs (Statistics Denmark, 2016). Annex B provides more information on Denmark’s productivity measurement methodologies.

47. A type of case mix adjustment might be applied to measure the productivity of other public services, but obviously requires sufficient and robust data to support such an approach. For example, Ciappi et al. (2015) suggested that a similar approach to DRGs could be applied to the justice system in Italy. A standard cost could be calculated for a specific type of offender. They define an Offense-Related Group (ORG) representing a classification system that groups offenders according to the consumption of resources and services required for their social reintegration, their personality profile and type of offense. In this sense it is a bottom-up approach to effective service management. Unfortunately, this type of approach requires a significant amount of data to be collected, as well as the on-going monitoring and evaluation of costs across regions. However, this approach would go some way to measuring performance and allow for evidence-based policies aimed at improving the efficiency and cost-effectiveness of the judicial system.

48. Benchmarking systems can either be designed through a bottom-up or top-down approach. A bottom-up approach collects objective performance indicators on specific services. In this case, aspects of performance are assessed from a service user perspective, to see how efficiently a service is provided. In recent years organisations have been encouraged to collect data at a patient level. While the relevance of patient-reported outcome measures to acute care is clear, their application to such areas as chronic disease and mental illness remain less well developed (Smith et al., 2008).

3.2. Cost approach to benchmarking

49. Benchmarking sub-national or regional government service providers through a cost or price approach can provide useful information on which operators or regional governments are most cost-efficient. Given the current pressure on government spending and tight austerity measures in some countries, cost-based approaches to benchmarking may seem even more useful, especially from a political perspective. Compared to the private sector, the estimation of the actual costs of public sector activities is relatively complicated. Public sector accounts are typically not as detailed or disaggregated as those in the private sector, making it difficult to obtain information on all input costs (Mandl et al., 2008). Alternatively non-monetary factors can be used, like the number of civil servants involved in a public activity, pupil to teacher ratios or construction costs per dwelling completed.

50. Cost benchmarking is useful for central governments where sub-national governments are more reliant on funding for the delivery of certain services. In this case, central governments often set spending

targets for key services like education and health to be met by sub-national governments, to ensure compliance with national objectives. In principle, sub-national governments also have incentives to ensure such cost-minimisation of service delivery, even without centrally-imposed rules.

51. On a more decentralised setting, cost benchmarking can be useful for both sub-national and central governments as it can encourage the use of the most cost-effective approaches.³ The dissemination of detailed regional cost information to sub-national governments may help reduce information asymmetries and reveal unjustifiable or avoidable differences in costs of services. Comparing the average costs across providers, some variations will be due to factors that are beyond the control of service providers, like geographical variations in consumer needs or differences in price and wage levels. Balaguer-Coll and Prior (2006) analysed the efficiency and quality levels of Spanish local governments, and showed that a great deal of cost differences are due to exogenous factors. In particular, they found that the size of the municipality, the per capita tax revenue, the per capita grants and the amount of commercial activity are some of the factors related with local government efficiency in Spain.

52. Also in Italy, the wide variations in prices of identical items (e.g. hospital syringes), unjustified by considerations of quality or quantity, suggest that more detailed information on costs is needed. There is also some evidence of occasional manipulation of procurement rules by purchasing managers who specify particular characteristics that are only met by one favoured supplier. In 2012, the Italian government took measures to strengthen the obligation on public administrations at all levels, to make use of competitive tenders and centralised procurement agencies to ensure projects and supplies are procured at the least cost (OECD, 2013a). 53. Ballanti et al. (2014) developed an interesting model to evaluate the performance of local government public services, using data on social services provided by Italian municipalities. Their Four Quadrant model was based on the joint graphical analysis of two measures: the expenditure gap (the difference between actual and standard expenditures), and the output gap (the difference between the actual and standard level of services). They identified four types of Italian local governments: over-standards (where both the output gap and the expenditure gap are positive); under-standards (both the output gap and the expenditure gap are negative); efficient (where output gap is positive and the expenditure gap is negative) and non-efficient (where the expenditure gap is positive and the output gap is negative).

54. The OECD carried out an analysis of variations across and within countries for some common hospital procedures (OECD, 2014). Some procedures were consistently ranked as “high” variation across geographic units (cardiac procedures, knee arthroscopy, MRI exams). Surgery/admissions after hip fracture and caesarean section were generally ranked as having low variation. Given the large variations found, governments should consider systematic health monitoring for at least a core set of high-cost diagnostic and surgical procedures.

55. A benchmarking system that only focuses on costs can create incentives for service providers to reduce costs at the expense of other important elements of public service provisions, including quality or safety of care. Giannakis et al. (2005) performed a benchmarking study of electricity distribution utilities in the United Kingdom in the 1990s and found that cost-efficient firms do not necessarily exhibit high service quality, indicating a possible trade-off between costs and quality of service. These findings show that it is desirable to also include output or quality of service measures in a benchmarking approach.

56. Small and rural sub-national governments may also find that a lack of private sector services results in citizens relying more heavily on government-provided services. International evidence suggests that there can be efficiency gains up to populations of around 10 000 to 20 000, mostly driven by capital

3. Labour is the single most important input to many production processes especially with regards to public service delivery. It is recommended that actual hours worked be the statistical variable used to measure labour input, as opposed to simple head counts of employed persons. Hours paid and full-time equivalent persons can provide reasonable alternatives (OECD, 2001).

intensity (Holzer et al., 2009). Labour-intensive services are considered more easily scaled but at very small sizes these services can also suffer. In Switzerland, for example, municipalities with less than 500 residents were found to have higher costs and lower quality services (OECD, 2016a). Urbanised regions may suffer from additional costs of service delivery like congestion or higher wages. For these reasons, cost adjustments will be needed to account for external factors. Alternatively, when undertaking cost benchmarking, benchmarking should involve comparing one regional or local service provider against peer group averages to determine whether their costs differ in any areas to a degree that warrants investigation.

57. Accurate measurement of inputs at any level requires cost accounting, which in turn builds on an accrual-based accounting system that registers costs and not only cash-flows (Lau et al., 2017). By 2016, around three-quarters of OECD countries had adopted full accrual accounting (Moretti, 2016). For sub-national benchmarking, sub-national governments need to have identical accounting practices in place in order to ensure comparability between indicators. This is often the initial step prior to implementing a benchmarking system. For example, following revisions to Norway's municipal accounting framework in 2001, Norway implemented a compulsory benchmarking system called the Local Government Data Registration and Information Scheme (KOSTRA). This system is designed to help the central government keep track of resources used by the local government sector and is discussed in more detail in Annex D.

3.3. The use of composite indicators to benchmark multiple metrics

58. A composite indicator is an aggregate index comprising individual indicators and weights that represent the relative importance of each indicator. The construction of a composite indicator is not straightforward and the methodological challenges raise a series of technical issues. However composite indicators are much easier to interpret than trying to find a common trend in many separate indicators. Composite indicators have proven to be useful in ranking regions or service providers in benchmarking exercises; however, they can send misleading policy messages if misinterpreted. Examples of composite indicators include 'star ratings' or 'report cards'. These composite performance ratings have taken on great importance as they are often used to reward or penalise organisations.

59. In 2000, the English government introduced a star rating system (from zero to three stars) for hospitals, of which waiting time targets comprised a central element. Failure to achieve such targets resulted in sanctions. Bevan and Hood (2006) investigated the use of these performance measures in the English health care system. They found that English health care managers were exposed to greater risk of being sacked under a public performance measurement system where hospitals were 'named and shamed'. There is evidence that this has dramatically improved performance indices, but Bevan and Hood (2006) argue that it is impossible to determine whether these improvements are genuine or whether they are due to gaming or a reduction in quality that is not measured (Verbeeten, 2008).

60. Arguments in favour of composite indicators include focusing attention on important policy issues, offering a more rounded assessment of performance and presenting the 'big picture' in a way, which the public can understand. Composite indicators should be supplemented by more detailed performance information. There are obviously merits in publishing a simple aggregate rating, but for individuals using care, and the staff in providers, it is important to represent the individual elements of the rating separately. They provide an attractive option for accountability purposes, as it is easier to track progress of a single indicator over time rather than a whole package of indicators. Ratings can aid choice among users, and may be more useful for choice of providers that offer relatively simple and more homogenous services, such as general practitioners, care homes and domiciliary providers as compared to more complex care in hospitals. The Indian central government, with technical assistance from the World Bank, has recently developed composite indicators to measure the level and annual improvements of public service performance of each State (NITI Aayog, 2017). More information on India's health and education composite indicators can be found in Annex C.

61. The construction of composite indicators can also involve subjective judgement: the selection of sub-indicators, the choice of aggregation model, the weights of the indicators, and so on. These subjective choices can be used to manipulate the results. It is thus important to identify the sources of subjective or imprecise assessments, and to appraise the reliability of rankings (Jacobs and Goddard, 2007). With regards to the choice of sub-indicators, ideally sub-indicators should include output, process and outcome measures. Outputs are more immediate and are more representative of what is important to users. However, final outcomes are obviously important because they are a better indicator of the impact of government services.

62. Jacobs et al. (2005) undertook a comprehensive study looking at how robust composite indicators were in assessing the performance of 117 English acute hospitals. They concluded that random variation beyond the control of organisations gives rise to considerable uncertainty in hospital scores. Composite indicators were also sensitive to changes made to the weighting system and to the aggregation rules, with some hospitals jumping almost half of the league table as a result of subtle changes. This concurs with previous research that suggested that rankings of performance indicators are unstable (Goldstein and Spiegelhalter, 1996). Composites need to be published with indications of uncertainty to communicate the sensitivity of the reported measure (Jacobs and Goddard, 2007).

63. With regards to aggregation, the primary decision hinges on whether providers should be allowed to compensate for poor performance in one metric with superior performance in another, that is, that providers are ‘average’ if good scores on one metric compensate for poor performance on other metrics. There are three principal choices: full compensation (additive), partial compensation (multiplicative), and no compensation (non-compensatory). Because of its simplicity, the additive aggregation technique is used widely. However multiplicative or non-compensatory aggregation methods deserve more consideration, as these techniques would make it more difficult (or impossible) to offset a bad indicator with a good one and help foster better quality. Poor quality services, even if only in some aspects, may have long-term consequences (Profit et al., 2010).

64. Different weighting techniques exist, ranging from statistical weights (through principal component or factor analysis), participatory methods (that are based on expert opinion) or conjoint analysis (that use decompositional multivariate analysis to determine individuals’ preferences). Ultimately the selection of appropriate weights depends on the theoretical framework and the data properties (OECD, 2008). Most composite indicators rely on equal weighting which implies that all variables are “worth” the same in the composite (OECD, 2008). Each sub-indicator is given a weighting of one, which can result in double counting if two sub-indicators are highly correlated (Profit et al., 2010). That said, equal weighting has the benefit of simplicity and may even be superior in some contexts, for example when the sample of sub-indicators is small (e.g. compared to weighting based on regression analysis) (Bobko et al., 2007).

65. Whatever aggregation and weighting method is chosen, governments must disseminate information on the statistical techniques employed to create the composite indicator, as well as information that allows users to investigate individual sub-indicators of the composite indicator. Composite indicators must be seen as a starting point for initiating discussion and attracting public interest. Transparency will go some way to avoiding the issues of oversimplification and misinterpretation. The usefulness of composite indicators for analytic purposes should not be discounted especially in the context of a growing appetite in the economically literate press for statistics-based narratives (OECD, 2008).

3.4. Quality assurance systems

66. Quality assurance systems are another way to strengthen the monitoring and evaluation of government services. Quality assurance systems, including external inspections, peer reviews and self-assessment tools, are used to monitor overall system performance and policy implementation as well as to hold service providers accountable (European Commission, 2017). The overall role of quality

assurance mechanisms is to identify improvement opportunities and then help facilitate these improvements, generally through follow ups. External inspectors should also assume the role of disseminating good practices across service providers. With regards to hospital inspections, the main aim is to create conformity across providers, mainly in terms of minimal requirements for safety. This section mainly looks at the role of inspections in schools.

67. External mechanisms provide data important for policy-level decisions and resource allocation, while internal evaluations provide more detailed and timely data. These two mechanisms may work best together: in New Zealand for instance, school evaluations have both internal and external components. Schools undergo self-evaluations, which then form the starting point for the external evaluation process. External evaluations are also a collaborative approach between the schools and external evaluators (ERO, 2016).

68. Peer review is a closed system for professional self-assessment and development, generally achieved through reciprocal visits of professionals. Inspections are on-site evaluations involving structured methods of enquiry. Self-assessments are where an organisation systematically (and generally routinely) reviews its processes and products against procedures or legal specifications. Audits, however, are completed by a third party, internal or external to the organisation (Eurostat, 2007).

69. The aim of school inspections is to ensure that educators are accountable and to identify the strengths and weaknesses of providers, in order to bring about improvement. There is yet no conclusive, empirical evidence that school inspections automatically increase quality. However, the process of school evaluations, either internal or external, commands educators to focus on possible areas of weakness that require development. They can also identify areas of service or quality that are valued by specific consumers. Inspectors examine a school's compliance with legislative requirements and regulations, as well as providing information about the quality of teaching and the learning environment. Depending on the particular inspection system, there may be financial or non-financial rewards or sanctions for the school or its staff.

70. In Iceland, governance of the education system is shared between the central and local authorities. The Icelandic Parliament sets the administrative frameworks of the education system. Municipalities are responsible for pre-primary and compulsory education, but most schooling decisions in lower secondary education are taken at school level (OECD, 2016b). Iceland has a system of external evaluators, which allow the head teacher to comment on the draft school report, after which the evaluators send their report to the Ministry of Education. The overall results are then made public and schools have a predefined time to send the Ministry information on how they intend to respond to the results of the evaluation (European Commission, 2017).

71. In Italy, different levels of government are responsible for different aspects of the education system. The central government sets minimum quality standards and the general framework for school autonomy, to ensure uniformity within the Italian education system. Schools have had teaching, organisational and research autonomy since 2000. Italy has made several education reforms over the last few years, including the implementation of a national self-evaluation reporting system that follows a three-year cycle. In spring 2015, every Italian school received a wide set of data on its resources, processes and outcomes, and was asked to produce a self-assessment report. These reports were based on a standardised template from the National Agency for School Evaluation, and focused on identifying strengths and weaknesses, and targets to meet over the following years. The reports were published in November 2015 on the Ministry of Education portal. Following this, in 2016, external inspectors began evaluating schools and identifying areas of improvement (European Commission, 2017).

72. In many countries there has been a move away from school inspections, towards systems based on self-assessment, partly as a result of strengthened school autonomy. Self-evaluation has the merit of being immediate, responsive to the school's specific needs and circumstances and its results are 'owned'

by the school. However, self-evaluations are less likely to challenge assumptions or to arrive at conclusions which threaten key actors in the school's hierarchy. The involvement of external bodies in school evaluation provides distance from the internal dynamics of the school and gives the kind of perspective and challenge to assumptions and to the interpretation of evidence (Santiago et al., 2011).

3.5. Consumer experience and satisfaction surveys

73. To achieve meaningful measures of service quality, central and regional governments are implementing measures of consumer experience and satisfaction, in order to identify specific domains of satisfaction or value, or to measure service performance against explicit national standards. Customer satisfaction measurement may also enable an organisation to understand the extent to which satisfaction with a service is influenced by factors outside its control and to understand what is really driving satisfaction with a service experience. Numerous researchers who focused on customer satisfaction have indicated that it is an emotional reaction influenced by the interaction of users' pleasure, expectations of performance, assessment of consumer experience, and consumer interests (Tan et al., 2017).

74. Consumer satisfaction survey results, such as those derived from in-patient surveys in hospitals, are a useful type of performance measurement system. Hospital performance is becoming more focused on health education, patient empowerment, comfort, complaint mechanisms and continuity of care. Patient experience measures are a complement to quantitative quality measures, as long as the information is collected using psychometrically sound instruments, employing recommended sample sizes and adjustment procedures. Surveys of patient experience directly evaluate the degree to which care is patient-centred, and thus capture an intrinsically important dimension of care quality. Although this section mainly relates to health surveys, there are examples of where surveys have been used in the education sector. In Denmark for instance, a national performance target relating to student well-being was introduced in 2014, which generated the implementation of a national survey on student well-being. Results from the survey are intended to help devise initiatives to enhance students' well-being. More information on Denmark's performance systems are provided in Annex B.

75. For transactional services, the outcome and therefore the measurement of the outcome is straightforward. For more complex public services including education and health, it is more difficult to untangle the impacts of satisfaction with the process, satisfaction with the outcome and the initial expectations the consumer had. Satisfaction surveys can help separate these variables.

76. There are four broad themes that are likely to be relevant to all organisations:

- Delivery of the service (how problems were handled, timeliness, security, access, reliability);
- Information (accuracy, enough information, kept informed);
- Professionalism (competent staff, fair treatment); and
- Staff attitude (friendly, polite, committed, courteous, communication strategies).

77. With regards to the compilation of surveys, there are certain things that can lead to respondent error in self-reported survey measures, including memory failure, unclear questions or the motivation by some respondents to present themselves in a positive light. Contextual cueing refers to the influence that subtle cues within the survey context can have on how individuals answer questions. Contextual cues can potentially lead to priming effects and socially desirable responding. Question construction requires consideration of the precise wording of a question as well as the reference period that respondents are asked to consider when forming their answers (e.g. "in the last week" versus "in the last month") (OECD, 2013b).

78. Early survey initiatives were often modelled on the work of the Picker Institute in the United Kingdom and the Consumer Assessment of Healthcare Providers and Systems (CAHPS) initiative in the United States (Lau et al., 2017). The Picker Institute developed the Principles of Patient Centred Care in 1987, which became a framework used internationally to support high-quality patient-centred care. In 2002, it designed and established England's first National Health Service (NHS) national survey programme for patient experience. Favourable Picker scores have shown correlations with significantly reduced complications and unexpected deaths in hospitals (WHO, 2003).

79. With regards to health care surveys specifically, patient-reported indicators of health system performance largely relate to patient-reported experience measures (PREMs, such as whether the patient feels they were adequately involved in important decisions about their care), and patient-reported outcome measures (PROMs, such as whether the patient is free of pain after an operation care). Some OECD countries are conducting PREMs surveys and, to a lesser extent, are experimenting with PROMs (Lau et al., 2017). Multiple studies across different countries revealed that the nurses' courtesy, respect, careful listening and easy access of care were the strongest drivers of overall patient satisfaction (Al-Abri and Al-Balushi, 2014). Otani et al. (2011) found that the courtesy and respect of healthcare providers affect patient satisfaction most, while communication and explanation are the second most important aspect. Further, research suggests that better patient care experiences are associated with higher levels of adherence to recommended prevention and treatment processes, better clinical outcomes and better patient safety culture within hospitals (Price et al., 2014). Several studies have examined relationships between patient-reported experiences and clinical outcomes. In a study by Meterko et al. (2010), they found that patient reports of better patient-centred hospital care were significantly associated with better survival rates after treatment. Similarly, controlling for hospitals' clinical performance, Glickman et al. (2010) found that higher patient ratings for hospitals were independently associated with lower hospital in-patient mortality rates. These findings suggest that many aspects of service delivery that are outside the usual realm of quantitative testing are important for consumer satisfaction narrowly, and important for consumer outcomes more generally.

4. Avenues for further research

80. This paper has provided insights into how benchmarking can be achieved, and provides a preliminary look at what countries are doing to ensure central governments can monitor the performance and productivity of public service delivery. Two key observations emerge from the analysis.

81. First, there is a need for further empirical research and analysis into the impact of key institutional variables on the performance of sub-national public service delivery and the administrative and political drivers of public sector productivity, including the uptake of digital government innovations and human resource management, especially across regions. Looking more closely at specific service areas, notably health, education and/or judicial services, would allow for more in-depth analysis and analytical work among comparable organisations. Given the trend towards decentralisation of health care and other expenditure in many OECD countries, there is a strong motivation for ensuring that this spending is used efficiently at the sub-national level. Many analysts believe that part of the reason health care spending has increased so rapidly is low productivity growth in the health care sector. However, other analysts maintain that health care productivity is actually rising, pointing to improvements in cancer survival rates and improved quality of life for people with heart disease, depression and other illnesses. The true position is critically important for the fiscal sustainability of health care.

82. One concrete project has already advanced. This effort is looking more closely at performance measurement systems in the health care sector, using a questionnaire that has been circulated to countries through the Joint OECD Network on Fiscal Sustainability of Health Systems. Future work that could build from the new information collected from this survey is manifold. For instance, data gathered from the questionnaire on the responsibilities of hospitals across levels of government could be associated with new unit-level hospital performance data, with policy impacts estimated using econometric techniques such as difference-in-difference estimation. Hospital-level performance data are being compiled for many OECD countries in a parallel effort by the Health Division of the OECD. More qualitative analysis aimed at offering policy recommendations to improve productivity measurement systems in the health sector across OECD countries could also be an avenue for further work, including specific country studies and examination of the impact of new technologies. Given the fledgling quality adjustments in public sector performance measurement, further information on the measurement obstacles countries face would be useful, such as privacy concerns, data limitations or legal barriers. Also of great interest would be projects that could extend earlier Fiscal Network analysis (Fredriksen, 2013) in primary and secondary education using the latest PISA data, seeking to understand management issues better; or examine less-examined sectors such as the interaction between community and legal services at the local level.

83. Second, the main benefit of performance benchmarking systems depends on how they are used and implemented. A second area of further work could be focused on how performance information is incorporated in determining inter-governmental grants or regulations on sub-national spending, and how the information is used by sub-national government to improve productivity. Consideration needs to be given to whether and how benchmarking systems are used by decision makers, and what incentives can be used to motivate decision makers to use the information. Similarly, the effects of quality and the cost-efficiency of performance budgeting or pay for performance systems could also be an avenue for further research. Finally, analysis on benchmarking could also be undertaken from a regional and/or cross-border perspective, for comparing local governments with peers, especially with regards to co-ordination procedures between sub-national governments, standardised accounting practices and information sharing between governments.

REFERENCES

- Adam, A., M. Delis and P. Kammas (2014), “Fiscal Decentralization and Public Sector Efficiency: Evidence from OECD countries”, *Economics of Governance*, Vol. 15/1, pp. 17-49.
- Al-Abri, R and A. Al-Balushi (2014), “Patient satisfaction survey as a tool towards quality improvement”, *Oman Medical Journal*, Vol. 29/1, pp. 3-7.
- Ashraf, N., O. Bandiera and B.K. Jack (2014), “No Margin, no Mission? A Field Experiment on Incentives for Public Service Delivery”, *Journal of Public Economics*, Vol. 120/C, pp. 1-17.
- Australian Productivity Commission (2017), *Report on Government Services 2017*, Australian Productivity Commission, <http://www.pc.gov.au/research/ongoing/report-on-government-services/2017>.
- Balaguer-Coll, MT. and D. Prior (2006), “Short- and Long-term Evaluation of Efficiency and Quality. An Application to Spanish Municipalities”, *Journal of Applied Economics*, Vol. 41/23, pp. 2991-3002.
- Balci, B., A. Hollmann and C. Rosenkranz (2011), “Service Productivity: A Literature Review and Research Agenda”, http://reser.net/materiali/priloge/slo/balci_et_at.pdf.
- Ballanti, D et al. (2014), “A simple Four Quadrants Model to monitor the performance of Local Governments”, CESifo Working Paper Series No. 5062.
- Bénabou, R. and J. Tirole (2006), “Incentives and Pro-social Behaviour”, *American Economic Review*, Vol. 96/5, pp. 1652-1678.
- Bevan G. and R. Hamblin (2009), “Hitting and Missing Targets by Ambulance Services for Emergency Calls: Effects of Different Systems of Performance Measurement within the UK”, *Journal of the Royal Statistical Society Series*, Vol. 172/1, pp. 161–190.
- Bevan, G. and C. Hood (2006), “What’s Measured is what Matters: Targets and Gaming in the English Public Health Care System”, *Public Administration*, Vol. 84/3, pp. 517-538.
- Blöchliger, H. and B. Égert (2013), “Decentralisation and Economic Growth - Part 2: The Impact on Economic Activity, Productivity and Investment”, *OECD Working Papers on Fiscal Federalism*, No. 15, OECD Publishing, Paris. <http://dx.doi.org/10.1787/5k4559gp7pzw-en>.
- Blomqvist, P. and P. Bergman (2010), “Regionalisation Nordic Style: Will Regions in Sweden Threaten Local Democracy?”, *Local Government Studies*, Vol. 36/1, pp. 43-74.
- Bobko P., P.L. Roth and M. A. Buster (2007), “The Usefulness of Unit Weights in Creating Composite Scores: A Literature Review, Application to Content Validity, and Meta-analysis”, *Organizational Research Methods*, Vol. 10/4, pp. 689–709.

- Bogumil, J. (2003), "Politische Rationalität im Modernisierungsprozess" [Political rationality in the modernisation process], in K. Schedler and D. Kettinger (eds), *Modernisierung mit der Politik, Ansätze und Erfahrungen aus Staatsreformen*, (Haupt, Bern).
- Braadbaart, O. (2007), "Collaborative Benchmarking, Transparency and Performance: Evidence from the Netherlands Water Supply Industry", *Benchmarking: An International Journal*, Vol. 14/6, pp. 677-692.
- Burgess, S. et al. (2017), "Incentives in the Public Sector: Evidence from a Government Agency", *The Economic Journal*, Vol. 127/605, pp. F117–F141.
- Ciappi, S. e al. (2015), "A Method to Measure Standard Costs of Juvenile Justice Systems: the example of Italy", *Working paper No. 15*, University of Verona, Department of Economics.
- Curristine, T., Z. Lonti and I. Joumard (2007), "Improving Public Sector Efficiency: Challenges and Opportunities", *OECD Journal on Budgeting*, Vol. 7/1.
- Daraio, C. and L. Simar (2007), "Chapter 2: the Measurement of Efficiency", in: *Advanced Robust and Nonparametric Methods in Efficiency Analysis: Methodology and Applications*, Springer Science & Business Media.
- Dumciuviene, D. (2015), "The Impact of Education Policy on Country Economic Development", *Social and Behavioral Sciences*, Vol. 191, pp. 2427-2436.
- Education Review Office (2016), "Integrating internal and external evaluation for improvement", www.ero.govt.nz/publications/effective-school-evaluation/integrating-internal-and-external-evaluation-for-improvement/.
- Eurostat (2007), *Handbook on Data Quality Assessment Methods and Tools*, <https://unstats.un.org/>.
- European Commission (2016), *Handbook on prices and volume measures in national accounts*, Eurostat, <http://ec.europa.eu/eurostat/documents/3859598/7152852/KS-GQ-14-005-EN-N.pdf/839297d1-3456-487b-8788-24e47b7d98b2>.
- European Commission (2017), "Quality assurance for school development: Guiding principles for policy development on quality assurance in school education", ET 2020 Working Group Schools 2016-18.
- Financial and Fiscal Commission (2013), "Chapter 11: Improving the performance of municipalities through incentive-based grants" in *Submission for the Division of Revenue 2014/15*, Financial and Fiscal Commission, South Africa.
- Fredriksen, K. (2013), "Decentralisation and Economic Growth – Part 3: Decentralisation, Infrastructure Investment and Educational Performance", *OECD Working Papers on Fiscal Federalism*, No. 16. <http://dx.doi.org/10.1787/5k4559gg7wlw-en>.
- Giannakis, D., T. Jamasb, and M. Pollitt (2005), "Benchmarking and incentive regulation of quality of service: An application to the UK electricity distribution networks", *Energy Policy*, Vol. 33/17, pp. 2256-2271.
- Glickman S.W. et al. (2010), "Patient satisfaction and its relationship with clinical quality and inpatient mortality in acute myocardial infarction". *Circulation cardiovascular quality and outcomes*, Vol. 3/2, pp.188–195.

- Goldstein, H. and D.J. Spiegelhalter (1996), “League tables and their limitations: Statistical issues in comparisons of institutional performance”, *Journal of the Royal Statistical Society*, Vol. 159(3): 385-443.
- Holzer, M. et al. (2009), “Literature Review and Analysis Related to Optimal Municipal Size and Efficiency”, Local Unit Alignment, Reorganisational, and Consolidation Commission, New Jersey.
- Jacobs, R. and M. Goddard (2007), “How Do Performance Indicators Add Up? An Examination of Composite Indicators in Public Services”, *Public Money & Management*, Vol. 27/2, pp. 103-110.
- Jacobs, R., M. Goddard and P.C. Smith (2005). “How robust are hospital ranks based on composite performance measures?” *Medical Care*, Vol. 43/12, pp. 1177-84.
- Koethenbuerger, M. (2008). “Revisiting the “Decentralization Theorem”—On the Role of Externalities”, *Journal of Urban Economics*, Vol. 64/1, pp. 116–122.
- Kouzmin, A. et al. (1999), “Benchmarking and performance measurement in public sectors: Towards learning for agency effectiveness”, *International Journal of Public Sector Management*, Vol. 12/2, pp. 121-144.
- Kristensen, R., R. McDonald and M. Sutton (2013), “Should pay-for-performance schemes be locally designed? Evidence from the Commissioning for Quality and Innovation (CQUIN) Framework”, *Journal of Health Services Research and Policy*, Vol. 18/2.
- Kuhlmann, S. and H. Wollmann (2006), “Transaktionskosten von Verwaltungsreformen— ein ‘missing link’ der Evaluationsforschung” [Transaction Costs of Administrative Reforms - a 'missing link' in evaluation research], in Jann, W. et al.(eds), *Public Management. Grundlagen, Wirkungen, Kritik* (Sigma, Berlin).
- Kuhlmann, S. and T. Jäkel (2013), “Competing, collaborating or controlling? Comparing benchmarking in European local government”, *Public Money & Management*, Vol. 33/4, pp. 269-276.
- Lafortune, G. (2015), *Developing health system efficiency indicators: Overview of key concepts, general approaches, and current and future work*, Meeting of OECD Health Data National Correspondents background document.
- Lau, E., Z. Lonti and R. Schultz (2017), “Challenges in the measurement of public sector productivity in OECD countries”, in *International Productivity Monitor: CSLS-OECD Special Issue from the First OECD Global Forum on Productivity*, Vol. 32, pp. 180–195.
- Lester, H. et al. (2010), “The impact of removing financial incentives from clinical quality indicators: longitudinal analysis of four Kaiser Permanente indicators”. *The BMJ*, Vol. 340/C, p. 1898.
- Mandl, U., A. Dierx and F. Ilzkovitz (2008), “The effectiveness and efficiency of public spending”, *Economic Papers* 301, European Commission.
- Meacock, R., S. Kristensen and M. Sutton (2014), “Paying for improvements in quality: recent experience in the NHS in England”, *Nordic Journal of Health Economics*, Vol. 2/1, pp. 239-255.
- Meterko, M. et al. (2010), “Mortality among patients with acute myocardial infarction: the influences of patient-centered care and evidence-based medicine”, *Health services research*, 45/5, pp. 1188–1204.

- Mizell, L. (2008), “Promoting performance – Using indicators to enhance the effectiveness of sub-central spending”, *OECD Working Papers on Fiscal Federalism*, No. 5.
<http://dx.doi.org/10.1787/22265848>.
- Moretti, D. (2016), “Accrual Practices and Reform Experiences in OECD Countries Results of the 2016 OECD Accruals Survey”, *OECD Journal on Budgeting*, Vol. 16/1, pp. 9-28.
- National Audit Office (2008), *The use of sanctions and rewards in the public sector: Performance measurement practice*, www.nao.org.uk/wp-content/uploads/2008/09/sanctions_rewards_-_public_sector.pdf.
- National Audit Office (2001), *Inpatient and Outpatient Waiting in the NHS* (HC 221, 26 July).
- NITI Aayog, (2017), “Working with the States: Cooperative Federalism”,
<http://niti.gov.in/content/cooperativefederalism>.
- OECD (2017a), *Understanding Public Sector Productivity*, Annual Meeting of OECD Senior Budget Officials background document.
- OECD (2017b), *Government at a Glance 2017*, OECD Publishing, Paris,
http://dx.doi.org/10.1787/gov_glance-2017-en.
- OECD (2016a), *OECD Economic Surveys: Czech Republic*, OECD Publishing, Paris.
- OECD (2016b), *Education Policy Outlook: Iceland*, OECD Publishing, Paris.
- OECD (2015a), *The Future of Productivity*, OECD Publishing, Paris. <http://oe.cd/GFP>.
- OECD (2015b), *Fiscal Sustainability of Health Systems: Bridging Health and Finance Perspectives*, OECD Publishing, Paris.
- OECD (2014), *Geographic Variations in Health Care: What Do We Know and What Can Be Done to Improve Health System Performance?*, OECD Health Policy Studies, OECD Publishing.
- OECD (2013b), *OECD Guidelines on Measuring Subjective Well-being*, OECD Publishing, Paris,
<http://dx.doi.org/10.1787/9789264191655-en>.
- OECD (2013a), *OECD Economic Surveys: Italy*, OECD Publishing, Paris.
- OECD (2009), “Components of Integrity: Data and Benchmarks for Tracking Trends in Government”, Global Forum on Public Governance, 4-5 May 2009, OECD, Paris.
- OECD (2008), *Handbook on Constructing Composite Indicators: Methodology and User Guide*, OECD Publishing, Paris.
- OECD (2001), *Measuring Productivity: Measurement of aggregate and industry-level productivity growth*, OECD Publishing, Paris.
- OECD, Eurostat and World Health Organisation (2017), *A System of Health Accounts 2011: Revised edition*, OECD Publishing Paris.
- Oates, W.E. (1999), “An essay on fiscal federalism”, *Journal of Economic Literature*, Vol. 37: 1120–1149.

- Otani, K., P.A. Herrmann and R.S. Kurz (2011), "Improving patient satisfaction in hospital care settings", *Health Services Management Research*, Vol. 24/4, pp. 163-169.
- Price, R. et al. (2014), "Examining the Role of Patient Experience Surveys in Measuring Health Care Quality", *Medical Care Research and Review*, Vol. 71/5, pp. 522-554.
- Propper, C. and D. Wilson (2003), "The use and usefulness of performance measures in the public sector", *Oxford Review of Economic Policy*, Vol. 19/2, pp. 250-65. Prentice, G., S. Burgess and C. Propper (2007), "Performance-pay in the public sector: A Review of the issues and evidence", Office of Manpower Economics.
- Profit, J. et al. (2010), "Improving benchmarking by using an explicit framework for the development of composite indicators: an example using pediatric quality of care", *Implementation Science*, Vol. 5/13.
- Robinson, M. (2007), "Introduction: Decentralizing Service Delivery? Evidence and Policy Implications" in *Decentralising Service Delivery?*, IDS Bulletin, Vol. 38/1, pp. 1-6.
- Santiago, P. et al. (2011), "OECD Reviews of Evaluation and Assessment in Education: Australia", OECD Publishing, Paris.
- Schreyer, P. (2010), "Towards Measuring the Volume Output of Education and Health Services: A Handbook", *OECD Statistics Working Paper No. 2010/02*, OECD Publishing, Paris.
<http://dx.doi.org/10.1787/5kmd34g1zk9x-en>
- Shadow Strategic Rail Authority (2000), A New Way of Measuring Performance, www.opraf.gov.uk.
- Sørensen, P. (2014), "Reforming public service provision: what have we learned?", CESifo Venice Summer Institute Workshop on Reforming the Public Sector, 25-26 July 2014.
- Smith, P.C., E. Mossialos and I. Papanicolas (2008), *Performance measurement for health system improvement: experiences, challenges and prospects*, WHO European Ministerial Conference on Health Systems background document.
- Tan, T., T. Chen and P. Yang (2017). "User satisfaction and loyalty in a public library setting", *Social Behavior and Personality: An international journal*, Vol. 45/5, pp. 741-756.
- Tiebout, C.M. (1956), "A Pure Theory of Local Expenditures", *The Journal of Political Economy*, Vol. 64/5, pp. 416-424.
- World Bank (2001), *Decentralisation and subnational regional economies*, www1.worldbank.org/publicsector/decentralization/what.htm.
- World Health Organisation (2003), "How can hospital performance be measured and monitored?", WHO Regional Office for Europe's Health Evidence Network.
- Wilson, D. (2003), "Which Ranking? The Use of Alternative Performance Indicators in the English Secondary Education Market", *CMPO Working Paper Series*, No. 03/058.
- Verbeeten, F.H.M. (2008), "Performance management practices in public sector organizations: Impact on performance", *Accounting, Auditing & Accountability Journal*, Vol. 21/3, pp. 427-454.

ANNEX A. AUSTRALIA'S REPORT ON GOVERNMENT SERVICES

84. Australia's federal system comprises of three levels of government: central; state and territories (six states and two territories with state-type powers); and local (around 550 local government authorities), which are established through state legislation. The Australian Constitution determines the areas of expenditure for which state governments have primary responsibility, resulting in quite narrow powers to the central government. State governments are also responsible for defining the powers of their local governments (note that the Australian Capital Territory does not have any local governments, with the territorial government administering tasks that are generally assigned to local governments, like waste disposal and recreational facilities).

85. The balance of power in Australia has shifted towards the central government over time and the growth of "tied" grants to the states increased federal control and led to a relatively high degree of shared functions between state and central governments, especially in key sectors such as health, education and community services (Koutsogeorgopoulou and Tuske, 2015). In 2017-18, Commonwealth payments to the states are expected to account for around 46 per cent of total state revenue. Commonwealth payments effectively support around 45 per cent of state expenditure (Australian Treasury, 2017). This has led to a "co-operative" federalism model.

86. Every year Australia's governments co-operate in producing the Report on Government Services (RoGS), which provides information on the equity, efficiency and effectiveness of government services delivered by Australia's state governments. It is a collaborative exercise in which the Commonwealth government plays a facilitative role rather than a directive or coercive one, where service objectives and indicators are identified through a consultative approach. RoGS reflects the co-operative nature of Australian federalism and its high degree of centralisation (Banks and McDonald, 2012).

87. The range of services that are analysed has grown since the first Report was published in 1995. Currently, RoGS analyses the following service areas:

- child care, education and training;
- justice;
- emergency management;
- health;
- community services; and
- housing and homelessness.

88. In 2009, the state and central governments agreed to a review of RoGS, including comparing the process of RoGS with international best practice. The Committee concluded that RoGS was a sophisticated performance measurement tool compared with international standards, which had driven considerable data improvement. The terms of reference were updated following the review, and emphasised the dual roles of

RoGS in improving service delivery, efficiency and performance, and increasing accountability to governments and the public (McGuire and O'Neill, 2013).

89. The RoGS emphasises the importance of all governments agreeing on the objectives of a service, and then creating robust indicators to measure the effectiveness, efficiency and equity of the services. Over time, the report has aimed at focusing on the outcomes influenced by those services.

90. Performance indicators include output indicators, grouped under equity, effectiveness and efficiency, and outcome indicators (see Figure A1) (PC, 2017). Each service area in the RoGS has a performance indicator framework and a set of objectives against which performance indicators report. Equity, effectiveness and efficiency indicators are given equal prominence in RoGS performance reporting, as they are the three overarching dimensions of service delivery performance. The Productivity Commission sees it as important that all three are reported on as there are inherent trade-offs in allocating resources and dangers in analysing only some aspects of a service (PC, 2017).

Figure A1. Generalised performance indicator framework

Source: Australian Productivity Commission (2017), *Report on Government Services*.

91. The RoGS does not establish benchmarks in the formal sense of systematically identifying best practice. There has also been direct resistance to provide summary information that ranks or rates states (Banks and McDonald, 2012). This is to encourage readers to seek out the details of indicators and understand the gaps in reporting and data collection. Historically, the ‘winner’ and ‘loser’ states tend to vary across the reported services, making it easier to report in those areas where jurisdictions perform relatively poorly (Banks and McDonald, 2012). Benchmarking in RoGs is in the sense of comparing cost-effectiveness by sector in order to compare value for taxpayer funds. This creates ‘yardstick’ competition, to improve productivity and responsiveness to users. Comparisons based on efficiency and effectiveness criteria are a substitute for market competition, and performance indicators substitute for market price signals (McGuire and O'Neill, 2013).

92. The RoGS is more geared towards policy-makers rather than citizens. Although the RoGS website includes some summary diagrams and interactive statistics, the ultimate use of the report is focused on strategic budget and policy planning, and for policy evaluation. The main aim is not to increase competitive pressure on jurisdictions or increase information to support consumer choice.

93. Each service area has a set of objectives against which performance is reported. The structure of objectives is consistent across service areas and includes three components:

- The high-level objectives or vision for the service, which describes the desired impact of the service area on individuals and the wider community.
- The service delivery objectives, which highlight the characteristics of services that will enable them to be effective.
- The objectives for services to be provided in an equitable and efficient manner.

94. To assist readers to interpret performance indicator results, the report provides information on some of the differences that might affect service delivery, including information for each jurisdiction on population size, composition and dispersion, family and household characteristics, and levels of income, education and employment. However, the report does not attempt to adjust reported results for such differences. An ‘interpretation box’ for each indicator provides the definition of the indicator measure, advice on interpretation of the indicator, any data limitations, whether the reported measures are complete and/or fully comparable.

95. Reporting aligns with nationally agreed data definitions and draws on data collected and verified by national statistical agencies wherever possible. At a minimum, all data have been endorsed by the contributor and subjected to peer review by a working group made up of representatives of relevant line agencies from all jurisdictions (Banks and McDonald, 2012).

96. An example of the Productivity Commission’s framework for school education reporting is given below (Figure A2).

Figure A2. RoGS School education performance indicator framework

Source: Productivity Commission (2017), *Report on Government Services*.

97. The Productivity Commission uses ‘retention’ to the final years of schooling as an indicator of governments’ objective that all students have access to high quality education and training necessary to complete education to year 12 or its equivalent. ‘Retention’ is defined as the number of full time school students in a designated year of education as a percentage of their respective cohort group (PC, 2017). A higher or increasing apparent retention rate suggests that a larger proportion of students are continuing to participate in school education, which is likely to result in improved educational outcomes.

98. The Productivity Commission suggests that ‘student engagement’ is an indicator of a governments’ objective that all students are able to improve educational outcomes through high or increasing levels of behavioural, emotional and cognitive engagement with schooling. Student engagement is regarded as relevant to student outcomes, and is closely connected to learning. ‘Student engagement’ is yet to be defined by the Productivity Commission. Attendance has sometimes been regarded as a proxy for student engagement and there is evidence of the relationship between poor attendance and poor student outcomes, particularly once patterns of non-attendance are established. However, the Productivity Commission suggests that measurement of attendance alone is not an adequate proxy for engagement with learning (PC, 2017).

REFERENCES

- Australian Government (2017), *Federal Financial Relations Budget Paper No. 3, 2017-18*, www.budget.gov.au/2017-18/content/bp3/download/bp3_consolidated.pdf.
- Banks, G. and L. McDonald (2012), “Benchmarking and Australia’s Report on Government Services”, in A. Fenna and F. Knupling, *Benchmarking in Federal Systems, Roundtable Proceedings*, Melbourne, 19–20 December 2010. Canberra: Productivity Commission.
- Koutsogeorgopoulou, V. and A. Tuske (2015), “Federal-State Relations in Australia”, *OECD Economics Department Working Papers*, No. 1198, OECD Publishing, Paris, <http://dx.doi.org/10.1787/5js4h5mx7bg0-en>.
- McGuire, L. and D. O’Neill (2013), “Institutionalising a National Approach to Performance Reporting: Lessons from Two Decades of Experience with the Report on Government Services”, *Australian Journal of Public Administration*, Vol. 72(4): 419–432.
- Productivity Commission (2017), *Report on Government Services 2017*, Australian Productivity Commission, www.pc.gov.au/research/ongoing/report-on-government-services/2017.

ANNEX B. DENMARK'S HEALTH AND EDUCATION PERFORMANCE SYSTEMS

99. Denmark is divided into 98 municipalities (local authorities) and five regions. Municipalities are managed by local councils, with all local councillors elected for a four-year period in local government elections. Denmark successfully merged its municipal structures in 2007, reducing the total number from around 300 to less than 100 and the number of regions (or counties) from 14 to five. Before 1970, Denmark was divided into 86 boroughs and approximately 1 300 parishes within 25 county council districts (MIH, 2006). One of the main drivers of the mergers was to reach a more adequate size for health care service provision, given new medical technologies, which called for larger regions (OECD, 2015). It was seen that larger municipalities could provide more locally based tasks and that democracy would be strengthened as more political decisions could be made locally (MIH, 2006).

100. In September 2014, Denmark adopted an output-based method for measuring the volume of government production, rather than the traditional input method (Statistics Denmark, 2016). Statistics Denmark collects and analyses a large number of volume indicators. Within health care for example more than 1 300 indicators are used. The outcome are 18 volume indicators, used to measure individual non-market services in the areas of health care, social protection, education, and recreation and culture. The input method is still used for some collective services like defence and environmental protection. In line with European Union requirements, the Danish government is currently not applying quality adjustments to the volume measures used (OECD, 2017).

Health

101. The Danish healthcare system is public, mainly tax-financed and organised across the three administrative levels of government. The national government sets the regulatory framework for health services and is in charge of general planning and supervision. The regions own, manage, and finance hospitals and finance the majority of services delivered by private general practitioners, office-based specialists, physiotherapists, dentists and pharmacists. Municipalities are responsible for financing and delivering nursing home care, specialised rehabilitation and home nurses, as well as health promotion and disease prevention (Commonwealth Fund, 2017; OECD, 2013)

Quality assurance systems

102. Denmark has had formal health care quality assurance mechanisms for over 20 years. The central government reaches agreements with the Danish regions on high-level service goals such as mortality or adverse event rates, without specifying more detailed targets for population-based health care outcomes. Similarly, the Danish Health and Medicines Authority (DHMA) develops service quality standards, such as disease management models, but allows them to be adapted to be operationally useful at regional level. The regions ensure that quality assurance and improvement are embedded in their activities (OECD, 2013).

103. Since 2000, Danish National Surveys of Patient Experiences on somatic hospitals have been implemented. In addition, national surveys of patients and relatives 'experiences with mental health have been implemented since 2002. Surveys of patient satisfaction with GPs and in the municipalities are currently voluntary.

104. The Danish National Survey of Patient Experiences was established in cooperation between the Ministry of health and the Danish regions. The surveys are conducted as annual nationwide surveys among in-patients and out-patients in the Danish hospitals investigating experiences related to: clinical services, patient safety, patient and staff member continuity, co-involvement and communication, information, course of treatment, discharge, inter-sectoral cooperation. Results are presented at unit, hospital, regional and national level, and they are available for the public on the official portal for the public Danish health care services. At the unit level, the results are used for identification of improvement areas, benchmarking and monitoring of development over time.

105. Denmark has unique opportunities for quality measurement and benchmarking since Denmark has well-developed health registries and unique patient identifiers that allow all registries to include patient-level data using the patients' unique patient identifier. Data are transmitted electronically and collected in clinical databases, so that trends and patterns of change, and types of improvements can be documented over time. Standardisation of more than 70 national-level clinical registries has secured efficient data collection and use of data and provided a good basis for improvement of the quality of care. Of these, three registries secure continuous data collection of mental healthcare including depression, schizophrenia and attention deficit hyperactivity disorder (OECD, 2013).

106. Digitally stored data open great potential for research, system level and everyday improvement in quality and safety. When shared and communicated, the data are the basis for accountability (Mainz et al., 2015).

Volume measures

107. Denmark measures the volume of output of health services as the quantity of health services provided to individuals adjusted for new products or services. Denmark divides healthcare services into:

- Hospital activity;
- Treatment by dentists; and
- Social provisions with and without institutional care.

108. The volume index for general hospitals is calculated on the basis of the Danish National Board of Health's Diagnosis Related Group database (DRG). In Denmark, this system is used as a tool for calculating fees to settle the accounts of patients treated in a different municipal area from the one in which they reside. The central health authorities and hospital owners also use the system to assess the correlation between activity and costs in hospital services. DRG's are increasingly used for budgeting and, particularly, as a tool for developing new methods of premises planning and management in administration and hospitals.

109. The DRGs were developed to create homogeneous cost groups in order to compare hospital activities. DRGs are good indicators of output volumes because they provide information on both (unit) costs per type of treatment and on the number of treatments. The DRG system includes a large number of categories (about 1 300), where each category denotes a rather homogenous service and thus, in principle, unit costs and volume indexes can be constructed at the most detailed level. The National Board of Health made a DRG for psychological diseases from 2008. Given cost weights and the number of treatments, a direct volume index can be constructed. The data is distributed according to a range of established main categories, known as MDC classification (Statistics Denmark, 2016).

Education

110. Although the national government implements the overall general objectives and national legislation, municipalities are responsible for the financial and organisational operation of public primary and lower secondary education including the appraisal of school principals (Nusche et al., 2016).

Quality assurance

111. Municipalities enjoy autonomy in designing quality assurance practices, specifying the local objectives and determining local guidelines for their schools. School principals at public schools are responsible for school-level administration. Most schools have developed goals or values for the wellbeing of students, the schools' educational performance and study and attainment targets for various subjects (Nusche et al., 2016).

112. Denmark's approach to performance evaluation combines a central legal framework specifying evaluation requirements and common objectives in compulsory education, with clear responsibility for municipalities to ensure quality control within this framework (OECD, 2011).

113. All public municipal primary and lower secondary schools share a set of binding learning progressions, achievement targets and curricular guidelines, the so-called Common Objectives. However, while the Common Objectives provide descriptions of how objectives can be reached, there is no tight curriculum at the national level. Common Objectives were introduced in 2003 and reduced and simplified in 2014, to ensure that learning objectives focus on learning outcomes and were goal-oriented. The Danish system deliberately allows for diversity across municipalities regarding, for instance, how to budget and allocate the appropriate level of resources needed in the area.

114. The students' educational performance in primary and lower secondary school is measured using both measurable and qualitative instruments, such as national testing, well-being surveys and the school leaving examination scores. Although an increasing amount of data have been made publicly available, school performance is not linked to rewards or sanctions and no official league tables exist.

115. The central government takes on a monitoring, supervisory role to school quality through the Agency for Education and Quality, whereas local governments are responsible for quality control procedures across their public schools, including through external evaluations.

116. In 2006, Denmark introduced the requirement for municipalities to produce annual quality reports, with reports now required biannually. In their quality reports, municipalities must describe how schools are attaining national and local objectives. The reports should describe their schools' quality of education (with a focus on outcome information), the measures the municipal board has taken to evaluate the quality of education, and the steps the municipal board has taken in response to the previous quality report. It is mandatory to publish the quality reports online, with the aim of transparency and to inform parental choice of schools (KORA, 2016).

117. A national performance target relating to student well-being was introduced in 2014, which generated the implementation of a national survey on student well-being. The first results from the national survey on well-being of students were made available in April 2015. Results are intended to inform municipalities, schools, principals, teachers, parents and students and to provide a basis for discussions and initiatives enhancing students' well-being. For municipalities, the results must be part of the quality reports: for instance, what is the state of student well-being in the schools of the municipality and what does the municipality do to follow up on results and promote well-being. While individual student results

are confidential, teachers, school management, the school board, and the municipality have access to all class results so they can collectively support further work on well-being.

118. National tests were implemented in Denmark following the introduction of Common Objectives in 2007. With respect to student outcomes, the introduction of the national tests seems to have implied learning effects beyond “teaching to the test” effects. In other words, the national tests generally seem to have improved the performance of the students, as shown by an evaluation from the Ministry of Education. At the same time, the evaluation also stressed that there was scope for an improved formative use of tests by teachers (European Commission, 2015).

119. Further, for each school the Ministry of Education publishes the average results from the compulsory school leaving examination, to act as an accountability mechanism. The results are controlled statistically for the student’s socioeconomic background. At the municipality level, while the awareness of information on student achievement is growing, many municipalities still do not fully utilise the current available data in order to follow up on the performance of individual schools (KORA, 2016).

Volume measures

120. Public non-market educational services in the national accounts are placed in four industries covering the various educational levels available in the education system:

- Primary education;
- secondary vocational education;
- higher education; and
- adult education.

121. When including primary and secondary education services as products in the National Accounts, Denmark uses a quantity indicator for primary education – a single unit of education service should be expressed as an hour of teaching received by a student at a particular level, that is, student-hours. There is no available data on student-hours for secondary education and for that reason, the number of students is used. Denmark acknowledges the shortfalls with using the number of students as a proxy.

122. Many issues remain: changes in education quality cannot be captured nor is the number of students an accurate proxy for the quantity of services provided. However, using the number of students as a measure for the activity level explicitly requires quality adjustment for the services delivered. This means that Denmark could improve the quality of volume indicators for secondary education by switching from measuring the number of students to the number of student-hours.

123. Higher education is organized differently from primary and secondary education. Compared to lower level education, attainments in higher education depend more on a student’s own efforts. Thus student-hours are not as useful a measure of output. An indicator that is often used to measure output is the number of students. Issues remain with this metric – changes in quality cannot be captured and participation in studies varies significantly (Statistics Denmark, 2016).

REFERENCES

- Bramesfeldt, A. et al. (2016), “Monitoring mental healthcare on a system level: Country profiles and status from EU countries”, *Health Policy*, Vol. 120(2016): 706–717.
- The Commonwealth Fund, (2017), The Danish Health Care System, <http://international.commonwealthfund.org/countries/denmark/>.
- European Commission, (2015), “Comparative Study on Quality Assurance in EU School Education Systems – Policies, procedures and practices”, www.schooleducationgateway.eu/files/esl/downloads/69_Comparative_Study_on_QA.pdf.
- KORA (2016), *Country Background Report – Denmark*, Danish Institute for Local and Regional Government Research.
- Mainz, J., Kristensen, S. and Bartels, P. (2015), “Quality improvement and accountability in the Danish health care system”, *International Journal for Quality in Health Care*, Vol. 27(6): 523–527.
- The Ministry of the Interior and Health (MIH), (2006), The Local Government reform – in brief, Department of Economics, Copenhagen.
- Nusche, D. et al. (2016), *OECD Reviews of School Resources: Denmark*, OECD Publishing, Paris.
- OECD (2017), *Understanding Public Sector Productivity*, Annual Meeting of OECD Senior Budget Officials background document.
- OECD (2015), *Fiscal Sustainability of Health Systems: Bridging Health and Finance Perspectives*, OECD Publishing, Paris.
- OECD (2013), *OECD Reviews of Health care quality: Denmark*, OECD Publishing, Paris.
- OECD (2011), *OECD Reviews of Evaluation and Assessment in Education: Denmark*, OECD Publishing, Paris.
- Statistics Denmark (2016), *General Government Output and Productivity, 2008-2014*, www.dst.dk/Site/Dst/Udgivelser/GetPubFile.aspx?id=18684&sid=gengov2014.
- Rambøll Management Consulting (2013), *Evaluering af de nationale test i Folkeskolen*, Copenhagen, Denmark, www.ramboll.dk/~media/Files/RM/Rapporter/Nationale%20test%202013.pdf (accessed August 2017).

ANNEX C. INDIA'S HEALTH PERFORMANCE INDEX

124. India's 1950 Constitution sets out powers for the central government, 29 states and 7 union territories, but confers maximum power to the central government, allowing the central government to take over the administration of states in times of emergency. In this context, India is a quasi-federation, combining features of both federal and unitary governments. States have jurisdiction over education, agriculture, public health, sanitation, hospitals and dispensaries and many other policy areas (Phukan, 2016). Local governments were also given powers under amendments to the Indian Constitution in 1992, although local governments are still dependent on allocations by State governments (Steytler, 2005).

125. The Constitution assigns the organisation and delivery of health services, including public hospitals, as the responsibility of State governments. The central government is responsible for international health treaties, medical education, prevention of food adulteration, quality control in drug manufacturing and national disease control. It also sets national health policy including the regulatory framework and supports the states (Gupta and Bhatia, 2016). Significant inequalities with respect to health care access and outcomes exist between India's states, rural and urban areas, socioeconomic groups, castes, and genders. There are also significant interstate differences in health outcomes. Also, the number of government hospital beds per population in urban areas is more than twice the number in rural areas, and urban areas have four times more health workers per population. Education is also a state responsibility according to the constitution, and controls held by the central government are very limited. However the central government monitors the delivery of education services by the states. Further, the National Council of Educational Research and Training (NCERT) was a resource organisation set up by the central government to assist governments on academic matters related to school education (TBC, 2014).

126. The National Institution for Transforming India (NITI Aayog), established in 2015, has been mandated the task of developing a shared vision of national development priorities with the active involvement of states. One of the main priorities of the NITI Aayog is to foster co-operative federalism amongst the states and to help the states act together in the national interest. NITI Aayog monitors the progress of central government policies and programmes and works closely with state governments. It aims to also disseminate best practices developed in one or more states or in other parts of the world to all states for possible adoption (NITI Aayog, 2017).

127. In 2016-17, NITI Aayog developed three indices for public service performance with the hope of invoking competition amongst State governments. The indices benchmark healthcare, education and water services, using an outcomes-based monitoring framework. The indices were developed through formal consultation with the Ministry of Health and Family Welfare, academic experts and state governments. The indices were also pre-tested in two states prior to finalisation. The data that underlies the indices are provided by the states and verified by an independent third party agency. The indices were published for the first time in September 2017, with the intention that they will be updated annually. The states are ranked on their overall performance on the indices as well as according to the sub-groups that form the index. Rankings are made publicly available through a web portal that allows users to generate custom of graphs and tables.

128. The indices are composite indicators developed by the NITI Aayog with technical assistance from the World Bank, and are calculated annually to highlight annual incremental improvement by the States. The composite indices are calculated by measuring the 'level' of public service performance of each State (calculated as a weighted average of the various sub-indicators). However the focus is on the incremental improvement in State performance, which is calculated as the change in the index from the base year to a reference year. The sub-indicators were selected based on their appropriateness and their availability through existing data sources.

129. India's 'Performance on Health Outcomes' index aims at rank States on the basis of their performance on measurable health indicators. The health index is constructed based on sub-indicators, grouped into 'health outcomes', 'governance and information' and 'key inputs/processes'. The 'health outcomes' category has the greatest weight (75 per cent) in the index. Indicators that form a part of the Sustainable Development Goal in health have been included in order to align these initiatives. Sub-indicators in the Health Index include:

- Number of still births per thousand live births;
- Proportion of low birth weight among newborns out of the total number of newborns weighed;
- Proportion of infants who have received certain immunisations during a specific year;
- Proportion of vacant health care provider positions in public health facilities.

130. Similarly, the School Education Quality Index is a composite indicator focused on outcomes: 29 sub-indicators are focused on educational outcomes grouped into learning, access and equity outcomes. And 15 sub-indicators are based on governance and measurement. It is hoped that the Education Index will encourage State-led innovation through cost-effective approaches, to shift the policy focus to educational outcomes and facilitate sharing of best practices across States. The majority of the indicators are based on published information. For the rest of the indicators, States submit information on the dedicated portal (with supporting evidence). Sub-indicators used as part of the Education Index include:

- Scores for language, mathematics and science on India's National Achievement Survey (which is a survey on a sample group of Indian students in different grades);
- Number of days taken by State government to release total Central share of funds to societies;
- The per cent of head-teachers/principals who have completed School Leadership training in the given financial year;
- Percentage differences between the boys' and girls' enrolments;
- Student retention rate at each school level.

131. The development of these composite indices will increase the information available to citizens on service performance, with the aim of improving the performance of these government services. This will be an iterative process, which will require evaluation, consultation and amendments as the performance system progresses.

REFERENCES

- Gupta, I. and M. Bhatia (2016), “The Indian Health Care System”, The Commonwealth Fund,
<http://international.commonwealthfund.org/countries/india/>.
- NITI Aayog (2017), “Working with the States: Cooperative Federalism”,
<http://niti.gov.in/content/cooperativefederalism>.
- Phukan, R.S. (2016), "What is State Government? Structure, Roles and Responsibilities",
www.mapsofindia.com/my-india/government/what-is-state-government-structure-roles-and-responsibilities.
- Steytler, N. (2005), “The Place and Role of Local Government in Federal Systems”,
www.kas.de/wf/doc/kas_8271-544-2-30.pdf.
- The British Council (2014), “Indian School Education System: An Overview”,
www.britishcouncil.in/sites/default/files/indian_school_education_system_-_an_overview_1.pdf.

ANNEX D. NORWAY'S KOSTRA SYSTEM

132. Norway has three levels of government, the Kingdom, counties (19) and municipalities (428). The municipalities and county governments represent a two tier-system of local government, with both authorities having the same administrative status. The central government has overriding authority and supervision over municipal and county administrations (NMLGM, 2014). Both the municipalities and the county authorities vary significantly in size, topography and population but all municipalities are given the same rights and responsibilities. Municipal responsibilities include primary and lower secondary school; primary healthcare and care for the elderly and disabled.

133. The Local Government Data Registration and Information Scheme (KOSTRA) is a data system designed to help the central government keep track of resource used in the local government sector, including the expenditures, activities and productivity. KOSTRA was established across some municipalities in 1995 and became compulsory for all municipalities in 2002. This was following revisions to the municipal accounting framework in 2001, which ensured accounting data included data on depreciation and operating costs. KOSTRA has ensured that a vast and complete range of financial and non-financial performance indicators are available to central and local governments in Norway (Statistics Norway, 2017). This information is quite detailed so that it can be seen not just how much money municipalities spend on different programmes but also how they spend it (OECD, 2012). Few other countries have such a system and KOSTRA is often held up as a model in international discussions.

134. The data collected through KOSTRA is used for research, analysis, planning and governing purposes. The key figures in KOSTRA provide information on most of the municipal and county municipal activities, including economy, schools, health, culture, the environment, social services, public housing, technical services and transport and communication (NMLGM, 2014). For example, municipalities must report the share of children having day care centre access, per capita expenditures on primary schooling, share of medical doctors per capita, and so on (Waeraas, 2014). Because of the vast amount of information, KOSTRA serves to benchmark Norwegian municipalities against similar peers. It also provides information for the general public, businesses, and the media that can be used for purposes of reputation rankings and decision-making. The KOSTRA database is available online and contains tables where one can find figures for each municipality, and comparisons with other local governments. It also contains information about how the key figures are defined and the data sourced.

135. KOSTRA also serves to aid local governments with planning and budgeting (Mizell, 2008). Johansen and Juuhl (2011) surveyed 140 Norwegian municipalities in order to deduce the use of KOSTRA data by Norwegian municipalities for internal management. The municipalities respond in the survey that KOSTRA is used to control underlying units, as a basis for decision making, for learning and for information to residents, users and the media. They also found that KOSTRA's data quality was not perceived as a major problem and that large municipalities use KOSTRA more than small municipalities. Further, by comparing their study with previous studies of KOSTRA, the use of KOSTRA for internal management has increased since the system became mandatory.

136. KOSTRA is based on electronic reporting of data from the municipalities, county authorities and data from other sources collected from within and outside Statistics Norway (SN), aiming to collect data

from different sources in a co-ordinated way. Each year, the local government sector submits finance and service production data in electronic form to SN. SN combines this material with various other types of information, for instance demographic data, and generates various key indicators for prioritisation of services, degree of coverage and productivity (NMLGM, 2014). Information on the past year is collected in February, and is published in a preliminary state in mid-March. After this time, municipalities have the opportunity to correct discrepancies or mistakes in their reported data prior to audited figures being published by SN in mid-June each year (Statistics Norway, 2017).

137. Its introduction has benefited central and sub-national governments and proved to be a useful tool for municipalities themselves. As well as informing policy-makers, publication of performance indicators and benchmarking exercises help to make regional and municipal governments more accountable to their residents, improve decision-making and drive service improvements (Lewis and Fall, 2016).

138. All reporting from municipalities and counties is on an electronic basis as well as the publishing of input and output indicators on local public services and finances. Publishing on the Internet includes a number of fixed indicators on the municipalities' priorities, productivity and coverage of needs. It is structured to enable comparisons of one municipality with the average for the comparable group, the region or the country. KOSTRA combines municipalities into comparable groups based on population, economical structure and demographics. In addition, the municipalities are organised according to revenue they have after covering costs for meeting minimum standards and statutory tasks. Comparisons of this type are useful to both central and local government authorities, since they may help identify potential service areas for which efficiency measures could be introduced in order to improve resource use. At least some municipalities are analysing these comparisons actively (OECD, 2012).

139. All of the local and county authorities report data. Reporting for the service areas varies from 88 to 100 per cent (Statistics Norway, 2017). The county authorities only lack a couple of forms. In addition to figures from Statistics Norway, data have been collected from several sources outside Statistics Norway, for instance from directorates that also collect data from the local authorities (Statistics Norway, 2017).

REFERENCES

- Johansen, M. and M. Juuhl, (2011), Norske kommuners bruk av KOSTRA til intern styring, Norges Handelshøyskole.
- Lewis C. and F. Fall (2016), “Enhancing public sector efficiency and effectiveness in the Czech Republic”, OECD Economics Department Working Papers, No. 1363.
- Mizell, L. (2008), “Promoting Performance – Using Indicators to Enhance the Effectiveness of Sub-Central Spending”, OECD Working Papers on Fiscal Federalism, No. 5, OECD Publishing, Paris.
<http://dx.doi.org/10.1787/5k97b11g190r-en>.
- Norwegian Ministry of Local Government and Modernisation (NMLGM),(2014), Local Government in Norway, www.mddsz.gov.si/fileadmin/mddsz.gov.si/pageuploads/dokumenti_pdf/-enake_moznosti/NFMStudijski2LokalnoUpravljanje.pdf.
- OECD (2012), OECD Economic Surveys: Norway, OECD Publishing, Paris.
http://dx.doi.org/10.1787/eco_surveys-nor-2012-en.
- Statistics Norway (2017), KOSTRA – Municipality-State-Reporting 2016, preliminary figures,
www.ssb.no/en/offentlig-sektor/statistikker/kostrahoved/aar-forelopige/2017-03-15?fane=om#content.
- Waeraas, A. (2014), “Municipal Reputation-building in Norway: A reputation common stragedy?”, in Organizational Reputation in the Public Sector, New York, NY: Routledge.

ANNEX E. THE USE OF PUBLIC PERFORMANCE DATA IN THE UNITED KINGDOM'S SCHOOL SYSTEM

140. Many countries have moved towards competitive benchmarking, with more countries releasing information about the performance of hospitals, schools and local governments through published league tables. This has been especially prevalent in the United Kingdom through the governments' use of performance information to encourage active citizen participation. Benchmarking through league tables can provide explanations and more detailed information than just raw numbers in order to help citizens choose among different service providers. This information, while not perfect, can at least provide some guidance with regard to the level of service provision and performance. The public availability of this information, and citizens actions based on these data, can serve to place the spotlight on under-performing service providers and thereby serve as a motivator for future action to improve performance (Curristine et al., 2007).

141. The English education system has historically been subjected to high levels of public performance monitoring. The current system involves two reporting streams: the publication of Office of Standards in Education, Children's Services and Skills (OFSTED) reports and the publication of summary performance indicators in the form of league tables. OFSTED inspects a range of educational (and other) institutions and rates schools on a four point scale from 'outstanding' to 'inadequate', further providing detailed reports on specific school aspects, including the 'quality of teaching, learning and assessment' and 'personal development, behaviour and welfare'. The league tables are available online, with easy tools available to compare similar schools across England.

142. The aim of the English performance league tables was twofold: to create competition so as to incentivise schools to improve performance and educational standards; and to provide information with the view that parents would choose the best schools for their child. Karsten et al. (2001) found that (middle-class) parents in England use the published performance information. However they could not conclude that it had a major influence on their choice of school. Burgess et al. (2015) investigate parents' references for school attributes and showed strong preferences for schools with high academic attainment, supporting the idea that competition among schools should help raise academic standards. However, they also found that families from lower socio-economic quintiles are more likely to choose schools with a lower academic standard, as other factors like location are important.

143. Bevan and Wilson (2013) analysed school performance in England and Wales following the abolition of the public dissemination of school performance information in Wales. They found a negative impact on school performance in Wales relative to England, with this effect concentrated in the lower 75 per cent of schools (as measured by student prior attainment and by poverty). They concluded that the key driver for improved performance in England came from the reputation effects of 'naming and shaming'.

144. It is widely acknowledged that non-cognitive skills also play an important role in determining social outcomes (Palczyńska and Świst, 2016; Almlund et al., 2011). It can be extremely difficult to measure these skills and unfortunately 'what gets measured gets done'. Studies of the English school system have shown that pupils moved from being 'learning orientated' in the 1980s to being 'performance-

orientated' (Pollard and Triggs, 2000). National testing in primary and secondary (which are generally focused on numeracy and literacy) is prevalent in many countries, and often these results are publicly released. A test is high-stakes when its results are used to make important decisions that affect students, teachers, administrators, communities, schools, and districts (Madaus, 1988).

145. High-stakes testing has been shown to improve test scores, but research suggests that this could be due to teachers focusing on preparing students for national tests. Some research has shown a 'narrowing' effect on the curriculum, with the proportion of time spent on subjects such as music, physical education and technology decreasing over time, pointing to the implementation of high-stakes testing (Fitz-Gibbon, 1997; Goodbody, 1999; Wiggins and Tymms, 2002; Au, 2007). However, much of the research in this area is based on surveys of teachers and educators, rather than using robust empirical techniques. Further, the specific construction of tests affects teachers' curriculum responses. Tests that do not promote drill and rote learning are obviously less likely to affect the design of curriculums.

REFERENCES

- Almlund, M., Duckworth, A., Heckman, J., and Kautz, T. (2011), “Personality Psychology and Economics”. In *Handbook of the Economics of Education*, Vol. 4: 1-181.
- Au, W. (2007), “High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis”, *Educational Researcher*, Vol. 36(5): 258–267.
- Bevan, G. and D. Wilson (2013). “Does “Naming and Shaming”, Work for Schools and Hospitals? Lessons from Natural Experiments Following Devolution in England and Wales”, *Public Money and Management*, Vol. 13(4): 245-252.
- Burgess, S. et al. (2015), “What Parents Want: School Preferences and School Choice”, *Economic Journal*, Vol. 125: 1262–1289.
- Fitz-Gibbon, C. T. (1997), *The Value Added National Project: Final Report* (School Curriculum and Assessment Authority, London).
- Goodbody, J. (1999), “Why PE is in decline”, *The Times* (27 October).
- Karsten, S., A. Visscher, and T. De Jong (2001), “Another Side to the Coin: The Unintended Effects of the Publication of School Performance Data in England and France”, *Comparative Education*, Vol. 37(2): 231-242.
- Madaus, G.F. (1988), “The Influence of Testing on the Curriculum”, in L.N. Tanner (ed.), *Critical Issues in Curriculum*, Eighty-seventh Yearbook of the National Society for the Study of Education (Chicago: University of Chicago Press).
- Palczyńska, M. and K. Świst (2016), “Measurement Properties of Non-cognitive Scales in the Polish Follow-up Study on PIAAC (POSTPIAAC)”, *OECD Education Working Papers*, No. 149, OECD Publishing, Paris, <http://dx.doi.org/10.1787/c533e448-en>.
- Pollard, A. and P. Triggs (2000), *Policy, Practice and Pupil Experience*, Continuum International Publishing Group, London.
- Wiggins A. and P. Tymms (2002), “Dysfunctional Effects of League Tables: A Comparison between English and Scottish Primary Schools”, *Public Money & Management*, Vol. 22(1): 43-48.

ANNEX F. THE UNITED STATES' NO CHILD LEFT BEHIND ACT, 2001-2015

146. In the United States, compulsory education is primarily a state and local-level responsibility, with around 90 per cent of funding for elementary and secondary schools coming from non-federal sources (USDE, 2017). Every state has its own laws regulating finance, the hiring of school staff, requirements for enrolment and graduation, student attendance, the number of years of compulsory education and the curriculum (Corsi-Bunker, 2014; USDE, 2017). Within the states, the education system is further divided into local school districts, which are managed by school boards, and are generally responsible for coordinating education policies. Individual schools can also be delegated various freedoms and responsibilities (Corsi-Bunker, 2014).

147. Given the United States' decentralised education system, it can be extremely difficult for the national government to drive systemic change and reform, especially given the differing competencies and administrative capacities across states. McGuinn (2012) defines this as the 50/14 000/130 000 problem – the United States has 50 different state systems that collectively contain around 14 000 school districts and almost 130 000 schools. States have developed vastly different education and accountability systems, resulting in variation in school quality within and across states.

148. The federal government can only indirectly attempt to drive education reform by incentivising states through funding or public criticism/praise. Although such incentives are in place, the federal government has often struggled to convince states to implement reforms fully or effectively (Shelly, 2008). This has greatly limited the federal government's capacity to drive educational reform and educational equity across the United States.

149. As the first major federal government involvement in elementary and secondary education, the *Elementary and Secondary Education Act* (ESEA) was launched in 1965, which involved a comprehensive set of programs including federal (Title I) funding to disadvantaged children (USDE, 2017). Unfortunately many states avoided the ESEA in various ways, by designing diffuse standards and creating unclear metrics of success (Solace, 2013). In order to overcome the weaknesses of previous ESEA policies, the United States Congress passed the *No Child Left Behind* (NCLB) *Act* in 2001, which became law in January 2002. NCLB was a reauthorisation of the ESEA, but included further requirements for testing, accountability, and school improvement.

150. The aim of the NCLB programme was to boost student academic achievement through a top-down, benchmarking strategy based on the assumption that setting high standards could improve student outcomes. The Act stated that the NCLB policy aimed “to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging state academic achievement standards and state academic assessments” (NCLB, Section 1001).

151. Below is a summary of the main requirements under the NCLB Act:

- that all students be taught by “highly qualified” teachers;
- that all students be judged at least proficient on State devised academic standards by 2013-14;
- that students be assessed through state-wide testing in reading and mathematics;
- that test results are publicised via school ‘report cards’;
- that a specified level of yearly progress be achieved; and
- that schools failing to make specified progress be subject to a variety of corrective actions that escalate each year, including technical assistance and the replacement of staff (Rudalevige, 2005).

152. Under the NCLB programme, states were required to test specified subjects and grades, report results and establish minimum performance standards for students in return for federal funding. NCLB required states to test students in reading and mathematics annually in Grades 3 to 8 and once in Grades 10 to 12; also, in science, once in elementary, middle, and high school (Shelly, 2012). Critics argued that high-stakes accountability policies through state standardised testing encouraged teachers to teach a narrow subset of skills. Teachers or schools may (sometimes unintentionally) narrow instructional effort by targeting the skills required by the high-stakes State assessment, at the expense of broader and more genuine improvements in cognitive achievement. The NCLB policies appear to have led elementary schools to increase instructional time devoted to mathematics and reading, although the majority of evidence on this issue is sourced from teacher and administrator survey data that are subject to potential bias (Dee and Jacob, 2010).

153. Student exam performance was categorised as ‘below basic’, ‘basic’, ‘proficient’, or ‘advanced’ (Solace, 2013). At least 95 per cent of students in each demographic subgroup was required to take the exams, and, at the end of the 2013-14 school year, all students were required to reach the ‘proficient’ level, or schools would risk losing federal funds (Manna, 2006). This incentivised schools to reallocate instructional effort away from high and low performing students and toward the students that were most likely to meet the proficiency standard with additional teacher attention (Dee and Jacob, 2010). Under the NCLB policy, schools had to also meet federally set “adequate yearly progress” (AYP) targets that ensured timely progress toward this goal, including specific AYP goals for demographic subgroups such as special education students, racial and ethnic minorities, and English language learners. The federal government could provide assistance or impose escalating sanctions/requirements on schools and districts that did not meet performance goals, including giving students the ability to transfer and providing supplemental educational services such as private tutoring (Shelly, 2012; Goertz, 2005).

154. In addition to the above requirements and consequences, benchmarks and results were widely disseminated to parents and the press. Federal policymakers argued that publicly disseminating performance results would weaken States’ ability to skirt student achievement, essentially shaming them into adopting the federal government’s focus on academic outputs (Solace, 2013). However, political resistance and capacity gaps at the sub-national level meant that educational reforms were often more artificial than substantive (McGuinn, 2012). As a result, the law did not generate as much meaningful school improvement in closing student-achievement gaps as was originally anticipated (Mintrop and Sunderman, 2009).

155. These incentives created a choice for states: avoid the NCLB sanctions by setting low standards for student performance or avoid NCLB sanctions by setting high standards and raising sub-national taxes to ensure that these standards can be reached (Duncombe et al., 2006). This arises because the NCLB did not actually set national standards for the States to meet, and was also likely due to inadequate funding to

meet NCLB objectives. Indeed, state governments were vocal critics of the NCLB, wanting much more federal money and much more freedom to spend it (Rudalevige, 2005). Duncombe et al. (2006) looked at the impact of the NCLB reforms in Missouri. In January 2006, the Missouri State Board of Education approved new cut-offs for the state tests that “should result in more students scoring at the ‘proficient’ and ‘advanced’ levels” (Missouri Department of Elementary and Secondary Education, MDESE, 2006). In 2005, 30 per cent of students reached proficiency in reading and 25 per cent in math; under the new standards these percentages were expected to rise to 44 and 43 per cent, respectively (Duncombe et al., 2006).

156. One other way that states aimed to avoid sanctions was by seeking amendments to state accountability plans from the federal government. Empirical work showed that larger, more affluent (that is higher mean income) states, and more Republican-governed states submitted more amendment requests to the federal government. The federal Department of Education granted a higher percentage of requests to more Republican states and states with a more developed system of standardised testing prior to NCLB implementation. The size and sophistication of state departments of education may explain these results. Larger and more affluent states possess the resources to dedicate more time to understanding NCLB and the areas in which a request for flexibility looks promising. Less populous and less affluent states may have smaller and under resourced education bureaucracies (Shelly, 2012).

157. States also had flexibility in setting the growth trajectory for student improvement and choosing their own accountability indicators, creating various indicators across States. California used a composite indicator called an Academic Performance Index (API) using standardised testing and exit exams on English, mathematics, and social studies tests (Goertz, 2005). The API assigned one number to a school on a scale of 200 to 1000, with a score of at least 800 as the goal (EDP, 2013). Most schools were required to improve their performance each year by 5 per cent of the difference between their API and the state-wide target of 800 (EDP, 2013). In contrast, North and South Carolina used a ‘value-added’ measure. This aimed to track the progress of individual students over time. Kentucky’s Accountability Index included student performance in seven academic areas, attendance, retention, and dropout rates (Goertz, 2005). This meant that Kentucky averaged test scores across subject areas beyond reading and mathematics.

158. Some studies have aimed to analyse the effect of NCLB on student performance. Unfortunately the lack of a credible control group that allows studies to distinguish the effects of the federal reforms from other time-varying factors makes this difficult. Dee and Jacob (2009) provide evidence that the NCLB reforms had some positive impacts on student achievement. They analysed state-level panel data on student test scores from the National Assessment of Educational Progress. The impact of NCLB is identified using a comparative time series analysis that relied on comparisons of the test-score changes across states that already had school-accountability policies in place prior to NCLB and those that did not. They show that NCLB generated statistically significant increases in the average mathematics performance of 4th graders as well as improvements at the lower and top percentiles. The benefits also appear to be concentrated among traditionally disadvantaged populations, with particularly large effects among Hispanic students. There was also evidence of improvements in 8th grade mathematics achievement. However, they found no evidence that NCLB increased reading achievement.

159. Benchmarking was in effect the only choice available to the national government, given the federal government’s lack of power in the education space and the states’ avoidance of previous legislative actions (Solace, 2013). However the fact that the NCLB Act did not posit national standards but instead required each State to develop their own educational standards ultimately diluted certain aspects of the programme. This top-down approach led to some States disregarding the spirit of the framework. Despite criticism, the NCLB policy was able to remain in place for around fourteen years, until it was replaced with the *Every Student Succeeds Act* in 2015.

REFERENCES

- Corsi-Bunker, A. (2014), "Guide to the Education System in the United States", International Student and Scholar Services, University of Minnesota.
- Dee, T. and B. Jacob (2010), "The Impact of No Child Left Behind on Students, Teachers, and Schools", Brookings Papers on Economic Activity, No. 2.
- Dee, T. and B. Jacob (2009), "The impact of No Child Left Behind on student achievement", NBER Working Paper No. 15531, National Bureau Of Economic Research Massachusetts Avenue.
- Duncombe, W., A. Lukemeyer and J. Yinger (2006), "The No Child Left Behind Act: Have Federal Funds Been Left Behind?", Center for Policy Research, Education Finance and Accountability Program.
- Education Data Partnership, (2013), Understanding the Academic Performance Index (API), [www.ed-data.org/article/Understanding-the-Academic-Performance-Index-\(API\)#a4](http://www.ed-data.org/article/Understanding-the-Academic-Performance-Index-(API)#a4).
- Goertz, M. (2005), "Implementing the No Child Left Behind Act: Challenges for the States", Peabody Journal of Education, Vol. 80(2): 73-89.
- Manna, P. (2006), "School's In: Federalism and the National Education Agenda", Georgetown University Press, Washington, DC.
- McGuinn, P. (2012), "Stimulating Reform: Race to the Top, Competitive Grants and the Obama Education Agenda" Vol. 26(1): 136-159.
- Mintrop, H. and G. Sunderman (2009). "The Predictable Failure of Federal Sanctions–Driven Accountability of School Improvement—And Why We May Retain It Anyway", Educational Researcher. Vol. 38(5): 353–64.
- Missouri Department of Elementary and Secondary Education (MDESE), (2006), State Board of Education revises scoring standards for MAP exams, News Release.
- No Child Left Behind Act of 2001, Pub. Law No. 107–110.
- Rudalevige, A. (2005), "Adequacy, Accountability, and the Impact of No Child Left Behind", Dickinson College, Research Paper No. 05-27, Harvard University, Department of Political Science.
- Shelly, B. (2008), "Rebels and Their Causes: State Resistance to No Child Left Behind", Journal of Federalism, Vol. 38(3): 444-468.
- Shelly, B. (2012), "Executive Federalism and the No Child Left Behind Act of 2001", Educational Policy, Vol. 26(1): 117-135.
- Solace Foundation Imprint (2013), "Benchmarking for Improvement", www.solace.org.uk/knowledge/reports_guides/Benchmarking_For_Improvement.pdf.
- US Department of Education (2017), "The Federal Role in Education", www2.ed.gov/about/overview/fed/role.html.