Abstract

Questions concerning *influence* are essential in the field of evaluation. This article argues for a dual perspective in this discussion. *Theories of influence* (Henry and Mark, 2003; Kirkhart, 2000) are combined with the idea of *The Evaluation Society* (Dahler-Larsen, 2012), and together these create a broad perspective on the use of evaluations and their possible impact. From this dual perspective, evaluation is recognized both as a method for planning and development and as an independent societal phenomenon. Norwegian participation in the Programme for International Student Assessment (PISA) is used to illustrate the essence of this dual perspective and how it can yield a more complete understanding of the complexity of evaluation impact.

Keywords:

evaluation influence, evaluation use, evaluation society, PISA, evidence based policy, social betterment

1. Introduction

One reason why evaluation should be a subject of research has to do with the questions concerning its potential influence. The way we define and understand evaluation determines how we answer these questions. Evaluation may be viewed as a method for planning and developing social improvements (Henry and Mark, 2003: 295), which reflects its intended or manifest function. Evaluation could also be seen as a societal phenomenon, as an omnipresent, powerful and inescapable trend (Dahler-Larsen, 2012: 9), criticized for its strong connection to neoliberal ideology and the new public management approach (Lindgren, 2014; Power, 1997). These two perspectives represent different academic traditions. The aim of this

This is an accepted manuscript of an article published by SAGE Journals in Evaluation on November 14, 2018.
Høydal, Ø. S. Evaluation: Method and societal phenomenon.
Evaluation: International Journal of Theory, Research and Practice (Vol 25, Issue 3), p. 381–398.© The Author(s) 2018.
http://dx.doi.org/10.1177/1356389018804263

article is to show why they are useful together and why such a combination could improve our understanding of evaluation's influence in modern societies.

The article is based on theoretical perspectives from *the evaluation society* (Dahler-Larsen, 2012) and the analytical frameworks outlined in the *theories of influence* by Kirkhart (2000), Mark, and Henry (Henry & Mark, 2003; Mark & Henry, 2004). While the theories of influence recognize evaluation as a valuable method for knowledge production, they represent an attempt to broaden the traditional understanding of the use and usefulness of evaluations by focusing on their total influence. This focus include an analytical framework developed to cover the most central aspects of evaluation influence: the intension behind evaluation use, the impact from the results and processes of evaluating, and the influence of these elements on different levels of society, over time.

While *the theories of influence* aim to reveal the complex pattern of evaluation consequences, *the evaluation society* is a theoretical sociological concept that seeks to describe *why* we are evaluating (Dahler-Larsen, 2012:21). According to Dahler-Larsen, the present western evaluation wave is a consequence of evaluation as a societal phenomenon. Dahler-Larsen reveals how, through different stages in the history of modernity, evaluation has been more of a cultural expression, reflecting the values of society, than a method for social change through knowledge production. This kind of broader cultural or societal perspective on evaluation is generally ignored in evaluation literature, which has focused primarily on theoretical aspects of evaluation as an instrument, or on the search for improved evaluative methods (Carden & Alkin, 2012:103; Dahler-Larsen, 2012:4).

When combining the two theories, the theories of influence reveal that evaluation is a complex process, often including elements that are not logical or planned for. The cultural perspective found in the theory of the evaluation society could help explaining these phenomena. For instance, Dahler-Larsen (2012) claims that the cultural aspects of evaluation

could explain why we keep on evaluating even when evaluations rarely lead to political change or when the evaluation results are not used at all.

I will first present the two perspectives and the analytical framework for this article. Then Norwegian participation in the Programme for International Student Assessment (PISA) is used as an example to illustrate the benefits of seeing evaluations from two sides. This study is not an empirical examination of all the effects of PISA, but rather, it uses PISA to illustrate how evaluation can be both an influential method and a societal phenomenon. The PISA evaluation provides a good illustration for several reasons:

- The example shows how evaluations can influence a national discourse and a political field.
- It provides an opportunity to follow the long-term effects of evaluation.
- The example illustrates how evaluations can have an impact on different levels in a society.
- The fact that PISA involves both national authorities and the international OECD is an example of the complexity involved in any clear outline of evaluation intentions.

2. Theoretical Framing

2.1. The evaluation society

"We live in the age of evaluation," Danish sociologist Peter Dahler-Larsen claims in his book *The Evaluation Society* (2012). According to Dahler-Larsen, in Western societies, all aspects of everyday life are constantly being measured and tested. We are surrounded by feedback loops in work-life, in the educational system, in the health sector and in our role as consumers (pp.1-2). Evaluation is impossible to escape or reject; it has become an institution and a "protected discourse", something taken for granted and not to be questioned. Evaluation is framed as a "ritual supported by societal norms" in which the meaning of these norms is more important than the empirical evidence for the instrumental use of evaluations. The

evaluation society, he notes, is mainly concerned with seeing that evaluations are done and less concerned about whether they are useful or of high quality (p.27).

The evaluation activity is part of a larger societal trend that also includes methods such as auditing, inspection and accreditation (p.3). As Dahler-Larsen argues, human beings may always have reviewed their efforts and made judgements about their activities, but such evaluative activity cannot be compared with today's *evaluation wave*, which is enriched with resources, attention and influence (p.17).

Dahler-Larsen looks for the forces behind this wave and asks what these forces might be when evaluative activities in fact do not seem to be driven by instrumental motives (p.21). The central task in the book is therefore to find the values and ideas behind evaluation and to understand the meaning produced in evaluation. Dahler-Larsen investigates this question through what he calls three *sociohistorical stages of modernity*: modernity, reflexive modernity and the audit society. Each stage represents the purpose and meaning of evaluation in light of the society where evaluation takes place. Sticking to the wave metaphor, Dahler-Larsen claims that these stages could be seen as defined epochs, or trends that have dominated certain times and then left sediments behind (pp. 99-100).

According to Dahler-Larsen, evaluation is a child of the rationalist belief in societal progress through the use of reason and science, characterizing the age of modernity (p.105) (see also Albæk, 1988). However, when planning for social betterment turned out to be practically complicated, the optimistic belief in progress through rationality was gradually replaced by doubt, uncertainty and suspicion – and *reflexive modernity* [1] (Dahler-Larsen, 2012:141). In reflexive modernity, society is struggling with modernity's side effects, such as pollution, stress and welfare dependency. As a result, public sector and other organizations have to evaluate so that they appear to take these side effects seriously (p.144). In a time of

[1] Beck, Giddens and Lash, 1994.

uncertainty and dilemmas, evaluation becomes a symbol of good management, knowledge-based decision-making, reflection and trustworthy leadership. However, evaluation also becomes a ritual because even if evaluation is demanded, the actual use of the results is not in focus (pp. 145-147).

While reflexive modernity is characterized by dilemmas and uncertainty, the essence of the audit society[2] is the handling of these consequences by risk management and control.. "The state strikes back" and takes a more active role when it comes to security issues, but also with regard to education. PISA reflects this new situation (pp.169-173). In the audit society, evaluation gives the illusion that problems are solved and under control, producing comfort (p.175). While the mismatch between the rationalist evaluation ideal and its role as a ritual gradually became problematic in reflexive modernity, evaluation in the audit society is standardized, routine-based and integrated in other organizational procedures (pp. 176-177). Dahler-Larsen uses the term *evaluation machine* as an analytical metaphor for the evaluation activity in the audit society (p. 176). Evaluation machines are characterized by emphasizing risk management, quantification and objective standards, and the reality has to adjust or be adjusted to fit the new evaluative procedures (p.182). Leading to constitutive effects of evaluation, a term referring to the way evaluations contribute to new definitions and practices, based on a construction of social reality (pp.198-199). Constitutive effects could be exemplified by the fact that standards and definitions used in evaluations of public programmes end up as new public standards.

By explaining the use of evaluation in relation to the shifting ideals in society, Dahler-Larsen also illustrates the flexible nature of evaluation and the challenge associated with a narrow definition of this concept. According to Dahler-Larsen, the ambiguities in the history

[2] Power, 1997.

of evaluation should also lead the evaluation industry to more modest ambitions and promises (p. 230).

2.2. Theories of influence

Definitions and language used to describe evaluation have changed over time. In the last decade, there has been a shift from the instrumental understanding of evaluations' *use* to a broader focus on their *influence* (Herbert, 2014). In the first evaluation wave after the Second World War, the use of evaluations were believed to have a direct and beneficial impact on political decisions (Albæk, 1988). During the 1970s, the narrowness of this instrumental concept of evaluation was recognized. Consequently, a system of sub-categories emerged: *symbolic*, *conceptual* (or *enlightenment*), and *process* use. *Symbolic use* refers to using the simple fact that an evaluation was done "… as a rational basis for action (or inaction), or to justify pre-existing positions …" (Henry & Mark, 2003:294). *Conceptual use,* or *enlightenment*, refers to a situation in which the evaluation results create new attitudes and knowledge among actors, but no direct action occurs. This influence may occur immediately after the evaluation or over years (Alkin &Taut, 2003:5; Weiss, Murphy-Graham, & Birkeland, 2005:14). *Process use* refers to the impact of the evaluation process itself, the fact that an evaluation occurred (Patton, 1998: 225-233).

Even after the introduction of sub-categories, the concept of use has been criticized for being too imprecise (Henry & Mark, 2003:11; Kirkhart, 2000:6). For example, the concept excludes anything about the possibly unintended consequences of an evaluation or the impact over time (Kirkhart, 2000:6). As a solution to these shortcomings, and to establish a broader perspective on the impact of evaluations, Kirkhart (2000) introduced the concept of influence:

> The term *influence* (the capacity or power of persons or things to produce effects on others by intangible or indirect means) is broader than use, creating a framework with which to examine effects that are multidirectional,

> incremental, unintentional, and non-instrumental, alongside those that are unidirectional, episodic, intended, and instrumental (which are well represented by the term *use*). (Kirkhart, 2000:7)

The theory of influence includes three dimensions: the source, the intention, and time. The *source* refers to the element of the evaluation that is presumed to generate change, such as the process and/or the results. The *intention* dimension has to do with both the intended and unintended influences of the evaluation, and the *time* refers to whether the evaluation has an immediate, end-of-cycle and/or long-term influence (Kirkhart, 2000:17). Thus, Kirkhart introduced a simple but effective system to pinpoint possible impact factors.

Henry and Mark (2003) use Kirkhart's work as a starting point to take the theory of influence a step further. On the existing dimensions, they add three different levels where evaluation activity and outcome may have influence: the individual, the interpersonal, and the collective level. The individual level refers to the evaluation's influence on each stakeholder or participant. The interpersonal level is about "…a process or outcome that predominantly takes place within interactions among individuals" andthe *collective level* refers to the influence of evaluation on practice and decisions on an organizational level (p.298).

2.3 Theories –limitations and overlap

The concept of influence was introduced to offer a "broader perspective on evaluations"(Kirkhart, 2000:7). However, like all frameworks, it could be criticized for presenting a too-neat picture of the evaluation process and for being limited when it comes to categories. Still, one could argue that the frameworks present the central aspects of evaluation influence and could be adjusted to match varied specifications.

Even though the ambition behind the concept was broad and inclusive, Kirkhart and Henry and Mark seem to have difficulties disentangling themselves from the historical, instrumental biased focus of their own field. Their primary concern seems to be how the

theory of influence could strengthen the brand of evaluations and increase their *positive* use.Henry and Mark (2003) stated their intention to …guide research on the influence of evaluation" and to provide a tool for "maximizing the influence of a specific evaluation" (p.311). According to Henry and Mark, *social betterment* is the ultimate purpose of evaluation (Henry and Mark, 2003; Mark and Henry, 2004, 2013). Social betterment refers to the improvement of social conditions, or "…to bringing about a state that would be considered as better than the state that existed before, as judged though deliberation and by public opinion" (Henry and Mark, 2003:295). Even when recognizing the complex aspects of defining social betterment for society as a whole, they still claim it should be "the guiding star" for evaluation (Saunders 2013:146).

Kirkhart's (2000:19) focus has been on the total impact of evaluation - good or bad, intended or unintended. However, she is also clearly concerned about the status of evaluation and the profession of evaluators. For instance, arguing that the concept of influence might help pinpoint evaluation practice's pervasive and unrecognized impact, which could build credibility and generate support among service delivery professionals (Kirkhart, 2000:20).

Maybe this pro-evaluation origin has hindered the spreading of the theories of influence. According to Herbert (2014), the research surrounding the influence concept is largely performed by evaluators who have examined their own projects and self-reporting by stakeholders who may seek to present their organizations as "receptive to evaluation evidence" (p. 412). Herbert argues that, because of this, theories of influence have not yet led to research with implications for evaluation practice, although they have been discussed and employed for more than 15 years (p.412). This could be used as an argument for a new and more critical use of the theories.

PISA could be described as what Dahler-Larsen (2012) calls an evaluation machine, which is commonly criticized for its many unintended effects (p. 201). However, contrary to

Kirkhart (2000) and Henry and Mark (2003), Dahler-Larsen is not interested in unintended effects but the constitutive effects following from evaluations, the fact that the world changes as being measured (p.216). He describes how constitutive effects might emerge, defining worldviews, social identities and relations and time frames, and also how constitutive effects "…may be displaced over time and across levels of analysis" (pp. 205-212). This framing of possible constitutive effects has much in common with the analytical framework presented in the theories of influence. Both Kirkhart (2000) and Dahler-Larsen (2012) believe influence is reciprocal, not unidirectional so that constitutive effects not only define, but also are defined by the context of use. This is what Dahler-Larsen labels *second-order construction.* To explain this phenomenon, Dahler-Larsen uses academic tests as an example. If teachers deliberate use strategies to influence children´s test scores, the meaning of these scores will change over time (p.217).

Dahler-Larsen's claim that we live in a world of evaluation may be no less true today than it was when written in 2012. However, on an international level, the notion of evaluation as a "protected discourse" that is "virtually sacred" has been disrupted by the election of political leaders who seem to have little respect for empirical inquiry of any kind. In the same way Dahler-Larsen´s use of society as singular, could be criticised for not taking into account the fact that society is not experienced the same by all its citizens.

One could discuss whether the use of evaluation within Dahler-Larsen's concept *the evaluation society* overlaps with what is referred to as *symbolic use*. For instance when evaluation is used to create an illusion of control or knowledge-based policy (Dahler-Larsen 2012:142, 185).

3. Analytical framework

The ideas of Kirkhart (2000) and Henry and Mark (2003) are used as an analytical framework for this article. The categories from the theories of influence are applied to explore and illustrate evaluation's significant role and impact in modern societies. Although Henry and Mark originally were concerned with the processes leading to change, for this article I have drawn upon their contribution to illustrate the actual changes following an evaluation. This approach can reveal how both the evaluation process and results might lead to change:

- The dimension of intention clarifies how the ideas behind an evaluation may vary among stakeholders and how evaluations can lead to both intended and unintended results.
- An evaluation is often understood to affect immediate results. The time dimension allows us to see how the evaluation's influence can extend for years and may even change in character over time.
- Henry and Mark's three levels of influence (individual, interpersonal and collective) can reveal how one aspect of an evaluation may affect different levels in the same society. One aspect might also be received differently across citizens who differ in social class, skin colour or religion.

While the theories of influence are used to reveal and investigate the complexity of evaluation, the concept of the evaluation society works like a meta perspective, framing the evaluative activity in the norms and values of society. In addition, illuminating how society itself is framed by the evaluation wave, and how evaluations are influencing life in general. The combination of the two perspectives also illustrates the reciprocal character of influence, meaning that constitutive effects not only define, but also are defined by the context of use (Kirkhart, 2000; Dahler-Larsen 2012).

The relationship between the theories of influence and the concept of the evaluation society

Model

Norwegian participation in PISA is applied as an illustrative example of the advantage of this two-sided perspective for understanding the influence that evaluation wields in modern societies. The PISA example is based on existing research, articles in the media concerning the PISA debate, and reports and public information from Norwegian authorities and the Organization for Economic Co-operation and Development (OECD).

4. Case presentation

4.1 PISA

PISA is an international survey-based evaluation organized by the OECD. The aim is to evaluate education systems by testing the skills and knowledge of 15-year-old students. Norway joined the first testing in 2000 and has implemented subsequent PISA evaluations every three years. All evaluations have indicated that Norwegian students are average in their skills and knowledge (Department of Teacher Education and School Research, n.d.). The PISA results presented in December 2016 were the first results showing Norwegian students ranking just above the OECD average (Kjærnsli and Jensen, 2016). However, because no major changes in the students' performance have taken place over the years, and because the influence from all the evaluations are tangled together, the whole sequence of testing is applied as an illustrative example.

4.2 Influence – source, intention, and time

4.2.1 The source of influence

Before PISA, Norwegian authorities assumed that a solid economic foundation would guarantee student performance at a high level (Norwegian Directorate for Education and Training, 2011). The first PISA results in 2001, indicating that Norwegian students ranked below the other Scandinavian countries and the OECD average (Baird et al., 2011:24; Norwegian Directorate for Education and Training, 2011), therefore induced a national trauma, named "the PISA shock" by the media. In short, Norway had spent a lot of money on education and believed its students would perform better than they did (Elstad and Sivesind, 2010: 22). PISA became what Dahler-Larsen (2012) describes as a *magic mirror,* where the society saw itself and its realities anew (p.205).

The surprising PISA result was the start of a new era in the history of Norwegian educational policy. Debates and disagreements regarding the quality of the educational system had occurred previously, but they were nothing compared with the storm that followed the first PISA results. Overnight, PISA was given the constitutive power (Dahler-Larsen, 2012) to determine that Norwegian students were not performing well enough. In the following years, every presentation of new PISA results have led to a massive debate about the quality of the educational system, but also about whether PISA should have the power to define this quality. The PISA results led to a series of reforms in both curriculum and assessment: a national quality assessment system (NKVS) introduced in 2004, and the national school reform, *Kunnskapsløftet,* in 2006 (Baird et al., 2011: 25; Norwegian Directorate for Education and Training, 2011). Later, a separate system of national and regional testing was introduced in reading, mathematics and English (Baird et al., 2011:23). These tests have been criticized for being too strongly inspired by the PISA concept. Some see them as rehearsals for the PISA evaluations rather than methods for mapping the students' achievement levels, needs and challenges (Kulbrandstad, 2010:193). Partly as a response to this criticism, diagnostic

"mapping" tests were established to offer better data at the individual level (Baird et al., 2011: 23).

The new focus on testing and the changes in the educational system following from this paradigm are themselves exerting influence every day. Especially in Oslo, the new national testing has led to a strongly results-orientated system where the schools are competing to present the highest scores (Marsdal, 2011:26-36). Teachers claim this competition has included extensive up-front drilling and even cheating (Marsdal 2011:62-66). When schools are adjusting their teaching and morality in order to fit evaluative procedures, this illustrates the constitutive effect of evaluations and the evaluation society (Dahler-Larsen 2012). While cheating clearly could be described as an unintended negative effect, teachers and school leaders have different opinions when it comes to the focus on up-front drilling (Marsdal, 2011). Either ways, this could lead to what Dahler-Larsen labels *second-order construction*. If the cheating goes too far or the upfront drilling leads to neglecting of other academic subjects, the meaning of the test scores will change over time.

The international enthusiasm for quantitative evaluations like PISA and the impact following could be seen as a reaction to the evaluations in the reflexive modernity, where the practical consequences of evaluating were believed to be limited. Creating a need for a more systematic and controlled way to conduct and use evaluations (Dahler-Larsen 2012:154). Illustrating the fact that the society is not just influenced by evaluations, but also influencing the field of evaluation.

4.2.2 The intention

Considering the number of actors and organizations involved in PISA, numerous intentions are likely to be at work. The discussion concerning intentions could start with the OECD. The OECD's purpose is to use "its wealth of information on a broad range of topics to help governments foster prosperity and fight poverty through economic growth and financial

stability" (OECD (1), 2016). PISA is based on the idea that economic growth depends on quality educational systems, and that international evaluations can provide information about how well a country's system is performing in comparison with others (Elstad and Sivesind, 2010:27–28, Langfeldt and Birkeland, 2010:92). According to the OECD's own website, their overall mission "…is to promote policies that will improve the economic and social well-being of people around the world" (OECD (2), 2016).

It is hard to say whether PISA fulfils the OECD's mission to promote economic and social well-being. Critics argue the contrary because PISA has led to a regime in which important traits such as creativity, independence, and compassion are valued less by the educational system (Ertesvåg, 2015; Hansen, 2013). Indeed, one could question the OECD's assumption of causality between high scores on the PISA test and "social well-being". Norwegian students, whom PISA has evaluated as "average" in knowledge and skills for more than 15 years, live in a country that has been rated five times as among the top five happiest in the world (Heiliwell, Layard, and Sachs, 2017). However, both the ranking of the world's happiest countries and PISA could be seen as expressions of "…the new international competition between global economies" (Dahler-Larsen, 2012:178), expressing the evaluation society.

As far as Norwegian authorities' intentions for PISA participation are concerned, initially, all the main political parties, except the Norwegian Socialist Left Party (SV), favoured this type of international testing. One of the reasons was a lack of quantitative data, and policy makers were eager to have comparative data on student performance levels over a time span (Kjærnsli, personal communication, September 29, 2016).[3] Most likely, it was not really an option for Norway, as a wealthy OECD member, to refuse to take part in PISA (Elstad, personal communication, August 9, 2016). Therefore, Norway's participation might

[3] Marit Kjærnsli is project manager for *PISA* Norway and an associate professor in the Department of Teacher Education and School Research, University of Oslo.

be described as a form of *imposed use* (Weiss, Murphy-Graham and Birkeland 2005). Illustrating the power of the norms involved in evaluation.

Participation in PISA did yield information about the national student performance levels that did not exist before. The first surprising results and the feedback loop that followed have led to increased political and public interest not only in testing but also in the overall quality of the national school system. This has spurred educational research. Internationally, there are mixed opinions when it comes to PISA's impact on research. While Gustafsson (2008) believes, the quantitative material from tests like PISA's have potential for improving the quality of international educational research, Rey (2010) notes that PISA has changed the focus in educational research from values and processes to efficiency (p.144). Either ways, PISA not only influences the educational system of today, it also has a constitutive effect on our perspectives when looking for new knowledge and trying to understand fundamental aspects like learning and education. Leading to second-order construction of our understanding of the educational system.

4.2.3 The time period

When we consider the time dimension in the analytical framework, the first PISA results initiated a process of influence that is ongoing. All though it is hard to separate the influence from each different test, there are both end of cycle effects and immediate effects involved (Kirkhart 2000:15). PISA has an immediate process effect on the students taking part in the test, and on the different parts of the educational system organizing the evaluation. When it comes to the actual test situation, the first students taking part in PISA answered the questions without knowing the fuss their results would lead to, whereas the complex PISA history has probably influenced the later participants. For the organizers, the process of testing is most likely both time and energy consuming, not just because of the practical workload, but also because of the debate that is always present when the testing is taking place.

End-of-cycle influence occurs at the conclusion of a summative evaluation like PISA (Kirkhart, 2000:16). However, the first PISA evaluation was also the start of what Dahler-Larsen (2012:1) refers to as a *feedback loop* in the Norwegian educational system. The first test led to new PISA evaluations and a general system of testing, and thus, the evaluation process itself clearly influenced the educational system. This continuous process of testing, symptomatic of the evaluation society (Dahler-Larsen, 2012), has also lead to an institutionalization of PISA, through new administrative and academic organizations and positions, including a unit at the Faculty of Educational Sciences at the University of Oslo and one in the Norwegian Directorate for Education and Training (Toft, 2008).

In the PISA case, it is complicated to separate the end-of-cycle from the long-term influence. Because of PISA, Norwegian students are learning new things in new ways every single day, and therefore it is hard to define the time limit of the evaluation influence. However, there are clear examples of end-of-cycle influence, for instance in the massive media coverage and public debate following the publication of each new PISA-report. In this heated environment, the political or symbolic use of the PISA results also becomes evident. The political right are generally supportive of PISA and the new test-based educational system and therefore tend to take every credit for improved test results. While the political left tend to highlight all forms of negative findings associated with the PISA-evaluations. For instance if the academic test results are improving, the political left will focus on data revealing more loneliness or social exclusion among students (see for instance Skårderud 2016).

While the end-of-cycle influence following the PISA-evaluation presumably is most evident on the collective level, the national tests have an immediate impact on the individual level because the students get to know their own score, and because these scores have consequences for their teachers and the schools involved (Marsdal 2011).

The time dimension and its sub-categories do allow us to look at influence as a phenomenon working across levels of analysis and time (Dahler-Larsen 2012), and to acknowledge that the degree of influence can change over time and continue for years to come. This indefinite time dimension is another indicator of the evaluation society, illustrating how it is almost impossible not to be affected by the evaluation wave.

4.3 The influence on the individual, interpersonal and collective levels

Henry and Mark (2003) have studied how evaluation influences attitudes and action: the pathways of change at the individual and societal levels. They claim that changes at the collective level are likely to follow from changes at the individual or interpersonal levels (p. 305). Influence at the collective level may occur not only through direct, instrumental use of evaluation and recommendations in an evaluation report, but also through a more indirect diffusion of knowledge and an overall change in attitudes (Henry and Mark, 2003: 305).

According to Elstad and Sivesind (2010), not just the results, but the entire PISA process, has led to changes in the public debate about education, the way Norwegians think about educational quality, and public perceptions of the national school system as a whole. This indirect impact could also be described as a form of imposed use (Weiss, Murphy-Graham and Birkeland 2005), and an example of how evaluation is a phenomenon impossible to escape (Dahler-Larsen 2012). Alternatively, it could be seen as a gradual process of enlightenment (Weiss, 1979), leading to new perspectives when it comes to schooling and the educational system. Either ways, this massive collective impact means that a stakeholder map of individuals affected by PISA could include more or less the whole population. Although the influence is far reaching, and hard to separate into neat categories, the levels provided by Henry and Mark (2003) still contribute to illuminating this complexity of evaluation influence.

4.3.1 Influence on the individual level

PISA affects individuals in the very process of evaluating, through the various use of the results and because of the consequences of this implementation. One of the consequences of PISA is the new national testing of younger pupils; prompting concerns among parents and professionals about the stress, this testing may inflict (Bromark, 2014; Ertesvåg, 2016; Marsdal, 2011; Ubøe, 2015). Reports have also cited incidents in which school officials requested that low-performing students stay home on the day of testing so that the school might score higher in the national rankings (Johansen, 2014; Marsdal, 2011; Sarwar, 2013). Both the stress symptoms among youngsters and the fact that school officials ask students to skip the test could be characterized as unintended negative process influences following from the PISA testing. This way, PISA not only has a direct influence through its own testing, but an indirect influence through a new test system that itself is a consequence of PISA. The fact that some kids find the testing stressful, while others accept it without any trouble also illustrates how the influence from testing varies on the individual level.

PISA has also brought about changes in curricula and in teaching. Any educational system is based on an unknown choice among alternatives, and during the last decades, these choices have been strongly influenced by large-scale testing (Jakhelln & Welstad, 2011: 253-254). Critics of testing maintain that teachers prioritize PISA subjects such as mathematics, natural sciences and Norwegian, and turn their attention away from the national school system's values and ideals (Jakhelln &Welstad, 2011:253-254). Another aspect of this discussion is the rehearsals for national testing. For instance, in Oslo, school leaders and teachers are rewarded for results in the national testing. As a consequence, , some focus less on subjects that are not a part of the testing, such as social science and religion (Marsdal 2011). These changes could be viewed as positive or negative, but either way they contribute to altering the knowledge, values and sense of achievement that each individual brings with them throughout their lives. This way, PISA and the new system of testing also have influence

on the collective level. The testing most likely also leads to a completely new idea about what it is like to be a student. Tests have become the new reality and the way schooling works, and this has impact on the individual, interpersonal and collective level.

Even though attitudes to PISA clearly vary among politicians, civil servants, researchers and journalists, PISA has also resulted in what could be described as unintended positive consequences on the individual level for some professionals. The general boost in educational interest has led to a lift in educational research, new career opportunities for bureaucrats and new attention and resources for politicians, journalists and others engaged in this specific field. However, there are also reports of frustrated teachers claiming that their workplace is changing from a school to a factory, where their main task is to ensure that their students achieve high scores in the testing (Marsdal, 2011).

4.3.2 Influence on the interpersonal level

There are potentially a number of relationships to consider when it comes to tracking PISA's influence on the interpersonal level. One important category is the dyads between students and their peers, teachers and parents. Another central category is the relationships formed by parents and teachers and between parents. A third category to consider is the communication between and among stakeholders representing political parties, organizations and the media.

PISA is a regular event, with a certain history, and every new PISA evaluation is influenced by previous results and debates. When parents communicate with their children, they might tell them to do their very best because PISA is an important international evaluation, or they might say that PISA is a meaningless test producing random numbers for international comparison and continuous economic growth. The student's presumably negative attitudes towards the testing and their lack of motivation have been used to explain the Norwegian results in the media. However, Hopfenbeck and Kjærnsli (2016) found that the Norwegian students where in fact motivated to do their best in the PISA-testing. Indicating

that lack of motivation cannot be used to explain the Norwegian results. Still, Eklöf et. al (2014) revealed a statistically significant relationship between motivation and test performance among Norwegian students taking part in TIMMS Advanced[4]. In this study, half the Norwegian students claimed that they did not give the test their full attention. These diverse findings could of course result from the fact that the studies report from two different tests (TIMMS and PISA), but they could also reveal methodological challenges involved when measuring self-reported motivation.

As previously discussed, the new system of testing has been criticized for creating stress among children and for changing the national curricula. Parents naturally react when their children are unhappy or worried. They might also take action if they believe the school is downgrading traditional values and subjects they regard as important. Further, the PISA results themselves could make parentsworry about the general quality of the educational system. These aspects could lead to tensions between teachers and parents, or between parents and school leaders. However, while resourceful parents might feel quite comfortable discussing such issues with the school, these aspects might be hard to address for less advantaged parents. This could reinforce the gap between the educational system and groups of students and their families. Contributing to further socio economic diversity in society.

According to a survey among teachers and school leaders,[5] 86 per cent of the respondents agreed that parents' perceptions of the school system were negatively affected by the public PISA debate (Midtbø & Stavik, 2009:23). Most likely, this shift complicates the day-to-day work for employees in the educational system, and according to the theories of influence, this could be described as a negative and unintended result of PISA.

The Norwegian PISA participation is financed by the Ministry of Education and organized by the Department of Teacher Education and School Research at the University of Oslo

[4] A large-scale, international comparative study measuring advanced mathematics and physics.
[5] The survey was organized by Union of Education Norway, the nation's largest teachers' union.

(Helsvig, n.d.). Because PISA has become immensely important, it has most likely led to a stronger relationship between researchers and bureaucrats - and maybe politicians. While close relationships between evaluators and clients are seen as a factor increasing the use of evaluation results (Nutley, Walter and Davies, 2007:68), such bonds could have a negative impact on the potential for criticism. One could argue that an academic institute organizing international tests is not in a position to criticize suchtesting in research.

4.3.3 Influence on the collective level

Henry and Mark (2003:304) cited policy change as a possible result of the influence from evaluations at the collective level. In Norway, PISA has created a new political focus on student performance, a new phenomenon in the history of the Norwegian educational system (Elstad, 2010:101; Norwegian Directorate for Education and Training, 2011). This has led to school reforms with a significant collective impact. While bureaucrats and school leaders tend to find the information from the tests useful. The teachers are mainly negative to the value of the results presented and the experience of teaching for testing (Marsdal 2011:43). Illustrating how the influence from evaluations can have different impact on different groups at the collective level. Even though the teachers as a group of professionals could be described as sceptic, each teacher has their own individual experience of the testing.. For instance a teacher reports how she can be excited and engaged if her class is doing well, somehow enjoying a system she does not support (Marsdal 2011:40). Illustrating how the interpersonal relations between teachers and students also areaffected by the testing.

On the collective level, PISA's *constitutive effects* lead to a general adoption of standards, indicators and definitions used in PISA, and PISA becomes a standard for measuring educational quality and how to define a successful school system (Elstad and Sivesind, 2010; Hanberger, 2014). Dahler-Larsen claims that constitutive effects occur more strongly when the data from evaluations are published (2012: 213-214). An element in the

discussion of PISA's role in shaping the educational discourse has clearly been the massive media interest. According to Langfeldt and Birkeland (2010), PISA offers a form of comparative international information that is very attractive to journalists. For instance, has there been an enormous interest for comparing the Norwegian results with the scores in neighbouring Scandinavian countries. However, in the massive media focus on numbers and rankings the nuances have a tendency to disappear. This is a reason why both teachers and school leaders find the public debate problematic. One example given is the media`s use of the PISA results as an indicator of the total quality of the Norwegian school system – when the evaluation is in fact just measuring a limited outcome of the educational activity (Eggen, 2010:287).

Because the media has played an important role in the PISA process, one might argue that media could be a dimension of its own in the influence framework. On the other side, most evaluations do not get this sort of attention, and this could be an argument for discussing the role of media within the existing framework.

5. Discussion

According to Dahler-Larsen (2012), we do not know enough about the societal consequences of evaluation. Using an analytical framework like the one presented in the theories of influence makes it clear that the influence from evaluations might be far more complex and considerable than what is captured through the traditional one-dimensional concepts of use. When analysing the complexity of evaluation influence, the evaluation society becomes evident, revealing how the values and norms of society also leave their marks on evaluation. This contributes to explaining why we evaluate, how we perform evaluations, and how the knowledge from evaluations is used and what it represents.

Dahler-Larsen (2012) claims that the evaluation wave, started as a reflection of the rationalist belief in societal progress through the use of reason and science, characterizing

modernity. This logic is still central in today's evidence-based policy and the public sector´s demand for evaluations. A political or bureaucratic interest in educational statistics seems to be an important motivation behind the Norwegian decision to join PISA. However, if the Norwegian authorities' intension were just to monitor their students' results for knowledge-informed policy, they might get a better information by using a national longitudinal study than joining PISA (Braeken, 2014). The fact that PISA does not build on national curricula could also be used as an argument for a national solution.

Instead of a search for quantitative information describing student results, taking part in PISA could be seen as an attempt to express Norwegian seriousness about research and progress through knowledge-driven policymaking, international cooperation and transparency. Reflecting the use of evaluation as a *symbol* of good management and knowledge-based decision-making, typical for reflexive modernity (Dahler-Larsen, 2012:145-147). Finally, taking part in PISA could also be understood as the authorities´ attempt to create and *illusion* of control by quantification and objective standards, presented as the central aspects of evaluating in the audit society (Dahler-Larsen 2012:169-173). Illustrating the complexity involved when trying to reveal the Norwegian intension behind PISA participation. This complexity is exactly the reason why it is important to analyse the evaluation process in order to understand the role of evaluation in society.

In the theories of influence, **time** is highlighted as a central factor. Norwegian PISA participation gives an opportunity to follow evaluation over more than 15 years. Making it possible to take both immediate, end-of-cycle and long-term effects of the evaluation into consideration. When looking at the international PISA ranking, there are only minor changes to be found in the Norwegian achievements, but PISA's influence on the Norwegian society has been considerable. The PISA **results** have been used instrumentally to implement school reforms and a new test-based educational system, while both the **results** and the **process** of

testing has caused a major shift in the public discourse concerning schooling and educational quality.

PISA has revealed that even in a wealthy, egalitarian country like Norway, socioeconomic background is the most important factor in school achievement (Haug, 2010: 255). According to Andreas Schleicher, OECD Director for Education and Skills, the Norwegian school system fails to lift low achievers (Helljesen, Grønli and Omvik, 2013). However, the importance of socioeconomic factors could be an argument for actions outside the educational system, such as better support for low-income families, or more focus on early pedagogical interventions, rather than a test-oriented educational system. In fact, it could be reasonable to assume that testing might have a positive effect on students who are already comfortable in the educational system, but it is counterproductive when it comes to students who are struggling. The school reforms could therefore be seen as an illusion of adequate action, a symbolic use of evaluation and as an expression of the evaluation society, rather than as a realistic attempt to lift low-performing students.

By participating in PISA, a nation signals not only its dedication to education, but also its belief in the values and policy approved by the OECD. For the PISA countries, corrective action is the answer to poor results; however, within the PISA framing, the number of solutions is limited. The answers are supposed to be found within the educational system (OECD (3), 2016) and not in social reforms. It seems somewhat complicated to define whether this framing of acceptable political solutions should be categorized as an intended or unintended consequence of PISA, but it could clearly be described as a constitutive effect of the OECD-initiated testing. Meaning that the evaluation itself defines acceptable response. According to Belgian philosopher, Isabelle Stengers (1999), there are no stronger arguments in a modern democracy than the ones built on research and science. The people may have the

power to decide, but science and research provide the alternatives from which to choose, and politicians promote these alternatives (p.5).

Like an evaluation machine, PISA is administered according to scientific standards and managed by an international organization, and this probably curbs critics at a national level. However, knowledge is never neutral and objective in the evaluation society and the OECD has been criticized for hiding political interests behind the PISA evaluation (Haugen, 2010:69). Although reports by the Norwegian PISA research team have reflected a clear division between empirical data and any possible explanations and recommendations (Langfeldt and Birkeland, 2010:34), the OECD's policy briefs based on the same results have been criticized for being political documents rather than evidence based recommendations (Elstad and Sivesind, 2010:80). OECD's recommendations have included statements like "close down small schools", "give teachers performance-based wages", and "do not increase the schools' funding" (Elstad and Sivesind, 2010: 34). This clearly indicates that the PISA example is not a story about possible misuse of data, but the significant meaning of values and norms in the evaluation society.

The fact that PISA involves both national authorities and the international OECD is an example of the complexity involved in any clear outline of evaluation **intentions** and the consequences of their **results**. On an international level, the same PISA results were used to argue for opposite solutions in Denmark and Norway: a centralized educational system in Denmark and a more decentralized one in Norway (Hernes, 2008:264). The rankings have also been used to legitimize existing educational systems in countries that did well in PISA, and to argue for reforms in those with lower scores (Bachmann, Haug and Myklebust, 2010: 298). Illustrating the need for analysing PISA both as a complex process of evaluation and as a social phenomenon reflecting national values and ideas.

PISA is an example of the power embedded in evaluations when it comes to influencing a political field and the public discourse (Hanberger, 2014). The reforms, the system of testing and the new educational discourse are influencing the Norwegian society on the **individual**, **interpersonal** and **collective** levels. In the evaluation society, evaluation is the norm. In Norway, the test-based school system has become the norm for new kids starting school and for teachers in the beginning of their careers. Testing is the way to ensure the quality of the educational system and to measure the academic potential in every student.. Evaluation has become a part of our mentality and nothing individuals can reject or oppose – we are in the loop (Dahler-Larsen 2012:3). The feedback loop may also be in effect at the collective level. Even though critical voices have been raised in the PISA debate, the major political parties have been loyal to the testing. The transparency that allows for international monitoring and comparison of results helps maintain the loop. If Norway were to drop PISA, the impression could be conveyed that the country cannot solve the challenges in its educational system and therefore does not want to draw attention to student achievement. In fact, when the Norwegian Centre Party in 2016 declared its intention to withdraw the nation from PISA their statement was immediately criticized by the Norwegian Conservative Party as an attempt to hide the mediocre national results (Sandvik, Grønli, and Myklebust, 2016). The ultimate consequence of this kind of lock-in effect (Osteloh and Frey, 2010 in Dahler-Larsen 2012:204) is that PISA ends up as the only legitimate way of evaluating educational systems.

6. Conclusion

The Norwegian PISA experience illustrates the need for a broad understanding of the influence of evaluation. Evaluation is a method for observing and valuing activities that otherwise may not be subject to analysis and therefore a potential source of useful knowledge.

However, evaluation also makes sense beyond utility – expressing the values and norms of society. Both aspects are important to remember when trying to present a complete picture of the use of evaluation in modern societies. When recognizing all aspects of evaluation, we are more likely to understand how evaluation can or cannot contribute to what the theories of influence call *social betterment*.

References

Albæk E (1988) *Fra sandhed til information : evalueringsforskning i USA før og nu,* København, Akademisk Forlag.

Alkin MC and Taut SM (2003) Unbundling evaluation use. *Studies in Educational Evaluation* 29.

Bachmann K, Haug P and Myklebust R (2010) Med rett til å prestere" i Eyvind Elstad. *Kirsten Sivesind (red.):" PISA–sannheten om skolen*.

Baird J, Isaacs T, Johnson S, Stobart G, Yu G, Sprague T and Daugherty R (2011) Policy effects of PISA.

Braeken J (2014) Et nyansert syn på PISA. Faculty of Educational Sciences, University of Oslo, Available at https://issuu.com/shanecol/docs/forsknings-bro-web?e=8763060/9855158 (accessed 6 May 2017).

Bromark M (2014) *Voksne ville blitt stresset av ungdoms skolehverdag* [Online]. Klikk.no. Available at http://www.klikk.no/foreldre/skolebarn/ungdom-og-stress-1488938.ece (accessed 9 November 2016).

Carden F and Alkin MC (2012) Evaluation roots: An international perspective. *Journal of MultiDisciplinary Evaluation* 8(17): 102–118.

Dahler-Larsen P (2012) *The Evaluation Society,* Stanford, CA: Stanford University Press.

Department of Teacher Education and School Research (ILS) , F. O. E. S., University of Oslo. Available: http://www.uv.uio.no/ils/forskning/prosjekt-sider/pisa/resultater/ (accessed 9 November 2016).

Eggen A (2010) PISAs gyldighet blant skoleledere og lærere. I E. Elstad and K. Sivesind (Red.). *PISA: sannheten om skolen*.

Eklöf H, Pavešič BJ and Grønmo LS (2014) A cross-national comparison of reported effort and mathematics performance in TIMSS advanced. *Applied Measurement in Education* 27, 31–45.

Elstad E (2010) PISA i norsk offentlighet: politisk teknologi for styring og bebreidelsesmanøvrering. *PISA: Sannheten om skolen*, 100–121.

Elstad E and Sivesind K (2010) *PISA: sannheten om skolen?*, Universitetsforl.

Ertesvåg (2015) Lærere fikk nok av prøvepress - sluttet i Oslo-skolen. *VG*, 24. September 2015.

Ertesvåg F (2016) Barneprofessor bekymret for skolestress: Stadig flere søker hjelp. *VG*, 11 October 2016.

Gustafsson JE (2008) Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal* 7(1): 1–17.

Hanberger A (2014) What PISA intends to and can possibly achieve: a critical programme theory analysis. *European Educational Research Journal*, *13*(2), 167-180.

Hansen A (2013) *Det skjeve tårnet: PISA* [Online]. Retriewed from http://www.forumnyskole.org/det-skjeve-tarnet-pisa/ (accessed 9 November 2016).

Haug BE (2010) Desentralisering og skoleprestasjoner. *PISA: Sannheten om skolen*, 246–255.

Haugen C (2010) For en konservativ modernisering av utdanning? *PISA: Sannheten om skolen*, 69–80.

Heiliwell, Layard, and Sachs (2017). *World happiness report 2016 update* (Vol. 1).NewYork: Sustainable Development Solutions Network.

Helsvig AK En linje i det norske Kunnskaps-departementets historie 1945–2015. Retrieved from http://www.uddannelseshistorie.dk/images/pdfer/2015-kim-helsvig-moderniseringsagent.pdf

Helljesen V, Grønli H and Omvik OR (2013) PISA-sjefen mener norsk skole har et stort, skjult problem. [PISA boss believes Norwegian school has a large, hidden problem]. NRK. Retrieved from https://www.nrk.no/valg2013/pisa-sjef_-_norges-store-problem-1.11186771

Henry GT and Mark MM (2003) Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation* 24: 293–314.

Herbert JL (2014) Researching evaluation influence: A review of the literature. *Evaluation Review* 38: 388–419.

Hernes G (2008) The interface between social research and policy making. *European Sociological Review* 24: 257–265.

Hopfenbeck TN and Kjærnsli M (2016) Students' test motivation in PISA: The case of Norway. *The Curriculum Journal* 27: 406–422.

Jakhelln H and Welstad T (2011) PISA-undersøkelsene – en dreining av verdigrunnlaget i norsk skole uten lovhjemmel? *Lov og Rett,* 50 ER.

Johansen AH (2014) Vi opplever hardt press for å frita elever. *Dagbladet*, 6 December 2014.

Kirkhart KE (2000) Reconceptualizing evaluation use: An integrated theory of influence. *New Directions for Evaluation* 2000: 5–23.

Kjærnsli M and Jensen F (2016) 1 PISA 2015–gjennomføring og noen sentrale resultater. In: *Stø kurs*. 2016. pp. 11–31.

Kulbrandstad LI (2010) Leseopplæring på ungdomstrinnet før og etter PISA 2000" i Eyvind Elstad og Kirsten Sivesind (red.)". *PISA–sannheten om skolen*.

Langfeldt G and Birkeland N (2010) PISA i lys av styringsteori. *PISA: Sannheten om skolen*, 83–97.

Lindgren L (2014) Nya utvärderingsmonstret: om kvalitetsmätning i den offentliga sektorn. Studentlitteratur.

Mark MM and Henry GT (2004) The mechanisms and outcomes of evaluation influence. *Evaluation* 10: 35–57.

Mark, MM and Henry, GT (2013). Multiple routes: Evaluation, assisted sensemaking, and pathways to betterment. *Evaluation Roots: A Wider Perspective of Theorists' Views and Influences*, 144-156.

Marsdal M (2011) Kunnskapsbløffen. Oslo: *Forlaget Manifest.*

Midtbø R and Stavik T (2009) PISA får for stor plass. *Bedre skole* 1-2009: 20-24.

Norwegian Directorate for Education and Training (2011) Internasjonale studer om norsk skole. [International study on Norwegian schooling]. Retrieved from http://www.udir.no/Upload/Rapporter/temanotat/Internasjonale_studier_om_norsk_skole_temanotat.pdf

Nutley SM, Walter I, and Davies HT (2007) *Using Evidence: How Research Can Inform Public Services*. Policy Press.

OECD (1). Available: http://www.oecd.org/about/whatwedoandhow/ (accessed 9 November 2016).

OECD (2). Available: https://www.oecd.org/about/ (accessed 9 November 2016).

OECD (4). Available: http://www.oecd.org/edu/skills-beyond-school/definitionandselectionofcompetenciesdeseco.htm (accessed 9 November 2016).

Patton MQ (1998) Discovering process use. *Evaluation* 4(2): 225–233.

Power M (1997) *The Audit Society: Rituals of Verification.* Oxford: Oxford University Press.

Ubøe J (2015) Et samfunn der alle er tapere. *Bergens Tidende (BT)*, 16 Feb. 2015.

Rey O (2010) The use of external assessments and the impact on education systems. *Beyond Lisbon 2010: Perspectives from Research and Development for Education Policy in Europe*, 137–157.

Sandvik B, Grønli H, and Myklebust S (2016) SP vil melde Norge ut av PISA. *NRK*.

Sarwar S (2013) Forsker: Skoler fritar svake elever for å blåse opp resultatene. *Dagbladet*, 15. Januar, 2013.

Saunders M (2012) The use and usability of evaluation outputs: A social practice approach. *Evaluation*, *18*(4), 421-436.

Skårderud J (2016) Gjør det bedre, har det verre. Klassekampen, 7. des.2016

Stengers I (1999) *For en demokratisering av vitenskapene* (Vol. Nr 4; Virus). Oslo: Spartacus.

Toft M (2008) På krigsstigen mot PISA. *Uniforum*, 5 mars 2008.

Weiss (1979) The many meanings of research utilization. *Public administration review*, *39*(5), 426-431.

Weiss CH, Murphy-Graham E and Birkeland S (2005) An alternate route to policy influence how evaluations affect DARE. *American Journal of Evaluation* 26: 12–30.