The quality of regulatory judgments of health care inspectors

Saskia M. Tuijn

The quality of regulatory judgments of health care inspectors

Published by

Ipskamp Drukkers

Postbus 333

7500 AH Enschede

The Netherlands

Phone: + 31 53 4826262

Fax:+31 53 4826270

e-mail:info@ipskamdrukkers.nl

http://www.ipskampdrukkers.nl

Cover photo: Saskia Tuijn

Cover design: Mark Klik

ISBN: 978-94-6259-145-5

The quality of regulatory judgments of health care inspectors

De kwaliteit van oordelen van inspecteurs in de Gezondheidszorg

(met een samenvatting in het Engels en Nederlands)

Proefschrift

ter verkrijging van de graad van doctor

aan de Universiteit Utrecht

op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan,

ingevolge het besluit van het college voor promoties

in het openbaar te verdedigen op dinsdag 9 september 2014 des middags te 12.45 uur

door

Saskia Maria Tuijn

Geboren op 27 december 1975 te Zaanstad

Promotoren: Prof. dr. H. van den Bergh

Prof. dr. P.B.M. Robben

Prof. dr. F.J.G. Janssens

Table of contents

Beoordelingscommissie: Prof. dr. R.A. Bal

Prof. dr. G.A.M. van den Bos

Prof. dr. F.L. Leeuw

Prof. dr. T.J.M. Sanders

Prof. dr. G. van der Wal

Chapter 1

General introduction and research questions

1.1 The importance of reliable and valid judgments in regulatory decisions on health care

Regulatory decisions on health care have a key position in the regulatory process. Based on the judgments of inspectors, health care institutions are asked to improve the quality of the care they deliver when necessary. If the improvements are not satisfactory, the Dutch Health Care Inspectorate (IGZ) can impose administrative sanctions and initiate penal measures, which may have serious consequences. When regulatory judgments are not reliable, institutions with similar characteristics are judged differently. When this happens, it is hard to explain why some institutions have to improve the quality of their care while others do not, and the regulatory process may appear subjective. When regulatory judgments are not valid, even though inspectors all assign the same judgment to institutions with similar characteristics, this judgment will not reflect the regulatory authority's corporate standards in any of these cases. The judgments can be too positive or too negative compared with the standards developed by the regulatory authority to identify safe, high-quality health care and evaluate the preconditions for this. If judgments are too positive, there is the risk that institutions will not be asked to improve their care, while this actually should have happened according to the corporate standards. For any regulatory authority, reliable and valid judgments are important requirements for preserving authority and being able to achieve improvements in the field they regulate.

1.2 Definition of regulation

The scope of the definition of regulation differs. Some definitions are more narrow, and others are broader [1]. For example, the Oxford English Dictionary defines regulation as "A rule prescribed for the management of some matter, or for the regulation of a conduct; a governing precept or direction; a standing rule." According to this definition, regulation means the introduction of laws or rules accompanied by mechanisms for monitoring and enforcing compliance [1]. Other definitions employ broader viewpoints, for example, the one that defines regulation as the privilege of the state (meaning government and its agencies) and refers to any form of direct intervention by the state, often through a public agency third party [1]. Supervision is included in the definition of regulation. The difference between regulation and supervision is determined by the reach of the activities. Regulation of health care includes all activities and accompanying instruments (like legislation, funding, and supervision) used to oversee and guide health care. Supervision is part of regulation. It includes a system for monitoring the quality and safety of health care, and for taking action when this quality and safety are at risk. On-site inspections are an example of an instrument available to regulators for assessing the quality of care. We

will use the definition of regulation drawn up by Ayres and Braithwaite (1992), which specifies the goal of and actors in regulation [1]. Ayres and Braithwaite define regulation as a function of governance that can be performed by state or non-state actors who use a variety of approaches (ranging from persuasion to coercion) to steer the flow of events. Here we will use both regulation and supervision to refer to all activities undertaken by the IGZ to monitor and improve the quality and safety of health care in the Netherlands.

1.3 Regulation of health care in the Netherlands

While regulation is not necessarily carried out directly by the government, it is usually undertaken with government support or authority, and regulators generally have a mission to protect, promote, or support the public interest [2]. In the Netherlands, government regulation of health care is performed by the IGZ, which is an independent agency within the Ministry of Health, Welfare, and Sport. The IGZ was set up in 1865, when the politician Thorbecke designed the Health Act [3]. The IGZ enforces the quality of care in the public interest. The IGZ safeguards the quality of care based on the Public Health Act[1], and enforces more than 25 laws, including the Care Institutions Quality Act [4]. The IGZ regulates the quality of care in a wide variety of health services, including hospital and nursing home care, mental health care, public health care, care for those with disabilities, and care provided along with pharmaceutical products and medical aids. The primary responsibility for the quality of care lies with the care provider, and this is the starting point of IGZ regulation [4,5]. To stimulate the quality of care, IGZ policy aims to standardize procedures and promote reliable and valid judgments. The IGZ also strives to justify its regulatory decisions and activities [3,6-10].

Regulators need methods for measuring and monitoring the performance of the organizations they regulate, a process described as “detection” [11]. For this purpose, the IGZ uses a combination of methods [4]. During the years this research was performed, the IGZ used three methods for its regulatory task: theme-based regulation, regulation in response to incidents, and risk-based supervision.

Theme-based regulation focuses on specific issues in health care. Sometimes the issues that require the attention of the regulator are put forward by a government minister or by parliament. The IGZ employs regulation in response to incidents in the event of emergencies that indicate structural shortcomings in health care. The IGZ has employed risk-based supervision to assess the quality of health care by means of indicators since 2002 [4,5]. As in countries such as Australia, the United States, Switzerland, Sweden, and Norway, quality indicators were introduced in the Netherlands to monitor and stimulate the quality of health care [12-16].

[1] Gezondheidswet

In risk-based supervision, a framework for the quality of care and also accompanying sets of quality indicators are drawn up, in cooperation with representatives from the health care sector. Subsequently, risk-based supervision consists of three phases: First, the IGZ analyzes the data collected with the indicators and selects institutions at risk. Next, inspectors visit the selected institutions. Institutions are obliged to improve their care based on the inspectors' judgment. Inspectors can decide to plan a follow-up visit if these improvements are not satisfactory. Finally, if the improvements are not satisfactory, the IGZ can impose administrative sanctions and initiate penal measures.

With risk-based supervision, inspectors visit a selection of health care institutions. This selection consists mainly of institutions at risk. Sometimes, this selection is made up of institutions that do not seem to be at risk based on their scores on the indicators. Because the institutions visited do not vary widely with respect to these risk scores, this selection of institutions at risk is a complicating factor, and implies that the inspectors visit and examine institutions that make up only a small part of the whole. As a result, it is necessary to make very accurate measurements to be able to reveal small differences between these institutions. This implies that both the regulatory instruments and the inspectors will need to comply with strict requirements.

In the Netherlands, there are those who advocate for regulating health care according to the principles of professional regulation. One of these principles is defined as standard procedures and transparency [6,7,9,17-19]: "A good regulatory authority explains the need for regulation, enlightens choices in regulation, and aims for uniform working processes. The regulatory authority publishes its findings as much as possible. Afterwards, the transparent regulatory authority justifies itself for its choices and regulatory results." The use of criteria or standards to assess the performance of the regulatee can facilitate a fair and transparent regulatory process [20]. Moreover, several juridical principles have been defined for regulating government conduct towards individuals in the Netherlands. These principles also apply to the regulation of health care. One of the principles is the principle of equal rights, and is defined by the Constitution of the Netherlands: the government treats similar cases equally. Other principles are defined by the General Administrative Law Act for example:

- The principle of accurateness: The government has to prepare and make a decision accurately. This includes proper treatment of the individual, an accurate investigation of the facts and interests, employing procedures and reliable decision making (article 3:2 and 3:4 subsection 1).
- The principle of argumentation: The government has to make a case for its decisions. The facts have to be correct, and the argumentation has to be logical and comprehensible (article 3:46).

- The principle of legal security: The government has to formulate its decisions comprehensibly, and apply the legal rules both correctly and consistently (article 3:40-3:45).

These principles apply to the IGZ, and IGZ employees are obliged to act according to these principles. They are particularly relevant in relation to reliable and valid decision making. Reliability refers to the consistency of ratings or to the ability of various raters to reach the same conclusion about a specific case. When a number of observers rate similar cases equally, the ratings are consistent. This is represented by high reliability. However, high reliability does not necessarily imply high validity. We will elucidate the difference.

Most regulators use (written) standards to communicate their regulatory objectives [11]. For the IGZ, these standards define the requirements for safe, high-quality health care. Moreover, criteria are included in the standards that describe which subjects are examined in on-site visits and which judgments apply in which situations. These are defined as corporate standards. When inspectors apply the corporate standards and assign judgments completely according to the corporate standards, the judgments are defined as corporate judgments. However, when raters all assign the same judgment to a similar case and this judgment does not completely correspond with the corporate standards, the judgments are reliable but not valid. Validity refers to accuracy, or to the extent to which a rater's measurement approximates the true value [21,22]. In the regulation of health care, this true value is represented by the corporate standards. Regulatory decisions have to be accurate, transparent, and have sufficient grounds, and similar cases have to be treated in a similar way.

1.4 Awareness of interrater reliability and the role it plays

The phenomenon of interrater disagreement has existed for a long time in a wide variety of professions. Research on judgment and decision making has significantly influenced research in several applied fields, including education [23-25], medicine [26,27], psychology [28], medical insurance science [29,30], law [31,32], public policy, and business (such as accounting and auditing) [33].

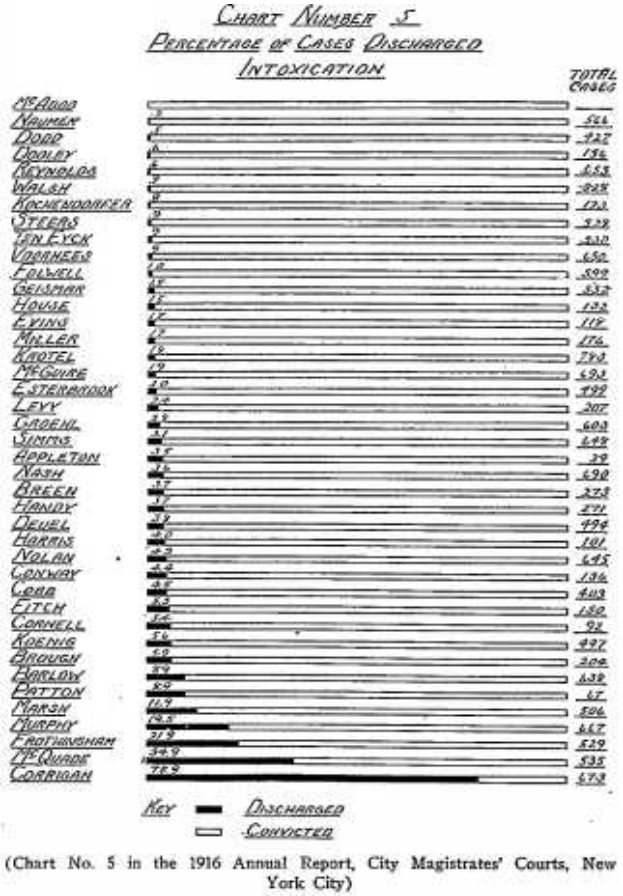
In medicine, Kilpatrick referred to the work of Sir Thomas Browne on the concept of human error, which dates back to 1646 [34,35]. Browne recognized several sources of error: the common infirmity of human nature, the erroneous disposition of the people, misapprehension, fallacy or false deduction, credulity, obstinate adherence to authority, the belief in popular conceits, and even the endeavors of Satan. In the early 1930s and late 1950s, many studies were carried out on interrater reliability in medicine [36,37]. The topic of reliability in medical practice is still relevant nowadays, and discussions about variation in medical interventions tend to flare up [38]. The focus of these discussions varies. For example, they may focus on the relationship between reliabil-

ity and the quality of care [39], between reliability and the costs of health care [40,41], or on the possibility of using interrater reliability to obtain a second opinion [42].

There has also been research on interrater disagreement in education. As early as 1888, Edgeworth stated that examination is a very rough, yet not wholly inefficient, test of merit, something that is generally accepted [43]. Empirical research on interrater reliability in education has been performed since the beginning of the nineteenth century [44,45]. These studies focused on grading work in English and mathematical papers. Studies conducted on education since the 1970s have investigated the complexity of decision making [24,25,46-48]. Extensive research has also been conducted on education in the Dutch language. This research focused on the reliability and validity of instruments used for student examinations [49]. Research has shown that, when evaluating their students' texts, the interrater reliability of teachers of the Dutch language is not always optimal [50,51]. Interrater reliability still plays a significant role in education [24,52]. An illustration of problems with the reliability of examinations is the 2012 national German examination, when it became clear that the first and second examiner did not agree. It appeared that the first examiner adjusted the work of his pupils, and made some corrections [53].

The concept of observer error is also apparent in penal regulation, and was discussed by Everson in the judiciary in 1919 [54]. Everson explained: "If one went about the courts from day to day, he would note a variation among the different magistrates. He would notice that one magistrate was particularly severe with some class of offenders, while not so severe with another. Another would be lenient with nearly all. While yet another would be uniformly severe, except in cases of some particular class of offenses." The reliability of magistrates was not only discussed but also examined in the early twentieth century. Everson exemplified the reliability of magistrates by the outcomes in the 1916 Annual Report of the City Magistrates' Courts in New York City. This report contained an analysis of the data for the cases in the City Magistrates' Courts. Figure 1 shows there was complete transparency in those days: the last names of the all of the magistrates were given. As illustrated in Figure 1, a great deal of disparity was found in the sentencing practices of different magistrates. The percentage of convictions of Magistrate Dooley is much higher compared with that of Magistrate Corrigan. Although these percentages all concern cases of intoxication, it is not explained whether the variance between magistrates can be explained by characteristics of the magistrates alone. For example, Dooley possibly had all of the seasoned delinquents because offenses were more serious in his part of the country.

A second aspect that remains unclear is the severity of the convictions. If we knew the true seriousness of the offenses, it would be possible to measure the differences between the magistrates' convictions and the



(Chart No. 5 in the 1916 Annual Report, City Magistrates' Courts, New York City)

Figure 1 Outcomes of the 1916 study on disparity in sentencing among magistrates in New York City, presented by Everson in 1919.

ground truth (true value). Subsequently, it would be possible to determine whether the standard deviation of the difference was comparable among magistrates. In other words, the systematic variance could be calculated and it would be possible to determine whether Dooley's variance is smaller compared with that of Corrigan. Although the percentage of convictions in Dooley's case is relatively high, his systematic variance might be smaller. Therefore, compared with other magistrates, his judgments might be more consistent.

Another aspect that remains unclear is the accurateness of the convictions. Because the true value is unknown for these cases, it is not possible to determine the validity of the convictions. Validity is always stated with reference to the criteria used [55]. Because these criteria are not known and we do not know whether the magistrates provided grounds for their convictions, it is not possible to determine whether the magistrates' convictions are precise and conform to the standards. Moreover, no explanation was given of whether the magis-

trates used instruments like guidelines or written criteria in their decision-making process, or about the role these instruments played with regard to the reliability of the magistrates.

Nor was it explained whether the magistrates were aware that their convictions would be used in an evaluation process. If they were aware of this, the Hawthorne effect could have affected the results. The Hawthorne effect is a form of reactivity whereby subjects improve or modify an aspect of their behavior that is being experimentally measured simply in response to the fact that they are being studied, and not in response to any particular experimental manipulation [56].

The effect of the convictions also remains unclear. Figure 1 does not explicate whether magistrates whose convictions were more severe had less repeat offending compared with magistrates whose convictions were more lenient. Lastly, it was not reported whether attempts were made to increase the reliability and validity of the magistrates' convictions.

Continuing on from the study by Everson, various series of studies were performed on the interrater reliability of judges [57-61]. These studies also showed differences in sentencing decisions. Attention has also been paid to the importance of the argumentations for convictions [62]. The discussion about sentencing decisions has recently flared up in the Netherlands as well [63]. For example, the Dutch penal code does not instruct judges how to determine penalties. The law books contain only the maximum penalties for every offense, and cases of concurrence. However, the discussion about the introduction of minimum penalties seems to have kindled a discussion about the professional discretion of judges in the Netherlands [64]. The council for the Judiciary fears there will be unexplainable differences in the penalties between offenses that have minimum penalties and those that do not.

1.5 The importance of insight into interrater reliability and validity of judgments in the regulation of health care

Despite the interest in explaining and increasing interrater reliability, there is still comparatively limited knowledge on interrater reliability and validity in regulation. Over the past number of years, the need for transparency in governmental decisions as well as for accountability in regulation has increased [65-69]. Consequently, the importance of reliable and valid judgments in the regulation of health care has also increased. It is less complicated to account for regulatory decisions if judgments are both reliable and valid. Moreover, earlier research has shown that, for institutions, the legitimacy of judgments is an important precondition for maintaining authority [70]. Therefore, it seems fair to assume that reliable and valid judgments are more effective. Further-

more, there is a growing need for transparency with regard to governmental decisions [66]. Reliable and valid judgments are an important precondition for being able to account for regulatory decisions and for preserving authority.

There is an international trend towards a greater use of government regulation in health care [65]. Just like accreditation, examining the reliability of judgments within the area of health care regulation is important, given the costs and prevalence of regulation [71-75]. Nevertheless, empirical research on interrater reliability and validity in regulation has not yet been developed. Most of the research has focused on risk regulation regimes [76] and surveyor styles, which offers important knowledge for increasing insight into the mechanisms of regulation [75,77-79]. Insight into the reliability and validity of regulatory judgments is a valuable addition because it offers the possibility of further professionalizing the regulation of health care. In the next section, these considerations will be expanded upon and placed in a theoretical model.

1.6 Theoretical framework and research questions

The reliability of a judgment depends on a variety of factors: the definition of the criteria being evaluated, the objects or persons being judged, the method used for making judgments, the setting, when the judgment is made, and the observers themselves [21,80]. The combined effect of these factors on an evaluative score is referred to as error of measurement [55].

If we focus on these various factors, we can identify some underlying factors that can also influence variability. The method used for making judgments mentioned above is a broad category, and includes instruments used for decision making. Most regulatory authorities employ standards for communicating their expectations to other stakeholders in regulation. These instruments are usually written statements used to explicate the regulatory objectives [11]. The regulatory instruments fall under the "method used for judging." When these instruments are strictly applied, inspectors use them during regulatory visits to assign and provide grounds for their judgments. They describe which judgment applies in which situation. This results in corporate judgments: judgments that correspond with the corporate standards. When the judgments do not correspond with the corporate standards, a validity problem arises.

Moreover, the content of the instrument also largely determines the validity of the judgments. The items or criteria described in the instrument are intended to represent risk in health care. However, when the criteria are not representative for risk in health care, the validity of the instrument is not yet optimal. For example, when the criterion "pressure ulcers" is not included in the regulatory standard even though this is an important criterion

for quality care in nursing homes, the standard cannot be considered to be representative. When developing regulatory standards, it is important that the standards describe the entire spectrum of the quality of care. Moreover, inspectors have to be able to use the standards to distinguish between institutions. This implies that the categories used in the instrument to assign scores to criteria have to largely describe actual situations in care. If the instrument has been well developed but is very impractical to use (for example, because the categories of scores do not correspond to the situations encountered during on-site visits), inspectors might not use the instrument during regulatory visits, or will only use it to a certain extent. The quality of the regulatory instrument is therefore important. Also, continuous evaluation and, when necessary, improvement of standards or instruments appear to be conditions essential to reliability and validity.

If we look more closely at the category "observers" (inspectors), we can distinguish factors related to the person, like previous job, age, gender, or other inspector characteristics. However, manipulating these factors to stimulate corporate judgments and reliability and validity does not seem very rational. Other factors, like training on becoming an inspector, on using regulatory instruments, on assigning scores to health care institutions, and on interventions to increase the reliability and validity of regulatory judgments can also be considered to be inspector characteristics. Because we did not yet know what kinds of interventions or training will be effective for stimulating reliability and validity, we defined these factors as training. The regulatory instrument and training of the observers are factors that were manipulated in this study, and as a result might have influenced reliability and validity. We study the effect of the type of instrument on reliability and validity. We subsequently explore how other professionals increase reliability. We study interventions performed to increase reliability and the effect these interventions have. Next, we investigate whether interventions that are effective for other professionals can also be effective for health care inspectors. In the scientific literature on reliability, the main approach to increasing reliability seems to involve increasing the number of observers and improving the instrument used [55].

In this study we investigate the reliability and validity of regulatory judgments within the system of risk-based supervision and we explore how we can improve both aspects. We are aware that focusing on the regulatory judgments covers only a small part of the entire regulatory system. However, this approach fits the concept of the learning organization, which allows people to examine small parts of working processes as units within a system (include regulatory systems) [81]. The concept of a learning organization was developed to clarify patterns and gain insight into possibilities for effectively changing patterns. The monitoring and improvement of the reliability and validity of judgments can be considered to be a component of the IGZ's performance. One

of the principles of learning organizations is that they develop as a result of the pressures faced by modern organizations, which enables them to adjust to new circumstances [82]. In the Netherlands, this can be illustrated by the introduction of market forces in health care in the Netherlands and the increased transparency of governmental organizations (including regulatory organizations). Both developments have influenced not only the content of the regulatory activities but also the system of regulation.

The main goal of this study is to identify and explore possibilities for improving the reliability and validity of IGZ regulatory judgments by investigating:

- the interrater reliability of nursing home care inspectors
- the validity of the judgments of nursing home care inspectors
- two types of regulatory instruments in relation with to accountability
- interventions for improving interrater reliability
- the effect of interventions for improving the interrater reliability and validity of IGZ nursing home care inspectors

We have the following research questions:

1 Do IGZ inspectors systematically differ in the regulatory judgments they assign to similar health care institutions? (Chapter 2)

2 Do IGZ inspectors assign judgments to health care institutions that conform to the corporate standards and thus result in valid judgments? (Chapter 3)

3 Do the reliability and validity of the regulatory judgments of IGZ inspectors vary between two types of regulatory instruments? (Chapter 4)

4 Which interventions are effective for increasing the interrater reliability of professionals? (Chapter 5)

5 Which interventions are effective for increasing the reliability and validity of the regulatory judgments of IGZ inspectors? (Chapter 6)

1.7 Methods and design

To answer the research questions, this study is divided into three parts. The first part of the study focused on analyzing interrater reliability and validity of regulatory judgments and the role of different types of regulatory instruments. The first three research questions were examined in this part.

A range of methods is available for researching interrater reliability. Experimental and quasi-experimental studies in which observers examine cases or patients are common [83-86]. Studies have also been

performed outside the laboratory setting so that participants would not be aware that data were being collected [38]. We performed the first part of the study by retrospectively analyzing regulatory reports. These reports were written by inspectors in regulatory practice in 2005 and 2006. The inspectors were not aware that the data would be used for a reliability study, and therefore the ecological validity of the data is good. Moreover, the Hawthorne effect did not apply in this analysis.

In the second part of our study, we examined which interventions are effective for increasing the interrater reliability of other professionals. We carried out a systematic review to answer the fourth research question. In the third part of the study, we investigated whether interventions that proved to be effective for increasing the reliability of other professionals were also effective for increasing the reliability and validity of the regulatory judgments of IGZ inspectors. We set up a case study and used a randomized controlled trial design and a before and after study to examine the effect of the interventions. Randomized controlled trials are frequently used in clinical studies. However, this design has rarely been used in reliability studies, although it has been done before [87]. The overview of the data used is depicted in Table 1.

Table 1 Overview of the data used

Data	Measure	Source	*Observations*
Data set 1	The reliability of IGZ inspectors	Analysis of judgments on the quality of care in nursing homes in 2005 and 2006.	4,914 judgments, 182 reports on nursing homes, 26 inspectors for the regulation of nursing homes.
Data set 2	Validity of judgments of IGZ inspectors	Analysis of the grounds for judgments on the quality of care in nursing homes in 2005 and 2006.	615 grounds for regulatory judgments, 182 reports on nursing homes, 26 inspectors for the regulation of nursing homes.
Data set 3	Suitability of regulatory instruments for providing accountability	Analysis of the suitability of two types of regulatory instruments for providing accountability for regulatory decisions.	4,914 judgments, 182 reports on nursing homes, 26 inspectors for the regulation of nursing homes. 615 grounds for regulatory judgments, 182 reports on nursing homes, 26 inspectors for the regulation of nursing homes. 520 judgments in 107 reports on hospitals, 11 inspectors for the regulation of hospital care.
Data set 4	Interventions effective for increasing the reliability of other professionals	Systematic review of literature about empirical studies on interventions for increasing the reliability of health care professionals.	57 studies.
Data set 5	Are interventions that proved to be effective for increasing the reliability of other professionals also effective for increasing the reliability and validity of the regulatory judgments of IGZ nursing home care inspectors?	Analysis of the reliability and validity of judgments on the quality of care in nursing homes using a case study in 2009.	32 vignettes, 25 inspectors.

The results of the study are presented in Chapters 2 through 6. The chapters were published as separate articles, and can thus be read independently of each other. Because of this, there will be some overlap in the content of the chapters. This is particularly the case in the introduction sections of the articles.

1.8 References

1 Healy J. Improving health care safety and quality. Reluctant regulators. Surrey, England: Ashgate Publishing Limited; 2011. p. 2.

2 Walshe K. Regulating Healthcare: A prescription for improvement? Philadelphia: Open University Press; 2003. p. 11.

3 IGZ (Dutch Health Care Inspectorate) Advisory Board. "Towards healthy trust" [Op weg naar gezond vertrouwen; in Dutch]. 2001:11.

4 IGZ. "Policy Plan 2012-2015. For justified trust in safe and appropriate care II" [Meerjaren beleidsplan 2012-2015. Voor gerechtvaardigd vertrouwen in verantwoorde zorg; in Dutch]. 2011.

5 IGZ. "Policy plan 2008-2011. For justified trust in safe and appropriate care" [Meerjaren Beleidsplan 2008-2011. Voor gerechtvaardigd vertrouwen in verantwoorde zorg; in Dutch]. 2007.

6 Steenhoven van der K. "Carry on! Organizational research on the Dutch Health Care Inspectorate" [Doorpakken! Organisatieonderzoek naar de Inspectie voor de Gezondheidszorg; in Dutch]. 2012.

7 Sorgdrager W. "From incident to effective regulation. Research on how the Dutch Health Care Inspectorate resolves files containing reports on health care incidents" [Van incident naar effectief toezicht. Onderzoek naar de afhandeling van dossiers over meldingen door de Inspectie voor de Gezondheidszorg; in Dutch]. 2012.

8 IGZ. "Regulating health care with confidence and authority" [Toezien met vertrouwen en gezag; in Dutch]. 2012.

9 Schippers EI. "Implementation of IGZ improvement plan" [Implementatie verbetertraject IGZ; in Dutch]. 2013.

10 Schippers EI. "Government reaction to the reports on the research reports on the IGZ " [Kabinetsreactie onderzoeksrapporten IGZ; In Dutch]. 2013.

11 Walshe K. Regulating Healthcare: A prescription for improvement? Philadelphia: Open University Press; 2003. p. 182.

12 Brennan TA. The Role of Regulation in Quality Improvement. The Milbank Quarterly 1998;76(4):709-731.

13 Luthi JC, McClellan WM, Flanders WD, Pitts S, Burnd-Hand B. Quality of health care surveillance systems: review and implementation in the Swiss setting. Swiss Med Wkly 2002;132:461-469.

14 Kollberg B, Elg M, Lindmark J. Design and Implementation of a Performance Measurement System in Swedish Health Care Services: A multiple case study of 6 development teams. Qual Manag Health Care 2005;14:95-111.

15 Pettersen IJ, Nyland K. Management and control of public hospitals – the use of performance measures in Norwegian hospitals. A case study. International Journal of Health Planning and Management 2006;21:133-149.

16 Lugtenberg M, Westert G. "Quality of health care and decision support-information for helping individuals to select health care. An international study on initiatives" [Kwaliteit van de gezondheidszorg en keuze-informatie voor burgers: een internationale verkenning van initiatieven; in Dutch]. 2007.

17 Dutch Parliament. "Framework approach of view to regulation II: Less of a burden, greater effect. Six principles of good regulation" [Kaderstellende visie op toezicht II, minder last meer effect. Zes principes van goed toezicht; in Dutch]. 2005;15.

18 Schippers EI. "Letter accompanying the IGZ research reports" [Aanbiedingsbrief onderzoeksrapporten IGZ; in Dutch]. 2012.

19 Scientific Council for Government Policy (WRR). "Overseeing the Public Interest: Towards a broader perspective on governmental regulation" [Toezien op Publieke Belangen. Naar een verruimd perspectief op rijkstoezicht; in Dutch]. 2013.

20 Walshe K. Regulating Healthcare: A prescription for improvement? Philadelphia: Open University Press; 2003. p. 183.

21 Nunnally JC. Psychometric Theory. New York: McGraw-Hill; 1978.

22 Uebersax JS. Validity inferences from observer agreement. Psychological Bulletin 1988;104(3):405-416.

23 De Groot AD. "Fives and Sixe's" [Vijven en Zessen; in Dutch]. Groningen: Wolters-Noordhoff; 1983. p. 122.

24 Bergh van den H, Zwarts M, Peter-Sips M. "Quality of the educational learning process" [Kwaliteit van het onderwijsleerproces; in Dutch]. Tijdschrift voor Onderwijsresearch 2000;25(1/2):20-39.

25 Meuffels B. "The maligned rater. On the expertise of the expert in Dutch as evaluator of essays" [De verguise beoordelaar. Over de deskundigheid van de neerlandicus als opstelbeoordelaar; in Dutch]. Tijdschrift voor Taalbeheersing 1989;11(3):161-176.

26 Krakenes J, Kaale B. MRI assessment of the alar ligaments in the late stage of whiplash injury-a study of structural abnormalities and observer agreement. Neuroradiology 2002;44:617-24.

27 Hilditch WG, Kopka A. Interobserver reliability between a nurse and anaesthesist of tests used for predicting difficult tracheal intubation. Anaesthesia 2004;59:881-4.

28 Dijkhuis JH. "Evaluation in psychology, in particular the evaluation of humans and human behavior" [Het beoordelen in de psychologie: in het bijzonder het beoordelen van mensen en menselijk gedrag; in Dutch]. Judgment in psychology. Especially the judgment of human and human behaviour.[Het beoordelen in de psychologie. In het bijzonder het beoordelen van mensen en menselijk gedrag.] [in Dutch] Utrecht: Bijleveld; 1961. p. 28.

29 Spanjer J. "Inter- and intrarater reliability of evaluations within the framework of the Dutch Disability Insurance Act" [De inter- en intrabeoordelaarsbetrouwbaarheid van WAO-beoordelingen; in Dutch]. Tijdschrift voor Verzekeringsgeneeskunde 2001;8:235-214.

30 Brouwer S, Dijkstra P, Gerrits PU, Schellekens EHJ, Groothof JW, Geertzen JHB, et al. "Intra- and interrater reliability for the 'FIS capability pattern' and 'list of funcitonal possiblities'" [Intra- en interbeoordelaarsbetrouwbaarheid 'FIS-Belastbaarheidspatroon' en 'Functionele mogelijkhedenlijst'; in Dutch]. Tijdschrift voor Verzekeringsgeneeskunde 2003(12):360:367.

31 Berghuis AC. "The heavy hand and the light touch: a statistical analysis of interrater reliability in the criminal justice system" [De harde en de zachte hand: een statistische analyse van verschillen in sanctiebeleid; in Dutch]. Trema 1992;15:84-93.

32 Brenninkmeijer AFM. "Similarity in punishment" [Gelijkheid van straffen; in Dutch]. Trema 1992;3:77-135.

33 Ashton RH, Ashton AH. Perspectives on judgment and decision-making research in accounting and auditing. In: Ashton RH, Ashton AH, editors. Judgment and decision-making research in accounting and auditing. Cambridge: Cambridge University Press; 1995. p. 5.

34 Kilpatrick G. Observer error in medicine. Journal of Medical Education 1963;38:38-43.

35 Browne T. Pseudodoxia Epidemica. Enquiries into very many commonly received tenets and commonly presumed truths. 1646.

36 Derryberry M. The reliability of Medical Judgments on Malnutrition. Public Health 1934;53:263.

37 Davies LG. Observer variation in Reports on Electrocardiograms. British Heart Journal 1958;20:153.

38 Jong de J. Explaining medical practice variation (dissertation). Ipskamp, Enschede: NIVEL; 2008.

39 Jacobs B, Duncan J. Improving Quality and Patient Safety by Minimizing Unnecessary Variation. Journal of Vascular and Interventional Radiology. 2009;20:157-163.

40 Köhler W. "Operations in the last year of life sometimes unwanted or unnecessary" [Operaties in laatste jaar soms ongewenst of onnodig; in Dutch]. *NRC Handelsblad* 2011.

41 Kelly AS. Treatment intensity at end of life – time to act on the evidence. The Lancet 2011;368(9800):1364-1365.

42 Lande van de S, Everdingen van JJE, Krol LJ. "Searching for a second opinion" [Op zoek naar een tweede mening; in Dutch]. NtvG 1993;137:1836-1840.

43 Edgeworth FY. The statistics of examinations. Journal of the Royal Statistical Society 1888;51:p. 600.

44 Starch D, Elliott ED. Reliability of the Grading of High School Work in English. School Review 1912;20:442-457.

45 Starch D, Elliott ED. Reliability of the Grading of High School Work in Mathematics. School Review 1913;21:245-459.

46 Schröter G. "The evaluation of essays in Germany" [Wie in Deutschland Aufsätze zensiert werden; in German]. Westermanns Pädagogische Beiträge 1970;22:408-417.

47 Meuffels B. "Contamination effects in the evaluation of essays" [Contaminatie-effecten bij het beoordelen van opstellen; in Dutch]. Tijdschrift voor Taalbeheersing 1991;13(1):15-29.

48 Ulshöfer R. "How objective are evaluations of essays" [Welcher Grad von Onbectivität lässt sich bei der Beurteilung deutcher Ausfsärtze erreichen?; in German]. Trema Straftoemetingsbulletin 2002:57-60.

49 Bergh van den H. "Examinations examined. A study of the validity of examinations in Dutch lower secondary vocational education (LBO) and junior general secondary education (MAVO)" [Examens geëxamineerd; in Dutch]. Amsterdam: University of Amsterdam; 1989.

50 Mellenbergh GJ "Studies on educational tests" [Studies in studietoetsen; in Dutch]. Amsterdam: Psychological Laboratory, University of Amsterdam; 1971.

51 Wesdorp H. "Measuring written language skills. Direct and indirect methods: 'evaluating essays' versus 'writing skills tests'" [Het meten van de produktief-schriftelijke taalvaardigheid: directe en indirecte methoden: opstelbeoordeling versus schrijfvaardigheidstoetsen; in Dutch]. Amsterdam, the Netherlands: Research Institute for Applied Psychology at the University of Amsterdam; 1974.

52 Eggen TJHM, Sanders PF. "Psychometry in practice" [Psychometrie in de praktijk; in Dutch]. Arnhem: Cito; 1993.

53 Better Education in the Netherlands (Beter Onderwijs Nederland). "LIA calls for external correction of secondary school exams" [LIA pleit voor externe examencorrectie in vo; in Dutch]. Available at: http://www.beteronderwijsnederland.nl/node/7915.

54 Everson G. The human element in justice. Journal of Criminal Law & Criminology 1919;10:90.

55 Feldt LS, Brennan RL. Reliability. In: R.L. Linn (Ed.), *Educational Measurement*. Third ed. New York: Macmillan Publishing Company; 1989. p. 105-146.

56 Franke R, Kaul J. The Hawthorne experiments: First statistical interpretation. American Sociological Review 1978;43:623-643.

57 Gaudet FJ, Harris GS, John St CW. Individual Differences in the Sentencing Tendencies of Judges. Journal of Criminal Law & Criminology 1933;23:811.

58 Frankel E. The offender and the Court: a statistical Analysis of the Sentencing of Delinquents. Journal of Criminal Law & Criminology 1940;31:448.

59 Smith AB, Blumberg AS. The Problem of Objectivity in Judicial Decision-Making. Social Forces 1967;46:96-105.

60 Duyne van P. "Simplicity in decision making" [Beslissen in eenvoud; in Dutch]. Arnhem: Gouda Quint; 1983.

61 Tata C, Hutton N. Rules in sentencing? Consistency and Disparity in the Absence of Rules. International Journal of the Sociology of Law 1998;26:339-364.

62 Frijda L. "The Supreme Court of the Netherlands and the grounds for sentencing" [De Hoge Raad en de motivering van strafoplegging; in Dutch]. D&D 1988;10.

63 Kas A. "Legal inequality in court " [Rechtsongelijkheid bij de rechter; in Dutch]. *NRC Handelsblad* 2012 14-03:6.

64 Salomon F. "Even the judiciary is treated with contempt" [Ook rechterlijke macht valt minachting ten deel; in Dutch]. *De Volkskrant* 2011.

65 Klazinga N. Re-engineering trust: the adoption and adaption of four models for external quality assurance of health care services in western European health care systems. International Journal for Quality in Health Care 2000;12:183-189.

66 Taylor BJ. Factorial Surveys: Using Vignettes to Study Professional Judgement. British Journal of Social Work 2006;36:1187-1207-1188.

67 OECD. Recommendation Of The Council On Regulatory Policy And Governance. 2012.

68 OECD. Principles For The Governance Of Regulators. 2013.

69 OECD. Public Consultation On Best Practice Principles For Improving Regulatory Enforcement And Inspections. 2013.

70 Bos van den K. "Justice and insecurity" [Rechtvaardigheid en onzekerheid; in Dutch]. In: Tiemeijer WL, Thomas CA, Prast HM, editors. "The human decision maker" [De menselijke beslisser; in Dutch] Amsterdam: Amsterdam University Press; 2009. p. 89.

71 Scrivens E. Putting continuous quality improvement into accreditation: improving approaches to quality assessment. Quality in Health Care 1997;6:212-218.

72 Wagner C, Merode van GG, Oort van M. Cost of quality management systems in long-term care organisations: an exploration. Quality Management in Health Care 2003;12:106-114.

73 Zwakkenberg SPM, Croonenborg van JJ, Drevers S, Barneveld van TA. "Costs of quality indicators. A brief study on the costs of developing indicators and collecting data on indicators in hospitals" [Een beknopte studie naar de kosten van indicatorontwikkeling en het uitvragen van indicatoren in ziekenhuizen; in Dutch]. 2007.

74 Sira Consulting. "Measurement of supervisory burden for hospitals. Research on administrative burden, costs of compliance, and hospitals' perception of supervision" [Meting toezichtlasten ziekenhuizen; in Dutch]. 2007.

75 Greenfield D, Braithwaite J. Health sector accreditation research: a systematic review. International Journal for Quality in Health Care 2008;20:172-183.

76 Hood C, Rothstein H, Baldwin R. The Government of Risk. Understanding Risk Regulation Regimes. Oxford, New York.: Oxford University Press; 2001.

77 Day P, Klein R. The regulation of nursing homes. The Milbank Quarterly 1987;65(3):303-347.

78 Hutter BM. Variations in regulatory enforcement styles. Law and Policy 1989;2:153-174.

79 Braithwaite J, Makkai T, Braithwaite V. Regulating Aged Care. Ritualism and the New Pyramid. Cheltenham, UK: Edward Elgar; 2007.

80 Feldt LS, Brennan RL. Reliability (p.107) . In: *Educational Measurement*. Eds: Linn, R.L. . Third ed. New York: Macmillan Publishing Company; 1989. p. 107.

81 Senge PM. The Fifth Discipline. London: Century Business; 1990.

82 O'Keeffe T. Organizational Learning: a new perspective. Journal of European Industrial Training 2002;26:130-141.

83 Craig WL, Dirschl DR. Effects of binary decision making on the classification of fractures of the ankle. Journal of Orthopeadic Trauma 1998;12:280-283.

84 Hofer T, Bernstein SJ, DeMonner S, Hayward RA. Discussion between reviewers does not improve reliability of peer review of hospital quality. Medical Care 2000;38:152-161.

85 Mikami Y, Manabe T, Epstein JI, Shiraishi T, Furusato M, Tsuzuki T, et al. Accuracy of Gleason grading by practicing pathologists and the impact of education on improving agreement. Human Pathology 2003;34:658-665.

86 Cook C, Braga-Baiak A, Pietrobon R, Shah A, Neto AC, Barros de N. Observer agreement of spine stenosis on magnetic resonance imaging analysis of patients with cervical spine myelopathy. Journal of Manipulative and Physiological Therapy. 2008;31:271-276.

87 Brorson S, Bagger J, Sylvest A, Hrøbjartsson A. Improved interobserver variation after training of doctors in the Neer system. A randomized trial. Journal of Bone & Joint Surgery (Br) 2002;84(7):950-954.

Chapter 2

Not all judgments are the same: a quantitative analysis of the interrater reliability of inspectors at the Dutch Health Care Inspectorate

This chapter is published as:

Tuijn SM, Janssens FJG, Van den Bergh H, Robben PBM. Not all judgments are the same: a quantitative analysis of the interrater reliability of inspectors at The Dutch Health Care Inspectorate. [Het ene oordeel is het andere niet: Interinspecteursvariatie bij inspecteurs van de IGZ: een kwantitatieve analyse] [In Dutch]. *Nederlands Tijdschrift voor Geneeskunde 2009(8)*;322:326

2.1 Abstract

To examine the interrater reliability of inspectors at the Dutch Health Care Inspectorate and to study the correlation between judgments and the presence of grounds for the judgments.

We collected and analyzed 4,914 judgments (and the grounds that accompanied them) made by 26 Dutch Health Care Inspectorate inspectors. These judgments were taken from 182 regulatory reports, and were assigned in 2005 and 2006 using 25 criteria for good care practices in nursing homes. To explain variation, the data were statistically corrected for institutional characteristics by analysis of covariance.

Poor interrater reliability is a cause for concern in risk-based supervision of nursing homes. Statistically significant differences were found for 22 of the 25 criteria. After correcting for institutional characteristics, these differences remained significant for 14 of the criteria. The presence of grounds for the judgments depended on the inspector as well as on the judgment.

Interobserver disagreement is a cause for concern in risk-based supervision of nursing homes. The results of this research will be used to further fine-tune regulation by the Dutch Health Care Inspectorate, and are also relevant for other regulatory authorities.

2.2 Introduction

In the Netherlands, health care regulation is performed by the Dutch Health Care Inspectorate (IGZ). When the personal views of individual inspectors are taken too much into account in the regulation of health care, this can be detrimental to the trustworthiness and authority of the IGZ [1]. Consequently, public trust in health care can be damaged. The IGZ acknowledges this concern, and aims to standardize its procedures and account for its regulatory activities [2]. Standardizing working processes is a familiar way of increasing uniformity among professionals in a wide variety of professions. In medical professions, guidelines and protocols as well as standardizing criteria, consensus meetings, and software that support medical decision making are common ways of stimulating uniformity among professionals.

To regulate health care in the Netherlands, the IGZ uses three types of regulation: proactive risk-based supervision, theme-based regulation, and regulation in response to calamities and incidents. Risk-based supervision consists of three phases. Before the first phase begins, the IGZ draws up the standards for safe, appropriate care and the corresponding quality indicators in consultation with different stakeholders. These indicators are proxies for measuring the quality of care. Next, in the first phase, the IGZ analyses the data collected with these quality indicators and selects institutions at risk (institutions in which health care risks are likely to occur). In the

second phase, inspectors visit the selected institutions and assign scores to predetermined criteria that represent the quality of care. Institutions are obliged to improve their care based on the inspectors' judgment. Inspectors can decide to plan a follow-up visit if these improvements are not satisfactory. In the third phase, the IGZ can impose administrative sanctions and initiate criminal proceedings [3].The IGZ uses risk-based supervision in the regulation of many health care sectors, such as hospital care, nursing home care, mental health care, care of the disabled, public health care, maternity care, public pharmacies, and private clinics [3]. Theme-based regulation is directed at specific issues in care that require the attention of supervisors. The IGZ employs regulation in response to incidents and in the event of emergencies that indicate structural shortcomings in health care. In these three forms of regulation, inspectors assign scores to aspects of care. Standardizing procedures is important to this.

Earlier research on how inspectors carry out their professional duties and their perception of regulation has shown that standardizing working processes does not automatically result in uniformity among IGZ inspectors. Inspectors differ in the way they work as well as in how they perceive the role of the IGZ [4]. These results emphasize the necessity of studying the interrater reliability of inspectors. Therefore, we analyzed data collected during regulatory visits for risk-based supervision, with the aim of answering three research questions: (a) Are there differences between inspectors in their regulatory judgments (b) Can these differences be explained by characteristics of the inspectors or institutions? (c) Do inspectors provide grounds for their judgments and to what extent does their presence depend on the inspector and on the judgment assigned?

2.3 Methods

This study focuses on analyzing the interrater reliability of regulatory judgments on nursing homes that were assigned as part of risk-based supervision in 2005 and 2006 (n=645). To examine the quality of care, inspectors (n=31) employed a regulatory instrument consisting of 27 criteria. The criteria are presented in Table 1. These criteria are a combination of measures of structure, processes, and outcomes. For example, for the criterion "pressure ulcers," inspectors assess whether the prevalence of pressure ulcers is recorded by the staff (process) as well as whether the staff has used a protocol for pressure ulcers (structure). During regulatory visits, inspectors use these criteria to examine the quality of care, and assign scores to the criteria on a four-point scale: "absent", "present", "operational" and "secured".

Table 1 Criteria for safe and appropriate care used to regulate nursing homes in 2005 and 2006.

Criteria		
Measurement of intensity of care	Procedure for complaints	Competences required by the Individual Health Care Professions Act (BIG)
Pressure ulcers	Working systematically with care plans	Professional conduct
Reporting incidents to commission	Recording deviations in individual care plans	Treatment of clients
Safety of materials	Procedure for assigning responsibility for content and coordination of care plans	Long-term policy
Satisfaction of clients	Procedure for multidisciplinary meetings to discuss clients	Annual work plan
Client board	Rights of clients regarding care plans	Organizational structure
Information for clients	Training for using care plans	Management information system
Sufficient help with eating and drinking	Availability and competence of care providers	System for monitoring an appropriate level of care
Continuous supervision in living rooms	Training plan	Privacy

These scores ascend from negative (absent) to positive (secured). The regulatory instrument describes exactly which judgment applies in which situation. Table 2 illustrates the corporate standards using the standard for the criterion "reporting incidents to commission." If the judgment "absent" or "present" applies, nursing homes are required to take measures to improve care.

Table 2 Framework with IGZ standards that define precisely which judgment applies in which situation for the criterion "reporting incidents to commission".

Criterion "reporting incidents to commission"	Absent	Present	Operational	Fulfilled
The organization uses procedures to record and evaluate incidents and accidents. If necessary, preventive and corrective measures are taken based on this. The organization has a system in place for recording and dealing with reports of incidents in health care and provision of services. Recording incidents is always part of the quality system.	Incidents are neither recorded nor evaluated.	Incidents are recorded.	The results of incident analysis are used to improve care.	The system for reporting incidents is systematically evaluated and adjusted if necessary.

An analysis was performed on a set of data that consisted of inspectors' (n=31) regulatory judgments of nursing homes (n=645) in 2005 and 2006. For this study, we used the judgments of inspectors who wrote at least seven regulatory reports on their regulatory visits in these years (n=26). These reports had to meet the requirements of the IGZ format. This format implies that inspectors substantiated and reported their judgments in the format's tables. This criterion guaranteed there would be sufficient observations per inspector and also that a maximum number of inspectors (26 of the 31) could be included in the study. From the reports of the inspectors who met this inclusion criterion (a minimum of 9 and a maximum of 41 reports per inspector), 7 reports per inspector were selected at random. In total, 182 reports were analyzed, consisting of 4,914 judgments. The reports described visits to long-term care facilities, which varied from nursing homes (n=123), care centers (n=42), homes for the elderly (n=11), and long-term care units (n=6). The different types of care institutions were randomly

divided among inspectors. Because all of the institutions offered long-term care and were examined using the same criteria, we will refer to all of the institutions as nursing homes.

2.4 Analysis

The 4,914 judgments were recorded in a data file. For each inspector, the mean judgment per criterion was calculated for the seven nursing homes. An analysis of variance (Anova) was performed to assess the interrater reliability of inspectors and determine sources of variation. To visualize how the mean judgments of inspectors related to each other, 80% confidence intervals were calculated for individual inspectors [5]. We performed an analysis of variance to investigate to what extent the grounds for the judgments depended on the inspector. To investigate the interaction between the independent variables "inspector" and "type of judgment," an analysis of variance was performed as well. An alpha level of 0.05 was employed in all analyses (Anova, SPSS 15). Not all of the 27 aforementioned criteria were included in the analysis. Two of the criteria ("privacy" and "training for using care plans") were not judged by the majority of inspectors, and so were excluded from analysis.

To examine the extent to which interrater reliability can be explained by characteristics of the inspectors or institutions, it would have been preferable if several inspectors had visited and judged the same nursing homes [5]. However, having groups of inspectors visit the same institution is not common practice in IGZ regulation. Therefore, to simulate the ideal situation as much as possible, we corrected for institutional characteristics in our analysis. Consequently, in the analysis of variance, we included the judgments on 24 of the 25 criteria as covariates. The mean judgment per inspector was corrected based on the judgments of all inspectors for the other 24 criteria that were included as covariates. Because of this, the heterogeneous group of nursing homes was homogenized on 24 criteria, and the precondition of correcting for institutional features was done to the greatest possible extent. We used the same correction to analyze differences between inspectors on the grounds they provided for their judgments.

2.5 Results

The analysis of variance showed significant differences in judgments for 22 of the 25 criteria. The effect size (η^{2}) varied between 0.2 and 0.4. This means that 20% to 40% of the total variance could be explained by differences between inspectors. After correcting for nursing home characteristics, significant differences remained for 14 of the 25 criteria. Of the total variance, 30% to 40% could be explained by differences between inspectors;

this represents a moderate effect. One of the criteria for which significant differences were found is "reporting incidents to commission" ($F = 1.652$, $p = 0.041$, $\eta^* = 0.3$, $df_1 = 25$, $df_2 = 133$).

To visualize how the mean judgments of inspectors relate to each other, 80% confidence intervals were calculated for individual inspectors. Figures 1a and 1b present the mean judgment of each inspector with the 80% confidence interval for the criterion "reporting incidents to commission." Figure 1a shows large differences between judgments; the mean judgment differs between inspectors. The results indicate that correcting for nursing home characteristics had a large effect on the judgments, as shown in Figure 1b. The correction resulted in changes in the mean judgments and in the scope of the confidence intervals.

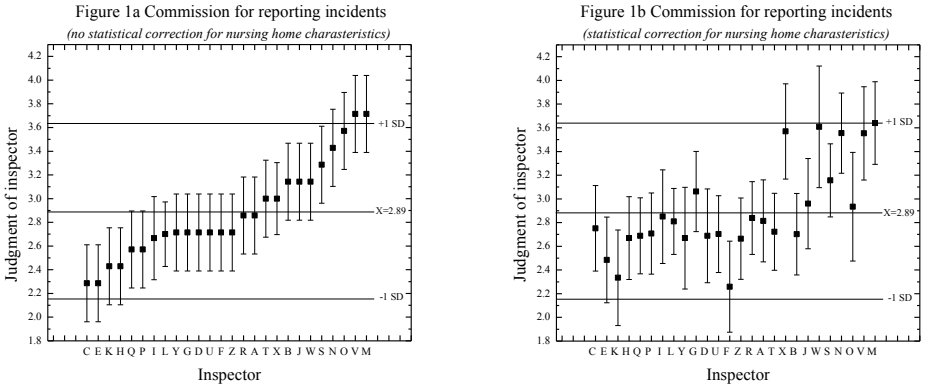


Figure 1a and 1b Average judgment for the criterion "reporting incidents to commission" for seven nursing homes represented in 80% confidence intervals for individual inspectors before and after correction for nursing home characteristics

For all 25 criteria, significant differences were shown for the presence or absence of grounds for the judgments. After correcting for nursing home characteristics, these differences remained for 22 of the 25 criteria. The results indicate that whether grounds are provided depends on the inspectors. One of the criteria that demonstrated this effect was "professional conduct" ($F = 6.699$, $p < 0.001$, $df_1 = 25$, $df_2 = 78$). The proportion of explained variance was large ($\eta^* = 0.5$). An interaction effect was found for 9 of the 22 criteria. For example, this effect was found for the criterion "measurement of the extent of care needed" ($F = 2.047$, $p < 0.001$, $df_1 = 75$, $df_2 = 78$). The proportion of explained variance was large as well ($\eta^* = 0.5$). For these nine criteria, the results indicate that whether grounds are provided depends on the inspector as well as on the type of judgment.

2.6 Discussion

The results of this study indicate that reliability issues are a cause for concern in the IGZ's risk-based supervision. The effect sizes we found varied from moderate to large (η^{*} > = 0.3) and were present in the majority of the examined criteria. Because we statistically corrected for nursing home characteristics, this variance cannot be explained by nursing home characteristics alone. Therefore, it is plausible that the reliability issues are based at least in part on differences between inspectors. The results indicate that for 22 of the 25 criteria, whether grounds were provided for judgments depended on individual inspectors. An interaction effect was found in 9 of the 25 criteria: grounds for judgments that were more positive were provided less frequently. However, this differed between inspectors.

Reliability issues are not unique to IGZ regulation. They also occur in the regulatory judgments of inspectors at the Netherlands Food and Consumer Product Safety Authority [5] and in those of inspectors at the Dutch Inspectorate of Education [4]. Reliability issues also appear in other professions, for example, judges [6-8], teachers [9], insurance physicians [10,11], and medical specialists [12-15]. In general, reliability issues are a major concern in all of these professions, and research is being conducted to examine how to improve reliability [16-19].

2.7 Limitations of the study

All research is characterized by methodological strengths and limitations, and this study is no different. First, in this analysis we treated ordinal data as discrete data: the categories "absent," "present," "operational," and "fulfilled" were changed into a score that ascended from one to four, which created an interval scale. However, calculating the average on this scale was problematic, because the average does not fit into one of the four semantic categories. Nevertheless, treating ordinal data as discrete or continuous data is a frequently used technique for analyzing ordinal data, and generally causes very minor distortions [4]. Second, we examined the reliability of regulatory judgments that were assigned during on-site regulatory visits. Because of this, the ecological validity of our study is high. As a consequence, we had to statistically correct for nursing home characteristics later on to be able to explain sources of variation. Third, the definition of the criteria and the method of judging influence the interrater reliability as well. It is possible that inspectors used different criteria to judge the quality of care because they worked with different definitions of the criteria. In other words, the validity of the judgments can also be a source of the systematic differences we found between inspectors.

Moreover, the regulatory instrument inspectors used was refined during the period in which the regulatory visits took place. Although the criteria remained unchanged, small adjustments to formulations may have caused interrater disagreement. Therefore, it is possible that the systematic differences we found can also be explained in part by the method of judging.

In this study we used judgments that were assigned to institutions during on-site visits. The institutions included differed slightly. In future research, we will examine the reliability of regulatory judgments using a case study in which inspectors all assign scores to the same cases.

2.8 Implications

The fact that systematic differences between the judgments of inspectors in the IGZ's risk-based supervision are a cause for concern implies that the chance of a positive or negative regulatory judgment depends not only on the health care characteristics according to predetermined criteria, but also on the individual inspector who visits the health care institution. Misclassification can result in requiring no improvement measures even though health care risks may be present. Subsequently, health care institutions that need closer monitoring are not monitored enough. On the other hand, institutions can also be monitored too closely and be wrongfully required to improve their care. It is important to emphasize that in the second phase of risk-based supervision, the IGZ never applies severe sanctions on institutions based only on the judgments on the criteria. Moreover, a lack of grounds is problematic mainly in the case of negative judgments. Providing grounds is important to the institution's acceptance of the judgment. If no grounds are provided, this can hamper acceptance. Moreover, such grounds frequently contain actual ways to improve care, which remain unclear when no grounds are provided.

2.9 Future research

This study is part of an extensive research program examining interrater reliability and validity of regulatory judgments. The objective of this program is both to gain insight into the reliability and validity of regulatory judgments as well as into factors that can explain suboptimal reliability and validity. This insight will be used to develop an intervention to increase the reliability as well as the validity of judgments. This intervention is in addition to current initiatives that aim to increase reliability, like training and regulatory audits. The effect of the intervention will be examined by means of a case study.

2.10 Acknowledgements

We would like to express our grateful thanks to J.A.H. van Veen, chief inspector for Nursing Homes and Long-Term Care (IGZ) and Dr. C.A.J. Ketelaars, inspector (IGZ) for their valuable comments on the manuscript.

2.11 References

1 Janssens FJG. "From research to evaluation. The methodology of the Dutch Inspectorate of Education" [Van onderzoek naar evaluatie. De methodologie van de Onderwijsinspectie; in Dutch]. 1997.

2 Hutschemaekers G, Kist S. "The profession of inspector at the Dutch Healthcare Inspectorate" [Beroep Inspecteur; in Dutch]. 2006.

3 IGZ. "Policy Plan 2012-2015. For justified trust in safe and appropriate care II" [Meerjaren beleidsplan 2012-2015. Voor gerechtvaardigd vertrouwen in verantwoorde zorg; in Dutch]. 2011.

4 Bergh van den H, Zwarts M, Peter-Sips M. "Quality of the educational learning process" [Kwaliteit van het onderwijsleerproces; in Dutch]. Tijdschrift voor Onderwijsresearch 2000;25(1/2):20-39.

5 Mascini P, Wijk van E. "This is just how fish smells. Responsive Regulation at the Dutch Food and Consumer Product Safety Authority" [Vis ruikt nu eenmaal zo'. Responsive Regulation bij de Voedsel en Waren Autoriteit; in Dutch]. Tijdschrift voor criminologie 2008;59(2):114-129.

6 Berghuis AC. "The heavy hand and the light touch: a statistical analysis of interrater reliability in the criminal justice system" [De harde en de zachte hand: een statistische analyse van verschillen in sanctiebeleid; in Dutch]. Trema 1992;15:84-93.

7 Duyne van P. "Simplicity in decision making" [Beslissen in eenvoud; in Dutch]. Arnhem: Gouda Quint; 1983.

8 Hendriks PAM. Towards consistent determination of penalties [Op weg naar consistente straftoemeting] [in Dutch]. Trema Straftoemetingsbulletin 2002:57-60.

9 Meuffels B. "Contamination effects in the evaluation of essays" [Contaminatie-effecten bij het beoordelen van opstellen; in Dutch]. Tijdschrift voor Taalbeheersing 1991;13(1):15-29.

10 Spanjer J. "Inter- and intrarater reliability of evaluations within the framework of the Dutch Disability Insurance Act" [De inter- en intrabeoordelaarsbetrouwbaarheid van WAO-beoordelingen; in Dutch]. Tijdschrift voor Verzekeringsgeneeskunde 2001;8:235-214.

11 Kerstholt JH, Boer de WEL, Jansen NJM. Disability assessments: Effects of response mode and experience. Disability and Rehabilitation 2006;28(2):111-115.

12 Gaasterland D, Blackwell B. The advanced glaucoma intervention study (AGIS):10. Variability among academic glaucoma subspecialists in assessing optic disc notching. Tr Am Opthth Soc 2001;99:177-186.

13 Hilditch WG, Kopka A. Interobserver reliability between a nurse and anaesthesist of tests used for predicting difficult tracheal intubation. Anaesthesia 2004;59:881-4.

14 Krakenes J, Kaale B. MRI assessment of the alar ligaments in the late stage of whiplash injury-a study of structural abnormalities and observer agreement. Neuroradiology 2002;44:617-24.

15 Raghoebar-Krieger H, Sleijffer D. The reliability of logbook data of medical students: an estimation of interobserver agreement, sensitivity and specificity. Medical Education 2001;35:624-31.

16 Stoop A, Berg M. "Mirrors for the physician" [Spiegels voor de arts; in Dutch]. Gezondheid: theorie en praktijk 1994:265-278.

17 Rogers K, Patel K. Intra- and Interobserver variability in interpretation of DMSA-scans using a set of standardized criteria. Pedriatic Radiology 1993;23(7):506-509.

18 Vet de HC, Koudstaal J, Kwee WS, Willebrand D, Arends JW. Efforts to improve interobserver agreement in histopathological grading. J Clin Epidemiol 1995;48(7):869-873.

19 Taylor M, Hipp JA, Gertzbein SD, Gopinath S, Reitman CA. Observer agreement in assessing flexion-extension X-rays of the cervical spine, with and without the use of quantitative measurements of intervertebral motion. Spine J. 2007;7(6):654-658.

Chapter 3

The relation between the employment of standards and judgments in the regulation of health care

This chapter is published as:

Tuijn SM, Van den Bergh H, Robben PBM, Janssens FJG. The relationship between standards and judgments in the regulation of health care. [in Dutch] (De relatie tussen normen en oordelen in het toezicht op de gezondheidszorg). *Tijdschrift voor Gezondheidswetenschappen 2009*;6:264-271.

3.1 Abstract

In this study, we examined the accuracy of regulatory judgments of inspectors at the Dutch Health Care Inspectorate (IGZ). The study concerns judgments assigned in 2005 and 2006 to four criteria developed to examine the quality of care in nursing homes by means of risk-based supervision.

We analyzed the grounds for the judgments (n=615) assigned to the criteria "pressure ulcers," "sufficient help with eating and drinking," "permanent supervision in living rooms," and "the extent of care needed." We analyzed to what extent the grounds for the judgments corresponded to the IGZ regulatory standards (corporate standards). By doing this, it was possible to study to what extent the actual judgments corresponded with the judgments that should have been assigned to the institutions based on the present arguments and strict employment of the IGZ standards (corporate judgments). Two independent observers analyzed the accuracy of the actual judgments.

The results of this study indicate a problem with the validity of the judgments: the meaning of similar actual judgments differed widely. For the four examined criteria, we found 52% false-positive judgments and 1% false-negative judgments. The results indicate no correlation between the percentage of false-positive judgments and the mean judgment for the four criteria. This implies that false-positive judgments are assigned by inspectors whose judgments are relatively negative as well as by inspectors whose judgments are relatively positive.

Because all inspectors included in this study assigned false-positive judgments, this is therefore a characteristic of these inspectors that is associated with the regulation of nursing homes. The percentage of false-positive judgments depends on the criterion judged. Because of this, the presence of false-positive judgments also depends on the quality of the regulatory instrument used.

3.2 Introduction

Issues surrounding the reliability of regulatory judgments of inspectors at the Dutch Health Care Inspectorate (IGZ) are a matter of concern. Earlier research has shown that interrater reliability is suboptimal in risk-based supervision of nursing homes [1]. It appears that the judgment assigned to the quality of care (or aspects of the quality of care) in nursing homes depends on characteristics of the institution as well as those of the individual inspector: some inspectors are systematically more stringent in their judgments compared to other inspectors. However, the accuracy of these judgments is still unknown. If inspectors do not judge a criterion according to the IGZ standards, these judgments can be considered to be inaccurate, because the meaning of the judgments

deviates from the IGZ corporate standards. In this study, we examined the validity of regulatory judgments. We evaluated to what extent the actual judgments corresponded with judgments that would have been assigned if the IGZ standards had been strictly employed (corporate judgment).

3.3 Criteria for safe and appropriate care in nursing homes in risk-based supervision by the IGZ

The IGZ performs three types of regulation: risk-based supervision, theme-based regulation, and regulation in response to calamities and incidents [1]. Risk-based supervision consists of three phases: collecting information, assigning judgments to the quality of care, and intervention [2]. In the second phase of risk-based supervision, inspectors assign scores to health care institutions using predetermined criteria. If necessary, based on these judgments, the institutions are obliged to take measures to improve care. If these measures are not satisfactory, the third phase of risk-based supervision begins. In this phase, the IGZ can impose administrative sanctions and start criminal as well as disciplinary proceedings.

In 2005 and 2006, inspectors examined the quality of care using a regulatory instrument consisting of 27 criteria. With this instrument, inspectors assigned scores on a four-point scale: “absent,” “present,” “operational,” and “secured.” These scores ascend from negative (absent) to positive (secured). When the judgment “absent” or “present” applies for certain criteria, nursing homes are obliged to take measures to improve care for these criteria. The regulatory instrument describes exactly which judgment applies in which situation. For example, according to the instrument, if a nursing home does not record the presence of pressure ulcers for its residents, the criterion “pressure ulcers” should be judged as “absent” (Table 1). If inspectors assign the judgment “present” in this situation, the meaning of this judgment does not correspond to the meaning of the judgment according to the IGZ standard. In this case, the judgment is considered inaccurate and a validity problem rises [3-6].

This study is a continuation of earlier research on the reliability of the regulatory judgments on 25 of the 27 criteria for the quality of care in nursing homes assigned by IGZ inspectors that showed significant differences in inspectors’ judgments for 14 of the 25 criteria [1]. However, we have not yet examined possible explanations for this disagreement. One explanation might be that inspectors assign different meanings to the corporate standards. If the actual judgments deviate from the corporate standards, a validity problem arises. In this case, the judgments are not unambiguous because it is unclear what the judgment “present” or “operational” means, and the meaning of judgments then depends on individual inspectors.

Table 1 Overview of the four analyzed criteria and corresponding IGZ corporate standards for regulation of nursing homes in 2005/2006.

Criterion	Definition	Absent	Present	Operational	Secured
The extent of care needed	The institution employs a system to record the extent of care needed.	The institution has no notion of the extent of care needed and/or the extent of care needed is not recorded systematically.	The extent of care needed is recorded systematically.	The data on the extent of care needed are used for management purposes.	The data on the extent of care needed are used for management purposes, and the system for recording the extent of care needed is systematically evaluated and adjusted if necessary.
Pressure ulcers	The institution employs a protocol for preventing, treating, and recording pressure ulcers. This protocol meets the quality requirements of the CBO. Both the system for measuring pressure ulcers and the protocol are evaluated systematically and adjusted if necessary.	Pressure ulcers are not recorded.	Pressure ulcers are recorded and the institution has a protocol for pressure ulcers.	Pressure ulcers are recorded and measurements of pressure ulcers take place. Employees are aware of the protocol, and are trained in its use.	Both the system for measuring pressure ulcers and the protocol are evaluated systematically and adjusted if necessary.
Continuous supervision in living rooms	A trained volunteer or employee is always present to guarantee supervision of psychogeriatric clients in the living rooms.	No guideline is available for adequate supervision of psychogeriatric clients in the living rooms by employees or trained volunteers.	Guidelines are available (including a guideline with instructions for those who will supervise psychogeriatric clients). Employees are not aware of these guidelines, or deviate from them on multiple occasions (more than 10% of the time). Or volunteers/relatives of clients who supervise psychogeriatric clients in living rooms are given no instructions.	Guidelines are available (including a guideline with instructions for those who will supervise psychogeriatric clients). Employees are aware of these guidelines, or seldom deviate from them (less than 10% of the time). Or volunteers/relatives of clients who supervise psychogeriatric clients in living rooms are given instructions.	Guidelines and the employment of the guidelines are periodically evaluated and adjusted if necessary.
Sufficient help with eating and drinking	There are sufficient trained employees/relatives of clients/ volunteers available to help with clients' food and drink.	No guideline is available on the presence of sufficient trained employees/relatives of clients/ volunteers to help with clients' food and drink.	Guidelines are available on the presence of sufficient trained employees/relatives of clients/ volunteers to help with clients' food and drink. Employees are not aware of these guidelines, or deviate from them on numerous occasions (more than 10% of the time). Or volunteers/relative of clients who help clients with food and drink receive no instructions.	Guidelines are available on the presence of sufficient trained employees/relatives of clients/ volunteers to help with clients' food and drink. Employees are aware of these guidelines, and seldom or never deviate from them (10% or less than 10% of the time).	Guidelines and the employment of the guidelines on the presence of sufficient trained employees/relatives of clients/ volunteers to help with clients' food and drink are periodically evaluated and adjusted if necessary.

3.4 Methods

To analyze to what extent the inspectors' actual judgments corresponded with the IGZ corporate standards (corporate judgments) in 2005/2006, we used the same data set as in the study on interrater reliability [1]. In this data set, only the judgments of inspectors (n=26) who had written a minimum of nine reports about their regulatory visits during 2005/2006 were made part of the file. Only reports that complied with the IGZ format were included. This format implies that inspectors substantiated and reported their judgments in the tables intended for this purpose. Seven reports were randomly selected for each inspector, and a total of 182 reports were analyzed. We studied the judgments and the accompanying grounds for 4 of the 27 criteria: "pressure ulcers," "sufficient help with eating and drinking," "permanent supervision in living rooms," and "the extent of care needed" (Table 1). We chose these criteria because they are important in regulating the safety of care. Furthermore, in 2005/2006, those criteria determined the public image of nursing home care in the Netherlands to a considerable extent. The judgments and accompanying grounds (n=615) were anonymously processed in a data file and analyzed. The arguments presented in the grounds were analyzed, and compared to the corporate standards. Subsequently, the corporate judgments were noted. These concern the judgments that would have applied if the corporate standards had been strictly employed.

3.5 Analyses

Two observers independently evaluated the grounds. One of the observers was a former nursing home care inspector knowledgeable about and experienced in using the criteria and corporate standards. The other observer (the first author of this article) has work experience as a researcher. As presented in Table 1, the corporate standards prescribe assigning the judgment "absent" when the nursing home staff does not record pressure ulcers. The judgment "present" applies when the nursing home staff records pressure ulcers and also has a protocol for pressure ulcers. The judgment "operational" applies when the nursing home staff is trained in the treatment of pressure ulcers, the nursing home has a protocol for pressure ulcers, and the staff is acquainted with this protocol. The judgment "secured" applies when the institution has a protocol for pressure ulcers, pressure ulcers are recorded, and the protocol and system for recording pressure ulcers are evaluated and adjusted if necessary. If the grounds state that "the nursing home has a plasticized card for pressure ulcers, but the staff is not aware of this," the observers assigned the corporate judgment "absent," because no arguments were present about the

existence of a protocol for pressure ulcers. This was the procedure the observers followed in evaluating the grounds for judgments on the four criteria.

When the observers disagreed, consensus was reached by discussion. In this, the arguments of the observer who was a former inspector were decisive. A high degree of agreement was shown for the criteria "sufficient help with eating and drinking" and "permanent supervision in living rooms." There was less agreement for the criteria "the extent of care needed" and "pressure ulcers" (Table 2).

Table 2 Agreement (%) between the two independent observers in analyzing the grounds for judgments for the four criteria used in the regulation of nursing homes in 2005/2006.

	Measurement of the extent of care needed	Pressure ulcers	Continuous supervision in living rooms	Sufficient help with eating and drinking
	(n=168)	(n=151)	(n=145)	(n=151)
% agreement	62	64	93	85

3.6 Results

The meaning of similar judgments varies widely. For example, the judgment "secured" for the criterion "sufficient help with eating and drinking" has the following meanings:

- "The deployment of staff during meals in the nursing units and enhanced care departments is not explicitly described."
- "The policy for meals and drinks, as well as the working procedures and the recording of this in the individual care plans, is transparent and thoroughly described according to the HKZ protocol. The policy of weighing clients is part of this as well. The protocol describes the responsibility of staff involved in helping with food and drink. The staff is acquainted with the procedures."
- "There is a guideline on the presence of sufficient help with food and drink for clients and this has been put into practice (secured)."

Moreover, the actual judgments do not always correspond to the corporate judgments (Table 3). Compared to the corporate judgments, the majority of the judgments (52%) is too positive, and result in false-positive judgments. The minority of the judgments (1%) are too negative compared to the corporate judgment, and result in false-negative judgments [5]. Table 3 shows the distribution of false-positive (dark grey) and false-negative judgments (light grey). Table 3 demonstrates that, compared to the corporate judgment, 22% of the false-positive judgments are two or more categories higher (too positive).

Table 3 Survey of the number of corporate judgments by the two independent observers in 2008 and the number of actual judgments by inspectors in 2005/2006.

		Actual judgment in 2005/2006				
		Absent	Present	Operational	Secured	Total
Corporate judgment in 2008	Absent	176	107	97	19	399
	Present	4	76	60	23	163
	Operational	0	1	30	12	43
	Secured	0	1	2	7	10
		180	185	189	61	615

False-negative judgments (1%)
False-positive judgments (52%)

This includes, for example, the actual judgments "operational" when the corporate judgment should have been "absent" or "present." The results show that 34% of the actual judgments in the category "operational" and "secured" are false-positive, and should have actually been the corporate judgment "absent" or "present." In these cases, incentives to improve the quality of care were not introduced even though they should have been.

Further analysis showed that false-positive judgments are not related to individual inspectors (x=51.1, median=52.8, Sd=19.3). Figure 1 shows the percentage of false-positive judgments for each inspector as well as the mean judgment on the four criteria for all inspectors.

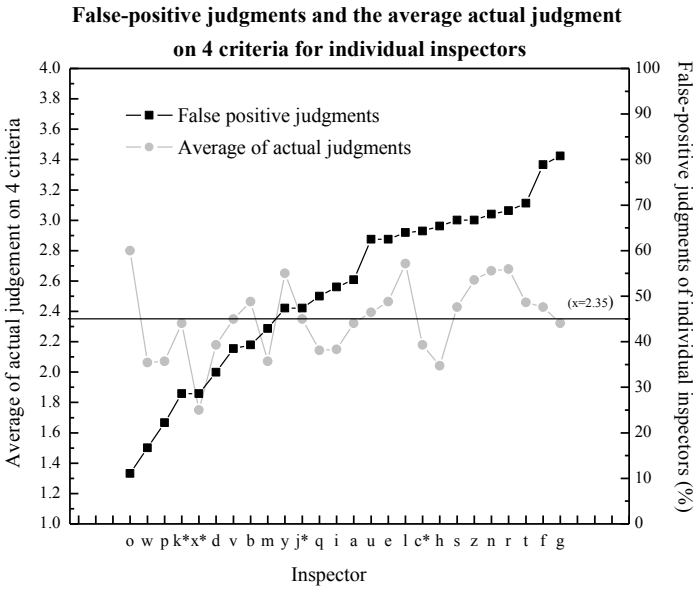


Figure 1 Percentage of false-positive judgments per inspector for the four analyzed criteria (x=51.1, median=52.8, SD=19.3) contrasted with the mean judgment for these four criteria per inspector (inspector in training indicated with *).

Although the percentage of false-positive judgments differs between inspectors (minimum=11.1, maximum=80), this is shown for all inspectors on the four criteria. Moreover, false-positive judgments are also shown for inspectors still in training. Another striking result is the lack of a correlation between the type of judgment and the percentage of false-positive judgments (Figure 1). Inspectors with relatively negative judgments have both relatively high and low percentages of false-positive judgments. The opposite can also be seen: inspectors with relatively positive judgments have high percentages of false-positive judgments as well as low percentages of false-positive judgments. Although the percentages of false-positive judgments are not inspector-dependent, these judgments do depend on the specific criteria (Figure 2).

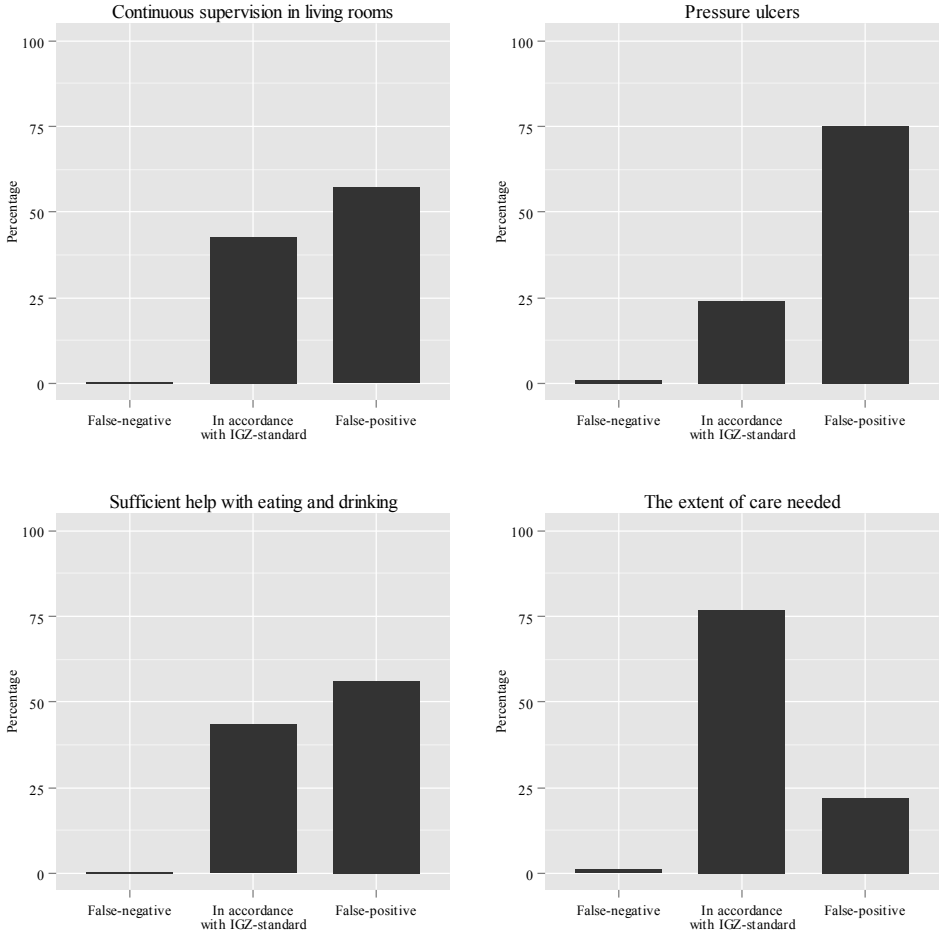


Figure 2 The percentage of judgments that correspond to the IGZ standards along with the percentage of false-negative judgments and false-positive judgments for the four analyzed criteria of the regulation of nursing home care in 2005/2006.

Figure 2 shows that the percentages are almost the same for the criteria “permanent supervision in living rooms” (57%) and “sufficient help with eating and drinking” (56%). Conversely, the percentage for the criterion “pressure ulcers” is relatively large (75%), and the percentage of false-positive judgments for the criterion “the extent of care needed” is relatively small (22%).

3.7 Discussion

The results of this study indicate validity issues in the decision-making process of nursing home regulation. Similar judgments appear to have a wide variety of meanings. Furthermore, the grounds for judgments do not always correspond to the corporate standards. This implies that in 53% of the cases, the actual judgments are inaccurate. This inaccuracy is characterized mainly by false-positive judgments. In general, inspectors assign scores that are too positive compared to the corporate standards. There was no correlation between the mean judgments on the four criteria per inspector and the percentage of false-positive judgments. Whether inspectors were still in training during the period their visits took place had no influence on the presence of false-positive judgments, and trainee inspectors also assigned such judgments. The results of this study indicate that the phenomenon of false-positive judgments is a feature associated with inspectors, because false-positive judgments were found for all inspectors included in this study. In contrast, the occurrence of false-positive judgments varies between criteria. Because of this, false-positive judgments seem to be related to the quality of the corporate standards. When standards are more ambiguous, inspectors have greater opportunities to individualize their decision making.

3.8 Explanations concerning the regulatory instrument

IGZ inspectors use an official document (which includes the corporate standards and the corresponding criteria) to examine and judge the quality of care in nursing homes. This instrument operates as a guideline to promote uniformity in regulatory judging processes. But even the best instrument will not be able to capture all of the situations observed in health care institutions. The professional skills of experts are needed to reach a professional judgment in cases where a gap exists between the instrument and reality. For example, if a nursing home does not record the presence of pressure ulcers for its residents, the judgment “absent” applies to this situation according to the corporate standards. If a nursing home does record the presence of pressure ulcers and also has a protocol for pressure ulcers, the judgment “present” applies when the corporate standards are used. If the nursing home only has a protocol for pressure ulcers and does not record the presence of pressure ulcers, the IGZ

corporate standard offers no appropriate judgment: neither the judgment "absent" nor the judgment "present" applies to this situation. In the corporate standards, there is a gap between reality and the categories used. Some inspectors find the judgment "absent" most appropriate, while others would rather choose for "present." A gap exists because the instrument does not fully reflect reality, and so inspectors may give health care institutions the benefit of the doubt, which results in judgments that are too positive. But the correspondence between the corporate standards and reality is not the only explanation for the results of this study. The extent to which the standards are unambiguous is also a possible explanation for the validity problems. Validity issues are more likely to occur when descriptions of the corporate standards and criteria are more ambiguous, because the criteria are employed in different manners. Both the interrater reliability of the two independent observers in their analyses of the arguments presented in the grounds as well as the percentage of false-positive judgments indicate the extent to which the IGZ corporate standards are unambiguous. High interrater reliability and low percentages of false-positive judgments indicate unambiguous standards.

This relationship between reliability and false-positive judgments was confirmed for the criteria "pressure ulcers," "continuous supervision in living rooms," and "sufficient help with eating and drinking." Two of the criteria ("continuous supervision in living rooms" and "sufficient help with eating and drinking") were characterized by relatively high percentages of agreement between the observers of this study (85% versus 93%) and relatively low percentages of false-positive judgments (57% versus 56%). Both of the observers found that these criteria seemed to be employed in a similar manner, which means they agreed with each other to a large extent. Moreover, for these two criteria, the percentage of false-positive judgments was relatively low.

This indicates that these criteria are relatively unambiguous. However, the opposite applies for the criterion "pressure ulcers," which is characterized by a relatively low percentage of agreement between the observers of this study (64%) and a relatively high percentage of false-positive judgments (75%). This indicates that this criterion is relatively ambiguous. These percentages confirm that the relationship between the level of interrater agreement and the percentage of false-positive judgments is an indication of the degree to which a criterion is unambiguous. This relationship was not confirmed for the criterion "the extent of care needed." This criterion is characterized by a relatively low percentage of interrater agreement as well as by a relatively low percentage of false-positive judgments rather than a relatively high percentage of false-positive judgments. This outcome might be explained by the way the independent observers reached consensus about differences in the judgment of the arguments presented in the grounds that accompanied judgments for this criterion. If the grounds stated that "measurements are performed to determine the extent of care needed," this was scored as

"the extent of care needed is recorded systematically." The observers defined "systematically" as "using a system to measure the extent of care needed." This definition was determined by the way inspectors use the standard in practice. With this definition, the percentage of false-positives is relatively low.

3.9 Explanations concerning characteristics of inspectors

Along with the quality of the regulatory instrument, the compliance of inspectors when using the instrument is also a significant factor in explaining the results of this study. If compliance with the instrument is limited, it is likely that inspectors are utilizing the instrument in an individual, inspector-dependent manner, which results in judgments characterized by an inspector-dependent meaning. Moreover, an inspector's personal frame of reference can also provide part of the explanation. When aspects in the quality of care that are important to an inspector are not represented in the regulatory instrument, this can influence the inspector's regulatory decision making. If, for example, inspectors consider hygiene to be an important criterion for the quality of care in nursing homes, they might include this criterion in their judgment even though this criterion is not part of the IGZ standards. In such cases, their judgment on hygiene has influenced their actual judgment on another criterion. It is a well-known phenomenon that experts use information other than the prescribed criteria in their judgment, and this can lead to the observation that professional judgments differ from corporate judgments [7,8]. The same applies if inspectors do not keep strictly to specific conventions, but let their overall impression of an institution affect their judgment. This results in what is known as the "halo effect" [8,9]. Research on teachers' judgments of evaluations shows that a positive expectation or positive impression of a student can result in good performances being included in the judgment more than less positive performances [8,9]. Also, the results of this study can be explained by types of enforcement styles [10-12]. Inspectors can employ a more deterrent approach or a more persuasive one. Validity and reliability issues can be linked to an inspector's preference for a type of enforcement style. Enforcement agencies do not always use their formal authorities. Inspectors prefer to use informal methods like advising, warning, and threatening to influence inspectees. False-positive judgments could be related to a more persuasive enforcement style [10,11,13]. In every regulatory relationship, there is a tendency to decrease the distance between the inspector and inspectees. Inspectors become advisors, partners, and confidant(e)s. After all, harmony is more pleasant than confrontation, and is possibly more effective in achieving regulatory goals [10].

In contrast to a repressive enforcement style, which is characterized by extensive controlling and sanctioning, a cooperative style is defined by persuading institutions to comply with the rules by means of consulta-

tions [11]. A cooperative enforcement style focuses on maintaining the good relationship between the inspectors and the inspectees. Drawbacks of this style are the humanness problem [14], client intimacy, capturing [15], and the shadow of the future [16]. Because of the involvement of inspectors with the inspectees, inspectors become encapsulated and are less able to judge objectively [14,15,17]. Inspectors and inspectees have a long-term relationship, and inspectors do not want to harm this relationship [16]. In a cooperative enforcement style, inspectors are more likely to assign judgments that are too positive than ones that are too negative [14].

3.10 Implications and solutions for the validity issues

The IGZ employs risk-based supervision in the regulation of health care in the Netherlands. In risk-based supervision, visits are made to a selection of institutions with suspected risks in the quality of care. During these on-site visits, inspectors examine and assign scores to, and also promote, the quality of care. In the case of negative scores, institutions are obliged to improve the quality of care, which should establish its promotion. However, when there are false-positive judgments, no measures to improve the quality of care are taken, and this might limit the effectiveness of regulation.

How can these reliability and validity issues be solved? Is the solution to develop a regulatory instrument that excludes any discretionary space for inspectors? Even if it were possible to develop such an instrument, it would not be a desirable solution for the aforementioned issues, because it does not reflect either the complex reality of health care or the professionalism of inspectors [18].

A multidimensional reality does not fit into a one-dimensional instrument [18]. Instruments with a unilateral accent on aspects that can be measured can result in an undesirable situation: political as well as normative considerations are determined only by the extent to which performance is measurable [19]. Consequently, only the aspects that are measurable are emphasized (for example, the percentage of pressure ulcers) instead of quality or what is being done to prevent the prevalence of pressure ulcers. This might result in accountability becoming a way to force institutions to only record those things that are measurable. This can surely be just a meager representation of reality [20]. Reality is more complex than an instrument, and inspectors need discretionary space to assign scores in this situation. If this space is not included in the instrument, inspectors become pollsters, and health care is reduced to a one-dimensional reality: an undesirable situation [18,21]. The solution for the ascertained issues seems to be a combination of factors. First, there should be a good regulatory instrument that does justice to the complexity of care. This instrument should include an explicit framework of standards that describe exactly which judgment applies in which situation [18]. Second, it is crucial that the inspec-

tors be committed to using the instrument as well trained in how to use it [22,23]. The training should include not only how to use the instrument itself, but in particular, how to the deal with the gap between the instrument and reality. If a situation requires deviation from regulatory standards and instruments, it is important that inspectors are trained to explain their deviations.

3.11 Limitations of the study

Some critical methodological notes should be mentioned here. In this study we used only the arguments reported in the grounds given for judgments. It might be possible, though, that inspectors discussed the presence of help with eating and drinking provided by trained volunteers, but did not report this in the grounds they gave. Arguments that played a part in the regulatory judgment process but were not reported might have resulted in a discrepancy between the actual judgment and the corporate standards. Therefore, the validity issues shown in this study might be explained to some extent by inaccurate reporting. This is not likely, though, because the results show that all inspectors have a tendency to judge inaccurately in the same manner: to make false-positive judgments. Moreover, nearly a quarter of the false-positive judgments are two or more categories higher (more positive) than the corporate judgment. Therefore, it seems more likely that the differences between the actual and the corporate judgments can be explained by inaccurate judgment rather than by inaccurate reporting.

3.12 Future research

It is as yet unclear which other factors might explain the reliability and validity issues of IGZ inspectors. Earlier research has shown that regulatory styles influence the judgment process in regulation [24-26]. Research on the interrater reliability of physicians has shown that organizational circumstances influence this [27]. We will examine which factors influence the interrater reliability of IGZ inspectors in addition to regulatory styles, characteristics of inspectors, and the quality of the regulatory instrument. We will conduct a systematic review to examine which interventions are effective in reducing interrater disagreement. Subsequently, we will study the effect of an intervention to reduce interrater disagreement in regulation.

3.13 Acknowledgments

We would like to express our thanks to Ms. K.N. Middendorp-Kolenbrander (IGZ), who participated in the research as one of the independent observers. We would also like to thank Dr. C.A.J. Ketelaars (IGZ), E. van Ankum (IGZ), and P. van Dyk (IGZ) for their valuable comments.

3.14 References

1 Tuijn SM, Janssens FJG, Van den Bergh H, Robben PBM. "Not all judgments are the same: a quantitative analysis of the interrater reliability of inspectors at the Dutch Health Care Inspectorate" [Het ene oordeel is het andere niet: Interinspecteursvariatie bij inspecteurs van de IGZ: een kwantitatieve analyse; in Dutch]. Nederlands Tijdschrift voor Geneeskunde 2009(8):322:326.

2 IGZ. "Policy Plan 2012-2015. For justified trust in safe and appropriate care II" [Meerjaren beleidsplan 2012-2015. Voor gerechtvaardigd vertrouwen in verantwoorde zorg; in Dutch]. 2011.

3 Uebersax JS. Validity inferences from observer agreement. Psychological Bulletin 1988;104(3):405-416.

4 Nijveldt MJ. Validity in Teacher Assessment. An exploration of the judgment processes of assessors. Enschede: Gildeprint; 2007.

5 Mehrens W, Lehmann I. Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart and Winston; 1973.

6 Bergh van den H, Zwarts M, Peter-Sips M. "Quality of the educational learning process" [Kwaliteit van het onderwijsleerproces; in Dutch]. Tijdschrift voor Onderwijsresearch 2000;25(1/2):20-39.

7 Gaeth GJ, Shanteau J. Reducing the influence of Irrelevant Information on Experienced Decision Makers. Organizational behavior and human performance 1984;33:263-282.

8 Meuffels B, Ansink M, Donselaar v J. "A unique halo-effect on the judgment of examinations" [Een bijzonder halo-effect bij het beoordelen van opstellen; in Dutch]. Tijdschrift voor Taalbeheersing 1986;8(3):177-193.

9 Zillig M. "Attitude and Evidence" [Einstellung und Aussage; in German]. Zeitschrift fur die Psychologie 1928;106:58-106.

10 Ridder de J. "A good advice for regulation" [Een goede raad voor toezicht; in Dutch]. Den Haag: Boom Juridische uitgevers; 2004.

11 Huisman W. "Between benefit and moral" [Tussen winst en moraal; in Dutch]. Den Haag: Boom Juridische Uitgevers; 2001.

12 Day P, Klein R. The regulation of nursing homes. The Milbank Quarterly 1987;65(3):303-347.

13 Gaasterland D, Blackwell B. The advanced glaucoma intervention study (AGIS):10. Variability among academic glaucoma subspecialists in assessing optic disc notching. Tr Am Opthth Soc 2001;99:177-186.

14 Albanese MA. Challenges in using rater judgments in medical education. Journal of Evalution in Clinical Practice 2000;6(3):305-319.

15 De Bruijn H, Heuvelhof ten E. "Regulation. The game between inspector and regulatee" [Handhaving. Het spel tussen inspecteur en inspectee; in Dutch]. Utrecht: Lemma; 2005.

16 Leeuw F, editor. "Behavioral mechanisms behind governmental interventions and rules of justice" [Gedragsmechanismen achter overheidsinterventies en rechtsregels; in Dutch]. ; 2008; Maastricht, The Netherlands: Oce business services; 2008.

17 Stoop A, Berg M. "Mirrors for the physician" [Spiegels voor de arts; in Dutch]. Gezondheid: theorie en praktijk 1994:265-278.

18 Janssens FJG. "From research to evaluation. The methodology of the Dutch Inspectorate of Education" [Van onderzoek naar evaluatie. De methodologie van de Onderwijsinspectie; in Dutch]. ; 1997.

19 Scientific Council for Government Policy (WRR). "Evidence of Good Service" [Bewijzen van goede dienstverlening; in Dutch]. 2004:p.198.

20 Tonkens EH. "Articulate citizens, domesticated professionals. Free market system, steering of demand and professionalism in the public sector" [Mondige burgers, getemde professionals. Marktwerking, vraagsturing en professionaliteit in de publieke sector; in Dutch]. Utrecht: Nizw; 2003.

21 Weiss DJ, Shanteau J. The vice of consensus and the virtue of consistency. New York: Cambridge University Press; 2004. p. 226.

22 Francke AL, Smit MC, Veer de AJE, Mistiaen P. Factors influencing the implementation of clinical guidelines for health care professionals: A systematic meta-review. BMC Medical Informatics and Decision Making 2008(38):1-11.

23 Ploeg J, Davids B, Edwards N, Gifford W, Miller PE. Factors Influencing Best Practice Guideline Implementation: Lessons Learned from Administrators, Nursing Staff and Project Leaders. World views on Evidence-Based Nursing 2007(4):210:219.

24 Gray C, Gardner J. The impact of schoolinspections. Oxfort Review of Education. 1999;25:455-469.

25 Ehren MCM. "Regulation and improvement of education" [Toezicht en schoolverbetering; in Dutch]. Delft: Eburon; 2006.

26 Hutschemaekers G, Kist S. "The profession of inspector at the Dutch Healthcare Inspectorate" [Beroep Inspecteur; in Dutch]. 2006.

27 Jong de J. Explaining medical practice variation. Ipskamp, Enschede: NIVEL; 2008.

Chapter 4

Evaluating instruments for regulation of health care in the Netherlands

This chapter is published as:

Tuijn SM, Robben PBM, Janssens FJG, Van den Bergh H. (2011) Evaluating instruments for regulation of health care in the Netherlands. *Journal of Evaluation in Clinical Practice,* 17, 411-419.

4.1 Abstract

Reliable and valid judgments are necessary for regulatory authorities to merit confidence from care institutions and society and preserve authority. Moreover, limited reliability and validity of regulatory judgments increase the risk of limited improvement of the quality of health care. The goal of the study is to obtain insight in (dis) advantages of different regulatory instruments for regulation of health care.

In this study, the reliability and validity of judgments generated by a lightly structured and highly structured regulatory instrument used by the Dutch Health Care Inspectorate are compared.

Results indicate that the lightly structured instrument causes a large variety in discussed topics in regulatory visits: indicators pointing out potential risks in care are not always part of these discussions, by which incentives to improve care remain unjustly undone. Both types of instruments show variations in the meaning of judgments, indicating validity problems.

The results of our study suggest that regulation of health care requires thorough appraisal of instruments. Several requirements are identified: first, an instrument that justifies the complexity of care with an accompanying explicit set of standards is necessary.

Second, commitment of inspectors to the instrument is essential. And third, training of inspectors is indispensable.

4.2 Introduction

Quality and continuous improvements have become a natural part of the conversation and activities of health services [1]. Regulation of health care also aims to stimulate quality improvement. There is an international trend towards the greater use of government regulation in health care [2]. The effects of regulation on the quality of health care are internationally heavily discussed and sometimes criticized [3-5]. Scientific research on the effects of regulation is relatively young and focuses generally on risk regulation regimes [6], the effects of enforcement and surveyor styles [1,7-10]. Studies on the quality of regulatory instruments and decision making in regulation are limited. As accreditation, regulation of health care is an important reliability judgment area to investigate given the investment levels and prevalence of regulation [11-16].

Regulators in supervision of health care essentially have three key objectives: improvement, assurance and accountability [17]. This type of accountability refers to the process of making health care organizations and the professionals who work within those organizations more directly accountable to patients and the general public. Besides these primary goals, regulators can also have principles to pursue effective regulation. One of

these principles is openness and transparency. Since public disclosure of performance data becomes common practice of regulators, information about (the design of) regulatory processes as well as results of regulation (including the findings and judgments about individual organizations) are liberally available. However, openness does not automatically flow over in accountability of the regulator. The aspect of accountability where we will refer to in this article, is the mechanism for holding the regulator accountable for its actions to those with an interest in the area being regulated. In the case of reliability and validity issues, accountability can be hampered. After all, it can be difficult for a regulator to explain why organizations with similar circumstances are judged variously. It is argued that in the policy of regulators, the major precondition for fair and transparent regulation is to work with an explicit set of standards [17]. But what exactly do we know about experiences with more or less explicit standards or instruments? To gain more insight in different regulatory instruments, a highly structured instrument (HSI) and lightly structured instrument (LSI) used in regulation of health care in the Netherlands will be empirically examined. This study offers the possibility to provide chances for further professionalism of regulation.

4.3 Regulation of health care in the Netherlands

In the Netherlands, regulation of health care is performed by the Dutch Health Care Inspectorate (IGZ). IGZ is an independent agency within The Ministry of Health, Welfare and Sport. With regards to government involvement in health care, the Dutch system can be positioned, on an international scale, somewhere between countries with a national health care system, such as the

UK, and countries where the market approach in organizing and financing health care is dominant, as in the USA [18]. IGZ guards the quality of care and enforces 25 laws, for example the Care Institutions Quality Act [19]. The policy of IGZ is to aim for standardized procedures and reliable and valid judgments to stimulate the quality of care and to justify her regulatory decisions and activities. Regulators need methods to measure and monitor the performance of the organizations they regulate: a process described as ‘detection’[17]. Detection can include regular inspections, responding to complaints with focused investigations and monitoring performance on a continuing basis. This last form can be realized by collecting, aggregating, analyzing and comparing performance data of regulated organizations. Like in countries such as Australia, the USA, Switzerland, Sweden and Norway, quality indicators were introduced in the Netherlands to monitor and stimulate the quality of health care [20-25].

Regulation of health care in the Netherlands is performed by IGZ by a combination of three methods of detection. First, theme based regulation is directed at specific issues in care, sometimes asked for by the minister or parliament, that require the attention of the regulator. Second, IGZ deploys regulation in response to calamities in the event of emergencies that indicate structural shortcomings in care provision. Third, IGZ has been developing risk-based supervision from 2002 to assess the quality of health care by means of indicators [26]. In risk based supervision, a framework for the quality of care and accompanying sets of quality indicators are drawn up, in cooperation with representatives of the health care sector. Subsequently, risk based supervision consists of three phases:

Table 1 Survey of the differences in application of risk based supervision on hospitals and nursing homes in the Netherlands

Characteristic	Analysis of the second phase of risk based supervision on hospitals	Analysis of the second phase of risk based supervision on nursing homes
Scope	Each of the hundred hospitals is visited once a year.	Inspectors visit a limited number of 2000 institutions. Institutions suspected of a risk are visited, as well as a random sample of institutions without any risk suspicions. The latter are visited in order to reveal 'best practices'.
The visit	During their visits, inspectors always have interviews with a member of the Executive Board, the Chairman of the Medical Council and often, the Quality Officer. The care providers whose work is related to the indicators discussed are usually also present. Inspectors do not tour the hospital, nor do they have interviews with patients or examine patient files. During this visit, data on complaints, information from incident supervision, theme-based supervision and data on indicators are discussed.	Inspectors visit different departments and interview the management of the institution, the nursing home practitioner and other care providers, the client council and sometimes the residents. Inspectors also study patient files and other documents.
Instrument	A set of 20 indicators is used which are all directed primarily at the hospital's care processes. The instrument consists of a list of signals based on the indicators. This list shows the relevant hospital's present score per indicator and scores of previous years, and the national average of Dutch hospitals per indicator. Inspectors decide which indicators will be discussed, but in principle indicators are discussed and assessed by inspectors in case of a 'signal'.	Inspectors use an instrument with defined criteria. In 2005/2006, this instrument consisted of 27 criteria. All of these criteria were assessed during regulatory visits.
Degree of explicitly of standards	Lightly structured instrument: Inspectors judge on the basis of the signal given by the indicator and the hospital statement regarding this signal. Inspectors assess indicators during annual visits in terms of a 3-point scale: 'no improvements necessary', 'minor improvements necessary' and 'adjustments needed'. Inspectors can also issue a finding of 'no judgment possible yet'. In that case, additional information from the hospital is requested before the indicator is judged.	Highly structured instrument: The instrument serves as a set of standards in which the criteria and the accompanying IGZ standards are defined. Inspectors judge these criteria on a four-point scale: 'absent', 'present', 'operational' and 'fulfilled'. This scale runs from a negative to a positive score, with 'fulfilled' being the most positive.
	No explicit set of standards is present stating which judgment applies in which case. Therefore, the statement that a hospital issues in response to a signal is vital in the judgment process.	This instrument includes a set of standards that defines exactly which judgment applies for which situation.

- IGZ analyses the data collected with the indicators and selects institutions at risk.
- Inspectors visit the selected institutions. Institutions are obliged to improve their care on the basis of the inspectors' judgment. Inspectors can decide to plan a follow-up visit if those improvements are not satisfactory.
- If the improvements are not satisfactory, IGZ can impose administrative sanctions and initiate penal measures.

In this manner IGZ has found a way to use the indicators in regulatory detection and enforcement, apart from publishing the indicators. In this manner, continuous quality improvement and a modern method of quality measurement by indicators are combined in a regulatory format [20]. However, the application of risk based supervision differs in the various health care sectors as can be seen for hospitals and nursing homes in Table 1.

One of the most obvious differences between risk-based supervision on hospitals and nursing homes is the use of a HSI in supervision on nursing homes versus the use of a LSI in supervision on hospitals. The HSI consisted of 27 criteria in 2005/2006. These criteria are presented in Table 2.

Table 2 The 27 criteria for safe and solid care used in supervision on nursing homes in 2005/2006.

Criteria		
Measurement of intensity of care	Procedure for complaints	Competences conform the law for professions in healthcare (BIG)
Pressure ulcers	Working systematically with care plans	Acting professionally
Commission for reporting incidents	Registration of deviations of individual care plans	Treatment of clients
Safety of materials	Procedure for responsibility for content and coordination of care plans	Long term policy
Satisfaction of clients	Procedure for multidisciplinary meetings to discuss clients	Annual work
Client board	Rights of clients in connection with care plans	Organizational structure
Information for clients	Education in using care plans	System for management information
Sufficient help with eating and drinking	Availability and competence of care providers	System to monitor a responsible level of care
Permanent supervision in living rooms	Educational plan	Privacy

These criteria are a combination of measures on outcomes, processes and structures. For example, if the criterion 'pressure ulcers' is judged, the prevalence of pressure ulcers is registered (outcome indicator) as well as the presence of a protocol of pressure ulcers (structure indicator). In regulatory visits in 2005/2006, those criteria were judged on a 4-point scale: 'absent', 'present', 'operational' and 'secured'. This scale runs from a negative to a positive view, with 'secured' being the most positive. This instrument includes a set of standards that define

exactly which judgment applies in what situation. Table 3 presents the standards for 4 of the criteria for regulation of nursing homes in 2005/2006. For the criterion 'pressure ulcers' the judgment 'absent' applies in the case that a nursing home does not register the presence of pressure ulcers.

In contrast to the HSI by which all criteria are discussed in nursing homes, topics of the LSI, which consist of 20 indicators, are only discussed in case of a signal. This is the case if the relevant hospital's score on an indicator is an outlier and too far from the national average to be caused by chance alone (p10, p90), an inexplicable trend or major changes in scores can be seen (for example, if the percentage of cancelled operations differs notably from preceding years) or the hospital has not provided the data, while more than 80% of hospitals have done so. The policy of IGZ is that, the specific health care sectors represented within IGZ develop their own regulatory instruments. Because of this, it is possible that the regulatory instrument used for regulation of nursing homes differs from the instrument used for regulation of hospital care. In this empirical study we will assess the effect of these types of instruments on the quality of regulatory output of the regulator (IGZ). First we investigate the effect of the type of instrument on the variation of discussed subjects during regulatory visits. Subsequently, we assess the reliability and validity of inspectors' judgments issued with a HSI and LSI. We start by the methodological section, and then we will present our findings, discuss their implications, their generalizability and draw conclusions.

4.4 Methods

Various data sets have been analysed. To analyse regulation with a HSI, we used data on risk based supervision of nursing homes in 2005/2006. This involved the data used for the study investigating interrater reliability of inspectors [27]. These data contain the judgments of inspectors (n = 26) who had written at least seven reports on their inspection visits to nursing homes. Only reports that complied with the IGZ format were included. This format implies that inspectors substantiated and reported their judgments in the tables of the format. Seven reports were randomly selected for each inspector. In total 182 reports were analysed.

For the analysis of the LSI, we used data on risk based supervision of hospitals in 2005, 2006 and 2007. Out of the available reports (n = 107) of regulatory visits in that period, reports by inspectors (n = 11) who had written at least four reports on different hospitals were analysed. Only reports on general hospitals (including academic hospitals) were included (n = 71). The use of indicators to discuss the quality of care in regulatory visits was analysed for the purpose of this study.

Table 3 Overview of the four analysed criteria and corresponding IGZ-standards that were used for supervision on nursing homes in 2005/2006.

Criterion	Definition	Absent	Present	Operational	Fulfilled
The extent of care needed	The institution employs a system to register the intensity of care.	The institution has no notion of the intensity of care and/or the intensity of care is not registered in a systematic manner.	The intensity of care is registered in a systematic manner.	The data of the intensity of care are used for management purposes.	Both the data of the intensity of care are used for management purposes as well as the system to register the intensity of care is systematically evaluated, and adjusted if necessary.
Pressure ulcers	The institution employs a protocol for prevention, treatment and registration of pressure ulcers. This protocol meets the quality requirements of CBO. Both the systematic of measurement of pressure ulcers as well as the protocol are evaluated in a systematic manner and adjusted if necessary.	Pressure ulcers are not registered.	Pressure ulcers are registered and a protocol for pressure ulcers is present in the institution.	Pressure ulcers are registered and measurements of pressure ulcers take place. Te protocol is known by the employees and training in the use of the protocol takes place.	Both the systematic of measurement of pressure ulcers as well as the protocol are evaluated in a systematic manner, and adjusted if necessary.
Continuous supervision in living rooms	An instructed person or employee is permanently present to guarantee supervision on psycho geriatric clients in the living rooms.	No guideline is available for the adequate supervision on psycho geriatric clients in the living rooms by employees or instructed persons.	Guidelines are available. And a guideline for instruction to persons who will supervise psycho geriatric clients is present. This guideline is not known by employees or deviation from this guideline takes place in the majority of times (more than 10% of the times). Or instruction of volunteers/relative of clients who supervise psycho geriatric clients in living rooms does not take place.	Guidelines are available. And a guideline for instruction to persons who will supervise psycho geriatric clients is present. This guideline is known by employees or deviation from this guideline takes place in the minority of times (less than 10% of the times). Or instruction of volunteers/relative of clients who supervise psycho geriatric clients in living rooms does take place.	Guidelines and the employment of the guidelines are periodically evaluated, and adjusted if necessary.
Sufficient help with eating and drinking	Sufficient employees/relatives of clients/ volunteers who are instructed are available to help with food and drink of the clients.	No guideline is available about the presence of sufficient employees/relatives of clients/ volunteers who are instructed to help with food and drink of clients.	Guidelines are available about the presence of sufficient employees/relatives of clients/ volunteers who are instructed to help with food and drink of clients. These guidelines are not known by employees or deviation from this guideline takes place in the majority of times (more than 10% of the times). Or instruction of volunteers/relative of clients who help clients with food and drink does not take place.	Guidelines are available about the presence of sufficient employees/relatives of clients/ volunteers who are instructed to help with food and drink of clients. These guidelines are known by employees or deviation from this guideline does not take place or takes place in the minority of times (10% or less than 10% of the times).	Guidelines and the employment of the guidelines about the presence of sufficient employees/relatives of clients/ volunteers who are instructed to help with food and drink of clients are periodically evaluated, and adjusted if necessary.

4.5 Analysis

Descriptive statistics (SPSS, 15) were used to study how many criteria and indicators were discussed during each regulatory visit. In order to analyse the correlation between the presence of a signal on hospital indicators and the discussion of indicators, the data on the 10 most frequently discussed indicators in 2007 were selected: 'care ICT, unplanned re-operations, high-risk interventions, intensive care, pressure ulcers, safety of medication, pregnancy, cancelled operations, diabetes and acute myocardial infarction'. Using logistical regression (SPSS, 15) the existence of an interaction effect between the presence of a signal and the discussion of indicators was investigated. We also analysed (aspects of) the validity of judgments with a HSI and LSI and investigated whether false positive or false negative judgments could be shown. For the judgments given with the HSI, two observers (working independently) analysed the arguments from the IGZ standards used by the inspectors in order to found their judgment (actual judgment). This analysis was performed on the judgments on the four criteria: 'pressure ulcers', 'sufficient assistance with food and drink', 'permanent regulation of living rooms' and 'measurement of intensity of care'. The first observer (the first author of this article) has survey experience but no experience as an inspector. The second observer has worked as an inspector for the regulation of nursing home care. This observer was familiar with the criteria used and the accompanying IGZ standards. The observers determined which judgment the inspector should have given if the IGZ standard had been strictly applied (prescribed judgment). This allows for comparison of the actual judgment given by the inspectors and the prescribed judgment that should be assigned on the basis of the arguments and the IGZ standards. The interrater reliability of inspectors was also investigated for the LSI. This was realized by analysing whether different judgments were given for similar indicators, with similar signals and similar hospital statements.

4.6 Results

The average number of indicators discussed by inspectors with the LSI varies widely between inspectors; between 5.5 and 10. In the period from 2005 to year-end 2007, an average of seven of the 20 indicators were discussed (M=7.0, $\bar{x}$ = 7.2, Sd = 1.41). In contrast to the LSI, there is little difference in the average number of criteria discussed per inspector with a HIS. The average number of criteria discussed with a HSI nearly approaches 27 (M=25.8, $\bar{x}$ = 25.7, Sd =.5). The non-discussed criteria are generally the same ones: 'privacy' was discussed in 30% of the visits; 'care plan training' was discussed in 67% of the visits. The variation in the aver-

age number of criteria discussed for the LSI (Sd=1.41), is larger compared to variation in the average number of discussed criteria with the HSI (Sd=0.5).

In addition to the variation in the number of indicators discussed with a LSI, the type of indicator discussed varied as well. In 2007, 10 of the 20 indicators for the LSI were discussed most frequently. These indicators were discussed in at least 42% of the hospital visits in 2007 (n = 52). Table 4 presents the lack of correlation between the 10 most frequently discussed indicators and the presence of a signal for these indicators.

Table 4 Cross table in which the discussion of 10 most discussed hospital indicators (care ICT, unplanned re-operations, high-risk interventions, intensive care, pressure ulcers, medication safety, pregnancy, cancelled operations, diabetes and acute myocardial infarction) in a lightly-structured instrument are related to the presence of signals (n = 520).

	Not discussed at annual meeting	Discussed at annual meeting	Total
No signal	31% (161)	27% (139)	58% (300)
Signal*	19% (99)	23% (121)	42% (220)
Total	50% (260)	50% (260)	100% (520)

* A signal occurs if a hospital's score is well above or below the national average (e.g. p90 or p10).

A two-way logistical regression analysis showed a significant difference between the frequency of indicators with signals and those without (χ^2 = 12.60, df = 1, p<0.01). There were more indicators discussed without a signal than with a signal. The differences in the frequency with which indicators were or were not discussed was not significant (χ^2 = 0.00, df. = 1, p=1.00). Results showed no significant interaction effect between the presence of a signal and the discussion of indicators (χ^2 = 3.80, df. = 1, p=0.051). In other words: there was no relationship between the presence of a signal and the discussion of indicators. The indicators discussed proved to be completely dependent on inspectors.

4.7 Validity and reliability of judgments

The meaning of the inspectors' judgments with a LSI varies to a great extent. For example, the signal on the indicator 'medication safety' was the fact that 'outpatient and extra-mural data were not digitally available'. In this case the judgment 'no improvements necessary' had been applied by several inspectors. The meaning of this judgment varied in the reports of the inspectors as presented in table 5. In contrast to the wide diversity of meanings of judgments with the LSI, the distribution of the type of judgments given with the LSI was less diverse. The judgment most frequently given to the 10 most frequently discussed indicators is 'no improvements necessary' (45%). Given less often are judgments 'further information needed' (11%), 'minor improvements necessary' (2%) and 'adjustments needed' (3%).

Table 5 Meaning of inspectors' judgment 'No improvements necessary' for the indicator safety of medication in the observed hospital situation that outpatient and extramural data were not digitally available.

Meaning of the score "no improvements necessary" on the indicator safety of medication in the observed situation that outpatient and extra-mural data were not digitally available.
'A pilot project with digital prescription is in progress'
'In the clinic we prescribe digitally, but expansion to primary care is difficult'
'In the clinic we prescribe digitally, but expansion to primary care is difficult'
'We conduct a project via "Better Faster" and integrating the results in the "Safe Reporting System"
'Safe Incident Reporting does not yet take place throughout the hospital but we do have a conventional reporting committee'
'We will be opening an outpatient dispensary but a digital medication system integrated with primary care is difficult'

In 36% of cases, no judgment at all is reported. In 3% of cases, a judgment is given that does not fit in the categories formulated by the IGZ. Because the LSI does not have an explicit set of standards, it was not possible to analyze whether any judgments were excessively positive (false positives) or excessively negative (false negatives).

Analysis of the HSI, in which an explicit set of standards is used, shows that similar judgments have many different meanings. Table 6 shows that the judgment 'Operational' given by inspectors actually means 'operational' in 16% of the cases, but in 51% of the cases means 'absent', in 32% 'present' and in 1% 'secured'.

Table 6 Judgments prescribed by the two independent observers on the basis of reported arguments for judgments of nursing home care and strict application of IGZ standards, compared with actual inspectors' judgment on nursing home care in 2005/2006.

		Actual inspectors' judgment 2005/2006			
		Absent	Present	Operational	Fulfilled
Judgments strictly in accordance with the IGZ standards of 2005/2006	Absent	98%	58%	51%	31%
	Present	2%	41%	32%	38%
	Operational	0%	0.5%	16%	20%
	Fulfilled	0%	0.5%	1%	11%
		180	185	189	61

In cases where the actual judgment 'operational' should have been 'absent' or 'present' according to the IGZ standards, a false-positive judgment occurs. In a situation where an actual judgment 'operational' should have been 'secured', a false negative arises. Overall however, the data show that false positive judgments appear much more frequently than false negative judgments.

4.8 Discussion

From our study we can conclude that with the use of a LSI, inspectors select the indicators to be discussed at their own discretion [10,28]. Indicators indicate a potential risk in provided care are not always discussed. In fact, every hospital is monitored with a different version of the same instrument. Because of this, it is difficult to justify why some institutions have to improve and others don't. This result could be explained by a number of factors. Firstly, earlier studies show that not all inspectors are familiar with the principles of risk-based supervision to the same degree. A perceived loss of depth in regulation as a result of standardization of the work appears to interfere with the implementation of risk based supervision [29]. Secondly, inspectors have different views regarding their role in regulation: some inspectors consider themselves advisors, others regard themselves as examiners. Thirdly, inspectors consider risk based supervision a supplement to the other sources of information and to their own knowledge of the institution(s) concerned. As in accreditation, inspectors use multiple methods by which to triangulate [30]. This might explain why they decide which indicators will be discussed according to their own views and independently of the existence of signals indicating potential risks. Earlier studies confirm that inspector's focus of attention differs as well [10]. Another remarkable result is the inspectors' tendency towards positive judgments over negative ones. This result could fit in Berwicks' theory of continuous improvement: the modern quality expert cares far more about learning and cooperating with the conventional worker than about censoring the truly deficient [31]. It could also be explained as a cooperative regulatory style [10].

Results indicate that with a HSI identical criteria were discussed at all institutions. However, with the HSI, the argumentation behind given judgments did not always correspond to the standard. In 52% of the judgments in terms of the criteria studied, the opinions proved to be more positive than was justified by the set of standards. This gives rise to a validity problem: the judgment criteria do not measure what they are designed to measure [32-34]. Consequently, it is hard to justify why some institutions have to improve care on certain criteria, while others don't. The same applies for the LSI: the meaning of judgments of the LSI varied widely as well. Because of the absence of an explicit set of standards with the LSI, inspectors have no guidelines regarding the margins within which each judgment applies. Results show that the most positive judgment ('no improvement necessary') is most frequently given with the LSI. Because of the lack of an explicit set of standards and the many different outcomes for similar indicators combined with the diversity of hospital explanations for these outcomes, the support for the views could not be compared with a standard. Due to this neither the percentage of false negative of false positive judgments, nor interrater reliability could be calculated. After all, in

order to make statements about the reliability, a comparable situation must be assessed by different inspectors. However, it is unlikely that, in the absence of explicit standards, such as with a LSI, meaning of judgments vary less compared to a HSI.

But how does the choice of types of regulatory instruments relate to the specific health care sector and the professional background of inspectors? The policy of IGZ on developing instruments is, that each health care sector represented in IGZ is allowed to develop their own instrument. On the introduction of risk based supervision in the IGZ in 2002, inspectors supervising nursing homes developed the HSI for this sector, while inspectors in hospital regulation developed the LSI. These developments appear to be based on both a match with the education and preceding working experience of the inspectors and on the complexity of the care institutions that they supervise. Inspectors supervising hospitals often have prior medical training and work experience in curative care, while inspectors supervising nursing homes often have prior training in nursing and work experience in facilities of long term care [29]. Nursing staff are trained more to work with protocols compared to physicians [35]. Furthermore, the range of care processes in hospitals varies more widely and the processes are more complex than the care processes in nursing homes. This complexity is recognized by the fact that different topics are discussed at regulatory visits, of which the indicators form only a (limited) part. The choice of the LSI for regulation of hospital care and of the HSI for regulation of care in nursing homes could be related to this.

4.9 Limitations of the study

We analyzed the use of two types of regulatory instruments in relation to reliability and validity. Both instruments are employed in different health care sectors. It would have been better to compare two types of instruments within the same sectors. However, this situation does not occur within risk based supervision. In interpreting the results, we therefore took in account explanatory factors specific to the regulatory field whenever possible.

4.10 Conclusions

Results indicate that in the presence of explicit standards, the meaning of similar judgments vary widely. Because of the lack of an explicit set of standards with the LSI, no statements could be made regarding the extent to which inspectors issued judgments in accordance with the set of standards, nor could interrater reliability be calculated. Earlier research showed that with the HSI, interrater reliability of inspectors was poor for 14 of the 25 criteria that inspectors assessed during regulatory visits [27].

4.11 What are the implications of those results for regulation of health care?

Regulation of health care with the HSI ensures that institutions are assessed according to the same set of criteria. In contrast to the HSI, the LSI leads to inspector-dependent discussion of indicators. This means that indicators that point to a potential risk in the delivered care are not discussed while they were supposed to have been discussed. Subsequently, enforcement arrangements are (unjustly) not requested. This implies that it is hard to justify regulatory arrangements and therefore accountability of the regulator is limited.

Compared to the HSI, the LSI has some important restrictions to pursue accountability. Firstly, because the criteria by which an institution is assessed vary from one institution to another, accountability in regulation is restricted. After all, indicators which point out a potential risk are not always discussed. Consequently some institutions have to improve and some do unjustly not have to improve the provided care on specific topics. Secondly, verifiable confidence is an important element of the regulation process [36]. Because inspectors measure different issues at different institutions and the regulatory subjects are announced preceding the regulatory visit, the extent to which this confidence can be verified is limited.

4.12 Solutions

How can increased reliability and validity be effected? Is the solution for reducing validity and reliability problems an instrument which allows inspectors no discretion? This does not appear to be the case, for the HSI also proved to have limitations: inspectors use their discretion in regulatory judgments as well. Leaving the possibility of the development of regulatory instruments which exclude discretion aside, this would not be a desirable solution. It neither justifies the complexity of judging care, nor justifies the professionalism of inspectors [37].

A multi-dimensional reality cannot be measured with a one-dimensional instrument [38]. In the case that instruments unilaterally focus on the quantifiable, one stands the chance of determining the normative considerations only by the extent that they are measurable [39]. In that way, accounting primarily becomes a way to force institutions to record measurable performance, which can be a very meager reflection of reality [40]. To put it differently, without discretion in regulation, inspectors become ‘poll-takers’ and care is reduced to a one-dimensional reality: [37,38] an undesirable situation.

The solution is offered in a combination of factors. Firstly, regulatory instruments that justify the complexity of the quality of care, with an accompanying explicit set of standards, are vital [38]. Through the use of such a set of standards, the arguments described in the standards are reflected in the foundations for judgments. A set of standards provides clear guidelines for inspectors and institutions, which makes the core of the founda-

tions of judgments consistent. As a result, judgments have a 'core meaning' which makes them easier to compare.

Secondly, the commitment of inspectors to the instruments, and the training of inspectors in the use of the instruments are essential [41,42]. This concerns training in the use of the instruments themselves, as well as dealing with the gap that exists between the instrument and reality. If deviations from the standards and the instrument occur, it is important that inspectors learn to motivate their reasons for doing so. Other professionals, such as care providers, lecturers and judges, deal with this competency as well. The types of intervention suitable and effective for increasing reliability and validity of inspectors' judgments will be investigated in more detail. Though our study focuses on regulatory instruments in the Netherlands, the outcomes have significance for health care regulators in other countries.

4.13 References

1 Greenfield D, Braithwaite J., Pawsey M. Healthcare accreditation surveyor styles typology. Int J of Health Care 2008;21:435-443.

2 Klazinga N. Re-engineering trust: the adoption and adaption of four models for external quality assurance of health care services in western European health care systems. International Journal for Quality in Health Care 2000;12:183-189.

3 Walshe K. Improvement through inspection? The development of the new Commission for Health Improvement in England and Wales. Qual. Health Care 1999(8):191-196.

4 Bevan G, Hood C. Targets, inspections, and transparency. British Medical Journal 2004;324:967-970.

5 Bevan G., Hood C. Have targets improved performance in the English NHS? British Medical Journal 2006;332:419-422.

6 Hood C, Rothstein H, Baldwin R. The Government of Risk. Understanding Risk Regulation Regimes. Oxford, New York.: Oxford University Press; 2001.

7 Day P, Klein R. The regulation of nursing homes. The Milbank Quarterly 1987;65(3):303-347.

8 Hutter BM. Variations in regulatory enforcement styles. Law and Policy 1989;2:153-174.

9 Braithwaite J, Makkai T, Braithwaite V. Regulating Aged Care. Ritualism and the New Pyramid. Cheltenham, UK: Edward Elgar; 2007.

10 Mascini P, Wijk van E. Responsive regulation at the Dutch Food and Consumer Product Safety Authority: an empirical assessment of assumptions underlying the theory. Regulation and Governance 2008;3:27-47.

11 Scrivens E. Putting continuous quality improvement into accreditation: improving approaches to quality assessment. Qual. Health Care 1997;6:212-218.

12 Wagner C, Merode van GG, Oort van M. Cost of quality management systems in long-term care organisations: an exploration. Quality Management in Health Care 2003;12:106-114.

13 Pomey MP, Contandriopoulos AP, Francois P, Bertrand D. Accreditation: a tool for organisational change in hospitals? International Journal of Health Care Quality Assurance 2004;17:113-124.

14 Zwakkenberg SPM, Croonenborg van JJ, Drevers S, Barneveld van TA. "Costs of quality indicators. A brief study on the costs of developing indicators and collecting data on indicators in hospitals" [Een beknopte studie naar de kosten van indicatorontwikkeling en het uitvragen van indicatoren in ziekenhuizen; in Dutch]. 2007.

15 Sira Consulting. "Measurement of supervisory burden for hospitals. Research on administrative burden, costs of compliance, and hospitals' perception of supervision" [Meting toezichtlasten ziekenhuizen; in Dutch]. 2007.

16 Greenfield D, Braithwaite J. Health sector accreditation research: a systematic review. Int J of Health Care 2008;20:172-183.

17 Walshe K. Regulating Healthcare: A prescription for improvement? Philadelphia: Open University Press; 2003. p. 183.

18 Lombarts MJMH, Klazinga NS. A policy analysis of the introduction and dissemination of external peer review (visitatie) as a means of professional self-regulation amongst medical specialists in The Netherlands in the period 1985-2000. Healt Policy 2001;58:191-213.

19 Embassy of the Netherlands. Health Care Quality in the Netherlands. *http://www.netherlandsembassy.org/printerfriendly.asp?articlere=AR00000249EN* 2009 last accessed 16-02-2010.

20 Brennan TA. The Role of Regulation in Quality Improvement. The Milbank Quarterly 1998;76(4):709-731.

21 Luthi JC, McClellan WM, Flanders WD, Pitts S, Burnd-Hand B. Quality of health care surveillance systems: review and implementation in the Swiss setting. Swiss Med Wkly 2002;132:461-469.

22 Shaw C. How can hospital performance be measured and monitored? 2003.

23 Kollberg B, Elg M, Lindmark J. Design and Implementation of a Performance Measurement System in Swedish Health Care Services: A multiple case study of 6 development teams. Qual Manag Health Care 2005;14:95-111.

24 Pettersen IJ, Nyland K. Management and control of public hospitals – the use of performance measures in Norwegian hospitals. A case study. International Journal of Health Planning and Management 2006;21:133-149.

25 Lugtenberg M, Westert G. "Quality of health care and decision support-information for helping individuals to select health care. An international study on initiatives" [Kwaliteit van de gezondheidszorg en keuze-informatie voor burgers: een internationale verkenning van initiatieven; in Dutch]. 2007.

26 IGZ. "Policy plan 2008-2011. For justified trust in safe and appropriate care" [Meerjaren Beleidsplan 2008-2011. Voor gerechtvaardigd vertrouwen in verantwoorde zorg; in Dutch]. 2007.

27 Tuijn SM, Janssens FJG, Van den Bergh H, Robben PBM. "Not all judgments are the same: a quantitative analysis of the interrater reliability of inspectors at the Dutch Health Care Inspectorate" [Het ene oordeel is het andere niet: Interinspecteursvariatie bij inspecteurs van de IGZ: een kwantitatieve analyse; in Dutch]. Nederlands Tijdschrift voor Geneeskunde 2009(8):322:326.

28 Bakker W, Waarden van F. Liberty around rules. Observation of regulation styles and policy realization (in Dutch). Amsterdam: Boom.; 1999.

29 Hutschemaekers G, Kist S. "The profession of inspector at the Dutch Healthcare Inspectorate" [Beroep Inspecteur; in Dutch]. 2006.

30 Hammersley M, Atkinson P. Etnography: Principles in Practice. London: Routledge; 1995.

31 Berwick DM. Continuous improvement as an ideal in healthcare. The New England Journal of Medicine 1989;320(1):53-56.

32 Mehrens W, Lehmann I. Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart and Winston; 1973.

33 Uebersax JS. Validity inferences from observer agreement. Psychological Bulletin 1988;104(3):405-416.

34 Nijveldt MJ. Validity in Teacher Assessment. An exploration of the judgment processes of assessors. Enschede: Gildeprint; 2007.

35 http://www.mednet.nl/actueel/artsen-verpleegkundigen-opener-over-incidenten-1291.html. Physicians: nursers are more transparant about incidents. 2007 last accessed 16-02-2010.

36 Official Supervision Committee II. Supervision: towards compliance for the society (in Dutch). 2003.

37 Weiss DJ, Shanteau J. The vice of consensus and the virtue of consistency. New York: Cambridge University Press; 2004. p. 226.

38 Janssens FJG. "From research to evaluation. The methodology of the Dutch Inspectorate of Education" [Van onderzoek naar evaluatie. De methodologie van de Onderwijsinspectie; in Dutch]. ; 1997.

39 Scientific Council for Government Policy (WRR). "Evidence of Good Service" [Bewijzen van goede dienstverlening; in Dutch]. 2004:p.198.

40 Tonkens EH. "Articulate citizens, domesticated professionals. Free market system, steering of demand and professionalism in the public sector" [Mondige burgers, getemde professionals. Marktwerking, vraagsturing en professionaliteit in de publieke sector; in Dutch]. Utrecht: Nizw; 2003.

41 Ploeg J, Davids B, Edwards N, Gifford W, Miller PE. Factors Influencing Best Practice Guideline Implementation: Lessons Learned from Administrators, Nursing Staff and Project Leaders. World views on Evidence-Based Nursing 2007(4):210:219.

42 Francke AL, Smit MC, Veer de AJE, Mistiaen P. Factors influencing the implementation of clinical guidelines for health care professionals: A systematic meta-review. BMC Medical Informatics and Decision Making 2008(38):1-11.

Chapter 5

Reducing interrater variability and improving health care: a meta-analytic review

This chapter is published as:

Tuijn SM, Janssens FJG, Robben PBM, Van den Bergh H. (2012) Reducing interrater variability and improving health care: A meta-analytic review. *Journal of Evaluation in Clinical Practice,* 18, 887-895.

5.1 Abstract

In the scientific literature about reliability, the main approach to increasing reliability seems to involve increasing the number of observers and improving the instrument used. Other aspects for improving reliability – like the training of raters – seem to receive less notice. It is worth asking whether this technical approach could be complemented by training the user of the instrument. A systematic meta-analytical review of the research literature was performed to answer this question and examine the effectiveness of planned interventions for improving interrater reliability of health care professionals.

In this study the databases of PubMed (MEDLINE), Embase, Omega, and PsycINFO were searched. The inclusion criteria were met by 57 studies. Details extracted from the studies included the study design, the number of observers and the number of observed cases, the intervention, the type of instrument (whether or not it was highly technical), and statistical information about the agreement before and after the intervention. Interventions were categorized into three groups: training of professionals, improving the diagnostic instrument, and a combination of training and improving the instrument. A meta-analysis was performed by means of linear regression.

The interventions were arranged according to their effectiveness in improving the diagnostic instrument (mean change: $|_i = 0.13$), training combined with improving the instrument (mean change: $|_i = 0.10$), and training (mean change: $|_i = 0.09$). Results indicate that on average, although all types of interventions are effective, improving the diagnostic instrument seems to be the most effective. Especially when highly technical instruments were concerned, improvement proved to be very effective ($|_i$ =0.52). Because instrumental variables constitute a major source of error, improving the instrument is an important approach. However, this review offers solid arguments that can complement the literature and practice, with a focus on training the user of the instrument.

5.2 Introduction

Variability in performance is an integral part of being human. No one operates fully consistently on all occasions. This is also true for health care professionals and inspectors who evaluate health care. This inconsistency in performance stems from a variety of factors, for example, variations in physical and mental welfare, external conditions, and the task to be performed as well as inconsistencies of the persons performing the task [1]. One of the specific competencies of health care professionals and evaluators or inspectors is to reach valid diagnoses

or judgments, taking into account that reliability is a necessary precondition for validity. However, the reliability of a judgment depends on a variety of factors: the differences in the professionals themselves and, just as importantly, the definition of the items evaluated, the objects or persons judged, the method of evaluating, and the setting and time of evaluation [1,2]. The combined effect of these factors on an evaluative score is referred to as error of measurement [1]. In many professions in which judgment plays an important part, the goal is often to minimize the error of measurement. In health care, the reliability of a diagnosis is significant because the patient's treatment is based on this judgment. In regulating health care, the reliability of judgments is of equal importance. Scriven is unequivocal about the role of evaluators: "Bad is bad and good is good and it is the job of evaluators to decide which is which" [3]. The evaluator must fulfill his or her role in serving the public interest, and this interest is not restricted to the evaluator's responsibility to clients, users, or stakeholders, but to all potential consumers [4]. After all, institutions have to improve their quality of care where necessary based on the judgments of an inspector or evaluator. There is more and more recognition for the importance of decision making within professional practices, which is confirmed by the increasing amount of research on this topic [5-8].

Earlier research on the regulation of health care in the Netherlands showed poor interrater reliability [9], and even showed a tendency to false positive judgments [10].In cases of false positive judgments, incentives for improving the quality of care remain unjustly undone. Furthermore, accountability and transparency in regulation are restricted by reliability and validity issues [11]. Some stress that the key to improving the quality of care is reducing variability [12]. In the literature on reliability, the main method for diminishing the error of measurement seems to be improving the instrumental reliability [1]. Less attention seems to be given to other aspects for improving reliability, such as the training of raters or observers. Because instrumental variables constitute a major source of error [1] , improving the instrument is an important approach. However, it is worth asking whether additional training of the raters could be a valuable complement to this approach. To answer this question it is essential to gain insight into the effectiveness of interventions to reduce interrater variance. Therefore, we performed a meta-analytical review to:

a) Identify the effectiveness of interventions for improving interrater reliability; and

b) Formulate recommendations for intervention(s) to reduce interrater variability of health care professionals as well as inspectors.

5.3 Methods

We collected data between the beginning of March 2009 and the end of June 2009. Because interrater variability occurs in a wide variety of professions, we searched medical databases (PubMed, MEDLINE, Embase) as well as sociological databases (Omega, PsycINFO) to identify papers published through June 2009; we did not introduce a lower limit for the publication year. Our search strategy consisted of three phases. First, the search strategy included a combination of MeSH terms (PubMed), Emtree terms (Embase), and free-text protocols (PubMed, MEDLINE, Embase, Omega, PsycINFO), as presented in Table 1.

Table 1 Search strategy

Search strategy	Results
1. PubMed (MEDLINE): “Observer variation”(Mesh) and “improving” (free text)	267 articles
2. PubMed (MEDLINE): “improving interobserver agreement” (free text)	232 articles
3. PubMed (MEDLINE): “interrater agreement” and “increasing” (free text)	18 articles
4. PubMed (MEDLINE): “improving" and "interobserver agreement” (free text)	30 articles
5. Embase: “Observer variation’ (Emtree) and “improving” (free text):	83 articles
6. Embase: “inter-rater agreement” and “increasing” (free text)	23 articles
7. Embase: “improving" and "interobserver agreement” (free text)	32 articles
8. Omega “Interobserver agreement”(stem of word) AND “improve” (whole word)	28 articles
9. Omega: “inter-rater agreement” and “increasing” (whole word)	800 articles
10. PsycINFO: “inter-rater agreement” and “increasing” (free text)	100 articles
11. PsycINFO: “improving" and "interobserver agreement" (free text)	1161 articles

5.4 Analysis

This strategy yielded 2774 studies. Subsequently, the first author scanned the titles and abstracts of potentially eligible studies, and excluded studies that clearly did not comply with the inclusion criteria. The first author also checked the papers’ reference sections to locate additional relevant articles. Articles were included if interrater reliability was described as a concern, an intervention was designed to increase reliability, empirical research was performed to determine the effect of the intervention, and the observers in the study were human observers (not computers). At the end of this process, 144 articles had been identified. The first author read and analyzed these articles carefully, and contacted the investigators if the statistics provided in the articles were incomplete or unsuitable. After careful analysis, 87 of them were excluded because they provided statistics that could not be transformed into κ-values, because investigators could not be contacted, or because the investigators who were contacted gave no additional information. If the first author doubted the extent to which the inclusion criteria had been met (n = 7), the other authors were consulted and disagreements were resolved by discussion. In the

end, 57 articles were selected. All of these articles were written in English, although this was not an inclusion criterion.

For descriptive purposes, we noted the study design (the presence of a control group), selection and allocation of observers and cases, setting, number of observers, number of cases, and type of intervention. If the intervention concerned improvement of instruments, we noted whether it concerned highly technical instruments like computed tomography scans, X-rays, and magnetic resonance imaging scans. For meta-analytical purposes, we noted pre-test and post-test statistics for observer agreement.

5.5 Meta-Analysis of Mean Change in Agreement

A meta-analysis was used to estimate the effects of different types of interventions. Because the majority of the studies reported κ-statistics, we transformed the statistics other than kappa into κ-values. When agreement was only presented in percentages, kappa was recalculated using a proportional (marginal) distribution of 50%. In this manner, we gave these studies the benefit of the doubt. We performed a linear regression analysis to estimate the effect of different types of interventions on interrater variability (SPSS 15.0, IBM, US). We weighted studies based on the number of observers and cases.

5.6 Results

Fifty-seven articles were identified. All of them concerned improving interrater reliability in health care professions. We found no empirical studies on interventions to increase interrater reliability in other professions like inspectors, teachers, or judges. The included studies incorporated a wide variety of medical and paramedical expertise. We categorized the literature in three groups and did so by consensus. The results of the analysis are presented in Table 2. As can be seen in Table 2, 38 (66%) of the studies assessed the effect of improving the diagnostic instrument on the reliability of professionals, 12 (21%) assessed the effect of training on the reliability of professionals, and seven (13%) examined whether training combined with improving the instrument increased the reliability of professionals.

Professional training focused mainly on identifying sources of variation, but varied in design. Some of the interventions focused on formulating consensus. Other interventions focused on practical teaching for using diagnostic classification systems or focused on lectures, sometimes by means of a web-based training module. Many of the improvements concerned mechanical adjustments of highly technical diagnostic instruments, while others focused on improving ranking scales. There was less variation among interventions that combined profes-

sional training and improving instruments. Most of these interventions combined the development of and training in medical decision criteria.

As can be seen in Table 2, in the group of studies on training, none of the training concerned highly technical instruments. This differs from the other groups. In the group concerned with improving instruments 61% of the interventions concerned highly technical instruments and in the group that combined training and improving instruments, 29% of the interventions concerned highly technical instruments. It can also be seen that the study design varied among studies in general and among types of interventions. A minority of the studies used a randomized clinical trial (0.05%), in contrast to the majority of the studies (95%), which used an experimental pre-test/post-test design. Studies on improving instruments used a control group less often (5.5%) compared with studies on training (50%) and studies on training as well as improving the instrument (29%). The mean pre-test kappa and the mean post-test kappa are also presented in Table 2. The dispersion of these values differs among studies. Studies on training have less dispersion (Sd pre-test = 0.12, Sd post-test = 0.12) compared with studies on instruments (Sd pre-test = 0.20, SD post-test = 0.21) and studies on training and instruments (SD pre-test = 0.23, SD post-test = 0.26). All of the studies had methodological issues; those present in the majority of the studies were the Hawthorne effect, the fact that most of the studies used the same cases in both the pre-test and the post-test, that the selection and allocation of observers and cases was not specified, and that there was no information about blindness for patient information.

Table 2 Interventions to improve interrater reliability of health care professionals.

Reference #	Cases (n)	Intervention	Highly technical instrument	Average pre-test and post-test kappa	Study design	Methodological issues
		Training				
[13]	51	The effect of formal lectures and practical training in stress echocardiography, image acquisition, and interpretation.	no	Pre-κ = 0.46 Post-κ = 0.49	Experimental, 1 group, pre-test-post-test	1,3,9,10,14
[14]	16	The effect of an independent reading and identification of sources of disagreement to formulate rules to define enhancing lesions in diagnosing gadolinium-enhanced lesions in multiple sclerosis.	no	Pre-κ = 0.52 Post-κ = 0.68	Experimental, 2 groups, pre-test-post-test	1,3,9,15
[15]	30	The effect of using the index of orthodontic treatment need as a tool on agreement.	no	Pre-κ = 0.45 Post-κ = 0.62	Experimental, randomized, 3 groups	1,3,9,15
[16]	42	The effect of training in the Neer classification system, on reducing variability in diagnosing proximal fractures of the humerus.	no	Pre-κ = 0.28 Post-κ = 0.64	Randomized clinical trail, 2 groups	1,3,15
[17]	88	The effect of a joint session behind the microscope and discussion about the observations and the interpretation on variability in histopathological grading.	no	Pre-κ = 0.42 Post-κ = 0.56	Experimental, 1 group, pre-test-post-test	1,3,9,15
[18]	95	The effect of discussion on agreement about adverse advents.	no	Pre-κ = 0.32 Post-κ = 0.47	Experimental, 1 group, pre-test-post-test study	2,3,9,14
[19]	30	The effect of an ultrasound workshop on reducing interobserver variation in the ultrasonographic evaluation of polycystic ovaries.	no	Pre-κ = 0.71 Post-κ = 0.84	Experimental, 1 group, pre-test-post-test	1,3,9,14
[20]	102	The effect of web-based training on the reliability of pressure ulcer risk assessments.	no	Pre-κ = 0.25 Post-κ = 0.35	Experimental, 2 groups, pre-test-post-test	1,3,5,9,14
[21]	16	The effect of training and an atlas on agreement of Gleason grading in prostatic adenocarcinoma.	no	Pre-κ = 0.44 Post-κ = 0.68	Experimental, 1 group, pre-test-post-test	1,3,6,9,15
[22]	15	The effect of two independent reading sessions, and a meeting to identify common sources of inconsistency,	no	Pre-κ = 0.42 Post-κ = 0.32	Experimental, 1 group, pre-test-post-test	1,3,9,15

		on agreement on diagnosing multiple sclerosis lesions.				
[23]	119	The effect of repeated slide conferences on rater agreement of nuclear atypia of breast cancer.	no	Pre-κ = 0.42 Post-κ = 0.42	Experimental, 1 group, pre-test-post-test	1,3,7,9,15
[24]	42	The effect of training on agreement in traditional Chinese medicine diagnosis of rheumatoid arthritis.	no	Pre-κ = 0.35 Post-κ = 0.46	Experimental, 1 group, pre-test-post-test	1,3,8,9,15
		Improvement of instrument				
[25]	23	The effect of computed tomography angiography compared to unenhanced CT on reliability of acute stroke assessment.	no	Pre-κ = 0.61 Post-κ = 0.76	Experimental, 1 group, pre-test-post-test	1,3,9,10,14
[26]	16	The effect of spin echo MRI images and of corresponding "multispectral" maps on agreement of white matter lesion detection.	yes	Pre-κ = 0.61 Post-κ = 0.72	Experimental, 1 group, pre-test-post-test	1,3,9
[27]	10	The effect of guidelines on interobserver reliability on diagnosing spine stenosis of patients with cervical spine myelopathy.	no	Pre-κ = 0.22 Post-κ = 0.06	Experimental, 1 group, pre-test-post-test	1,3,9,10,14
[28]	264	The effect of the brain observer microbleed scale on reducing variability of diagnosing brain microbleeds.	no	Pre-κ = 0.54 Post-κ = 0.67	Experimental, 1 group, pre-test-post-test	1,9,10,14
[29]	50	Effect of binary decision making on reliability of diagnosis of ankle fractures.	no	Pre-κ = 0.69 Post-κ = 0.7	Experimental, 1 group, pre-test-post-test	1,3,9,10,15
[30]	25	The effect of binary classification system on medical decision making on fractures of the tibial plafond as a model.	no	Pre-κ = 0.35 Post-κ = 0.46	Experimental, 1 group, pre-test-post-test	1,3,9,15
[31]	30	The effect of 3D computed tomography on reliability of distal humeral fractures characterization.	yes	Pre-κ = 0.24 Post-κ = 0.38	Experimental, 1 group, pre-test-post-test	1,3,9,14
[32]	142	The effect of a comprehensive drug screening method as a first line tool on clinical decision making on the emergency department for suspected drug overdose.	no	Pre-κ = 0.28 Post-κ = 0.49	Experimental, 1 group, pre-test-post-test	1,6,9,14
[33]	20	The effect of measurement technique on interobserver reliability of ovarian volume calculation from 3D ultrasound data.	yes	Pre-κ = 0.98 Post-κ = 1.00	Experimental, 1 group, pre-test-post-test	1,9,11,15

[34]	90	The effect of a web-based atlas of lumbar spine dual-energy X-ray absorptiometry images with options to exclude vertebrae.	yes	Pre-κ = 0.41 Post-κ = 0.54	Experimental, 1 group, pre-test-post-test	1,3,9,10,12
[35]	67	The effect of a new scoring scheme for the diagnosis of noninvasive endocervical glandular lesions.	no	Pre-κ = 0.57 Post-κ = 0.71	Experimental, 1 group, pre-test-post-test	1,3,6,8,9,10
[36]	104	The effect of computer-aided diagnosis on radiologists' interpretation of mammograms depicting microcalcifications.	yes	Pre-κ = 0.19 Post-κ = 0.41	Experimental, 1 group, pre-test-post-test	1,3,9,10
[37]	194	The effect of the marker p16ink on diagnosis of cervical lesions.	yes	Pre-κ = 0.60 Post-κ = 0.81	Experimental, 1 group, pre-test-post-test	1,3,9,10
[38]	30	The effect of different classifications systems for fractures of the pelvic ring on interrater agreement.	no	Pre-κ = 0.51 Post-κ = 0.52	Experimental, 1 group, pre-test-post-test	1,3,9,10,14
[39]	72	The effect of a checklist in which the symptoms were recorded in plain language instead of in abstract diagnostic terms on agreement of diagnosis of transient ischemic attacks.	no	Pre-κ = 0.65 Post-κ = 0.77	Experimental, 1 group, pre-test-post-test	1,3,9,10
[40]	93	The effect of guidelines on the diagnosis of pancreatic intraepithelial neoplasia and intraductal papillary-mucinous neoplasms.	no	Pre-κ = 0.24 Post-κ = 0.36	Experimental, 1 group, pre-test-post-test	1,3,9,10,14
[41]	70	The effect of addition of a spectrogram when rating pathological voices.	no	Pre-κ = 0.33 Post-κ = 0.27	Experimental, 1 group, pre-test-post-test	1,3,9,10
[42]	55	The effect of the addition of in vivo quantitative hydrogen 1 magnetic resonance spectroscopy can improve the radiologist's interpreting breast MR images to distinguish benign form malignant lesions.	yes	Pre-κ = 0.44 Post-κ = 0.56	Experimental, 1 group, pre-test-post-test	1,3,9,10,11,14
[43]	36	The effect of contrast echocardiography in the improvement of segmental quality and interobserver agreement during stress real-time 3D echocardiography.	yes	Pre-κ = 0.47 Post-κ = 0.76	Experimental, 1 group, pre-test-post-test	1,3,9,10
[44]	34	The effect of a 3D analysis tool that makes it possible to anatomically align 3D rest and stress data systemat-	yes	Pre-κ = 0.35 Post-κ = 0.68	Experimental, 1 group, pre-test-post-test	1,3,9,10

		ically on interobserver agreement.				
[45]	30	The effect of addition of oblique radiograph views to improve reliability of diagnosing fractures of the acetabulum.	yes	Pre-κ = 0.55 Post-κ = 0.34	Experimental, 1 group, pre-test-post-test	1,3,9,10,14
[46]	528	The effect of combining visual analysis of CMR cine sequences with corresponding parametric images of myocardial contraction on interobserver variability in assessing segmental function.	yes	Pre-κ = 0.64 Post-κ = 0.68	Experimental, 1 group, pre-test-post-test	1,3,9,10,14
[47]	459	The effect of adding an airway pressure signal to pressure tracings of central venous pressure and pulmonary artery occlusion pressure on reliability in the measurements of central venous pressure and pulmonary artery occlusion pressure in critically ill patients.	no	Pre-κ = 0.43 Post-κ = 0.62	Experimental, 1 group, pre-test-post-test	1,8,9,10,15
[48]	48	The effect of magnetic resonance on reliability of neonatal brain imaging of term infants.	no	Pre-κ = 0.68 Post-κ = 0.88	Experimental, 1 group, pre-test-post-test	1,3,9,10,15
[49]	59	The effect of computer-assisted diagnosis system on observer agreement of the presence of bone metastases.	yes	Pre-κ = 0.47 Post-κ = 0.57	Experimental, 1 group, pre-test-post-test	1,3,13
[50]	100	The effect of p16 immunohistochemistry on diagnosis of cervical biopsies.	yes	Pre-κ = 0.46 Post-κ = 0.65	Experimental, 1 group, pre-test-post-test	1,9,10,14
[51]	24	The effect of 3D-radiographic analysis compared to plain radiographs of proximal humeral fractures.	yes	Pre-κ = 0.38 Post-κ = 0.46	Experimental, 1 group, pre-test-post-test	1,3,9,10,15
[52]	23	The effect of 3D gadolinium enhanced technique in the iliac arteries on reliability.	yes	Pre-κ = 0.79 Post-κ = 0.92	Experimental, 1 group, pre-test-post-test	1,3,9,10,11
[53]	20	The effect of MRM combined with MRI in diagnosing myocardial ischemia resulting from coronary artery disease on reducing variability.	yes	Pre-κ = 0.74 Post-κ = 0.79	Experimental, 1 group, pre-test-post-test	1,8,3,9,10
[54]	16	The effect of image registration on interrater agreement in the visual detection of active multiple sclerosis lesions from serial magnetic resonance scans.	yes	Pre-κ = 0.23 Post-κ = 0.41	Experimental, 1 group, pre-test-post-test	1,9,10

[55]	75	The effect of quantitative measurements of intervertebral motion of the cervical spine on interobserver agreement.	yes	Pre-κ = 0.17 Post-κ = 0.77	Experimental, 1 group, pre-test-post-test	1,3,9,14
[56]	40	The effect of the operationalisation of NINDS-AIREN criteria for vascular dementia on observer agreement.	no	Pre-κ = 0.28 Post-κ = 0.39	Experimental, 1 group, pre-test-post-test	1,3,6,9,14
[57]	10	The effect of the addition of PET-CT on reducing variability in identification of the gross tumor volume in patients with gastro-oesophageal carcinoma.	yes	Pre-κ = 0.38 Post-κ = 0.45	Experimental, 1 group, pre-test-post-test	1,3,9,10,14
[58]	556	The effect of standardization of images display and objective criteria on silent myocardial ischemia by interpretation of planar thallium-201 imaging.	yes	Pre-κ = 0.62 Post-κ = 0.71	Experimental, 2 group	1,3,9,10
[59]	15	The effect of real-time 3d echocardiography on agreement.	yes	Pre-κ = 0.78 Post-κ = 0.79	Experimental, 1 group, pre-test-post-test	1,3,9,10,14
[60]	20	The effect of tissue harmonic imaging on reliability of wall motion analysis on dobutamine stress echocardiography.	yes	Pre-κ = 0.81 Post-κ = 0.92	Experimental, 2 groups, pre-test-post-test	1,3,9,10
[61]	15	The effect of dermoscopy on interobserver agreement of melanoma and melanocytic naevi.	no	Pre-κ = 0.52 Post-κ = 0.54	Experimental, 1 group, pre-test-post-test	1,3,9,10
[62]	50	The effect of immunohistochemical markers for vascular and lymphatic channels on interobserver agreement of diagnosing Lymphovascular invasion in colorectal cancer.	yes	Pre-κ = 0.23 Post-κ = 0.33	Experimental, 1 group, pre-test-post-test	1,3,9,14
		Training and Improvement of instrument				
[63]	54	The effect of didactic material, including mammographic and pathologic correlation of individual BI-RADS features for calcifications, masses, and asymmetric densities, and another half-hour of questions and answers on interrater agreement.	yes	Pre-κ = 0.35 Post-κ = 0.42	Randomized, 2 groups	1,3,9,10
[64]	20	The effect of anchors and training on the reliability of perceptual voice evaluation.	no	Pre-κ = 0.19 Post-κ = 0.32	Experimental, 1 group, pre-test-post-test	1,3,9,15
[65]	78	Refresher training in histological criteria by studying a standard document of histological criteria for endome-	no	Pre-κ = 0.65 Post-κ = 0.63	Experimental, 1 group, pre-test-post-test	1,3,9,10,15

		trial dating used by a reference panel.				
[66]	56	The effect of a computer-aided detection tool and training on agreement on acute pulmonary embolism.	yes	Pre-κ = 0.74 Post-κ = 0.78	Experimental, 2 groups, pre-test-post-test	1,3,9,14
[67]	20	The effect of a decision tree with diagnostic criteria and 40 minute lecture on agreement on diagnosing Gleason grading of prostatic cancer.	no	Pre-κ =0.33 Post-κ = 0.41	Experimental, 1 group, pre-test-post-test	1,3,9,10
[68]	56	The effect of standardized histologic criteria and image based training that outlined the criteria on reliability of diagnosing breast carcinoma.	no	Pre-κ = 0.53 Post-κ = 0.93	Experimental, 1 group, pre-test-post-test	1,3,9
[69]	58	The effect of a structured interview on agreement on diagnosing changes in activity and lifestyle after stroke.	no	Pre-κ = 0.78 Post-κ = 0.93	Experimental, 1 group, pre-test-post-test	1,3,9,10

Legend:
1=Hawthorne effect
2=correction for Hawthorne effect
3=same cases in pre-test and post-test: recall effect
4=bias because borderline cases were excluded
5=researcher attended pre- and post_test
6=observers might differ in knowledge
7=the composition of groups of respondents who attended conferences differed
8= bias because only the most difficult cases were selected
9= selection of observers not specified
10=allocation of cases not specified
11= bias because observers were researcher or authors
12=observers attended consensus meeting between pre- and post-test
13= bias because of voluntary participation of observers
14=cases blinded for patient information
15= blindness for patient information in cases not specified

CMR=Cardiovascular Magnetic Resonance; MRM=Magnetic Resonance Myelography; MRI=Magnetic Resonance Imaging; NINDS-AIREN=National Institute of Neurological Disorders and Stroke and Association Internationale pour la Recherché et l'Enseignement en Neurosciences; PET-CT=Positon Emission Tomography-Computed Tomography; BI-RADS= Breast Imaging Reporting and Data System.

5.7 Results of Meta-Analysis

We used κ-statistics to express agreement of health care professionals in both pre-test and post-test. We interpreted the κ-values according to the guidelines proposed by Landis and Koch [70]. According to these guidelines, κ-values of 0.00 to 0.20 represent slight agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, and 0.61 to 0.80 substantial agreement. Values over 0.80 are considered almost perfect agreement. Pre-test overall agreement for health care professionals was fair (κ = 0.31). Table 3 shows the estimated average agreement before and after implementing three groups of interventions. On average, the post-test agreement was moderate (κ = 0.43), an increase of 0.12. This increase varies between studies. Health care professionals' agreement with interventions that concern improving the diagnostic instrument was moderate (κ = 0.50) before and substantial (κ = 0.63) after intervention. The agreement of professionals with interventions based on professional training was fair before (κ = 0.27) and after (κ = 0.36) intervention. The agreement of professionals with interventions based on professional training and improving diagnostic instruments was moderate before (κ = 0.41) and after (κ = 0.51) intervention.

Table 3 Results from calculation of agreement and linear regression analysis on agreement of health care professionals

	Agreement before intervention	Agreement after intervention	Mean change in agreement (p)
Overall agreement	0.31	0.43	0.12*, p<.000
Improvement of instrument	0.50	0.63	0.13*, p<.001
Control group (+)	0.65	0.70	0.05
Highly technical instrument (+)	0.14	0.66	0.52
Training	0.27	0.36	0.09*, p<.001
Control group (+)	0.39	0.65	0.26
Training and Improvement of instrument	0.41	0.51	0.10*, p<.001
Control group (+)	0.52	0.68	0.16
Highly technical instrument (+)	0.28	0.32	0.04

* Significant differences between the mean change in agreement of the separate group and the overall change of agreement (p<0.05).

The effect of the three types of interventions is significant for the three groups of interventions (p < .001). However, the largest effect on agreement can be gained from improving instruments, especially highly technical ones (|i = 0.52). The presence of a control group in the study design is of particular importance in studies that concern training (|i = 0.26) or training as well as improving the instrument (|i = 0.16).

5.8 Discussion

What can be concluded from this review? It is striking that although we searched both medical and sociological databases, empirical studies on the effects of interventions for improving reliability were found only for medical professions. One possible interpretation is that health care professionals are leading the way on this subject. It can also be concluded that interventions work. The highest percentage of studies (66%) concerned interventions to improve the instrument. This seems to confirm that, in attempts to improve reliability, the emphasis is on improving instruments. Both an overall effect and the effects of three separate groups of interventions are shown. There is slight variation in the magnitude of the effect of interventions. Improving diagnostic instruments seem to be slightly more successful compared to the other interventions, particularly because the baseline agreement of these studies is already higher, which complicates increasing the level of agreement.

Results indicate that improving instruments would be the best choice for increasing the reliability of health care professionals. But can these results be generalized to other professionals, and to inspectors in particular? If so, can this be done unconditionally? There appear to be a couple of factors that constrain generalizability. First, the majority of studies on interventions to improve instruments concern highly technical ones. In contrast to these instruments, more subjective instruments occur as well. For example the regulatory instruments used in the Netherlands to regulate health care are highly subjective. They consist of descriptions of criteria which are judged on the basis of situations, documents, and interviews conducted during regulatory visits in situations with contrasting interests. Just like health care professionals, inspectors make their judgments in uncertain circumstances. In general, the technical instruments like computed tomography scans have already been in use for an extended period of time, and improvements are concerned with further fine-tuning. Training in the use of these instruments was not part of the intervention, because this was already an element of the relevant medical education. This differs substantially from the instruments used in regulation of health care in the Netherlands. Both the instruments as well as training in the use of these instruments are relatively new, and participating in training has been mainly voluntary up till now.

5.9 Limitations of the study

There are a number of methodological issues in the majority of the studies that might limit their internal validity. First of all, in all of the studies the Hawthorne effect could have had an impact on the results. In addition, most of the studies used the same cases in both the pre-test and the post-test, and therefore do not account for recall bias. Therefore, the increased reliability could also partly be explained by the learning effect of the respondents'

pre-test judgment on the post-test judgment. Furthermore, in the majority of the studies neither the selection and allocation of observers and cases nor blindness for patient information was specified. Lack of random selection or allocation and lack of blindness for patient information could have introduced biases that might have affected the results. Moreover, studies differed in the level of experience and number of included observers. The degree of improvement may vary with the experience of the observers, and this might have influenced the results. In addition, small numbers (for example, two observers) are often not representative.

Next, the review was based on a sensitive search strategy, and we find it unlikely that any study we overlooked would change our conclusion. However, we acknowledge that this review is subject to selection bias because studies presenting negative results are published less often. We performed pooling of data that differed in design and setting, and this is unusual. However, the topic of our study (reducing interrater variability) applies to a wide variety of (para) medical decision-making settings. Therefore, the inclusion of a broad range of studies in this review adds to the validity of the study, as our aim is to describe a phenomenon that is apparent in all (para) medical professions. Although the authors realize that this innovative approach is an uncommon method of data pooling, it is nevertheless necessary for investigating general interrater variability. Understanding of the effect of recall bias would be desirable. However, quantifying the effect of recall bias was not possible because the presence of recall bias differs in three conditions used in the meta-analysis. Therefore the comparison of the conditions on this aspect was not possible. When agreement was only presented in percentages, kappa was calculated on the basis of a proportional distribution of 50%. This could have induced an overestimation of the effect of interventions. Because this was only the case in five (8.8%) of the included studies, we presume that this has little or no effect on our results. Finally, it would have been better if the selection of the articles had been performed by two of the authors independently. Because there was doubt about the satisfaction of the inclusion criteria in only seven articles, we presume this will scarcely affect our results.

5.10 Conclusions

What implications does this review have for theory and for practice? Should reliability theory focus mainly on improving instruments to increase reliability? Despite the methodological restrictions of our study, we think that our study does have some implications. Because instrumental variables constitute a major source of error [1] , improving the instrument is an important approach. However, this meta-analytical review offers solid arguments which can complement the literature and practice, with a focus on training the user of the instrument. Moreover, this review offers knowledge about possibilities for increasing reliability in practice (including regulatory prac-

tice): Training the professional is a valuable way to increase reliability, particularly when (highly) subjective instruments are used. Professionals increasingly individualize their decision processes to a greater extent the longer they are out of their training programmes [22,24] , and as a result, obligatory re-education is important for professionals regardless of discipline [24]. A change in corporate culture might be an important precondition for realizing this. Moreover, the training should not only encompass the use of the instruments themselves, it should also deal with the gap that exists between the instrument and reality. Like health care professionals, inspectors have to learn to substantiate deviations from the standards. Certain situations will not fit into regulatory instruments, and substantiated deviation prevents regulation from becoming static. Moreover, substantiated deviations from standards induce reproducibility, which is important for transparency and accountability. Reproducibility is not only the cornerstone of good science [71] , it is the cornerstone of good regulation and health care as well. This review shows that the amount of studies on the effect of interventions to increase reliability is still modest and methodological issues are often present. Therefore, much more research on this topic is needed in the future.

5.11 Acknowledgments

We would like to extend our gratitude to Rinske van den Berg (Nivel) for her help with composing the right search strategy. We would also like to thank Marjolein Garretsen (IGZ) for quickly processing our requests for articles.

5.12 References

1 Feldt LS, Brennan RL. Reliability. In: R.L. Linn (Ed.), *Educational Measurement*. . Third ed. New York: Macmillan Publishing Company; 1989. p. 105-146.

2 Nunnally JC. Psychometric Theory. New York: McGraw-Hill; 1978.

3 Scriven M. New frontiers of evaluation. Evaluation Practice 1986;7:19.

4 Scriven M. Evaluation bias and it control. In: G.V. Glass (Ed.). *Evaluation studies review annual* (p.). Berverly Hills: CA: Sage; 1976. p. 220.

5 Brennan TA, Russell JL, Nan LL. Reliability and validity of judgments concerning adverse events suffered by hospitalized patients. Medical Care 1989;27(12):1148-158.

6 Kahn KL, Rubenstein LV. Structured Implicit Review for Physician Implicit Measurement of Quality of Care: Development of the Form and Guidelines for its Use. 1989;N-3016-HCFA.

7 Hayward RA, McMahon LF, Bernard AM. Evaluating the Care of General Medicine Inpatients: How good is implicit review? Ann Intern Med 1993;118(7):550-556.

8 Taylor BJ. Factorial Surveys: Using Vignettes to Study Professional Judgement. British Journal of Social Work 2006;36:1187-1207-1188.

9 Tuijn SM, Janssens FJG, Van den Bergh H, Robben PBM. "Not all judgments are the same: a quantitative analysis of the interrater reliability of inspectors at the Dutch Health Care Inspectorate" [Het ene oordeel is het andere niet: Interinspecteursvariatie bij inspecteurs van de IGZ: een kwantitatieve analyse; in Dutch]. Nederlands Tijdschrift voor Geneeskunde 2009(8):322:326.

10 Tuijn SM, Van den Bergh H, Robben PBM, Janssens FJG. "The relationship between standards and judgments in the regulation of health care" [De relatie tussen normen en oordelen in het toezicht op de gezondheidszorg; in Dutch]. Tijdschrift voor Gezondheidswetenschappen 2009;6:264-271.

11 Tuijn SM, Robben PBM, Janssens FJG, Van den Bergh H. Evaluating instruments for regulation of health care in the Netherlands. Journal of Evaluation in Clinical Practice 2011 DOI: 10.1111/j.1365-2753.2010.01431.x:411-419.

12 Jacobs B, Duncan J. Improving Quality and Patient Safety by Minimizing Unnecessary Variation. J. Vasc. Interv. Radiol. 2009;20:157-163.

13 Anand DV, Theodosiadis ID, Senior R. Improved interpretation of dobutamine stress echocardiography following 4 months of systematic training in patients following acute myocardial infarction. Eur J Echocardiogr 2004;5:12-17.

14 Barkhof F, Filippi M, Waesberghe van JH, Molyneux P, Rovaris M, Lycklama à Nijeholt G, et al. Improving interobserver variation in reporting gadolinium-enhanced MRI lesions in multiple sclerosis. Neurology 1997;49(6):1682-1688.

15 Bentele MJ, Vig KW, Shanker S, Beck FM. Efficacy of training dental students in the index of orthodontic treatment need. Am J Orthod Dentofacial Orthop 2002;122:456-462.

16 Brorson S, Bagger J, Sylvest A, Hrøbjartsson A. Improved interobserver variation after training of doctors in the Neer system. A randomized trial. J Bone Joint Surg Br 2002;84(7):950-954.

17 Vet de HC, Koudstaal J, Kwee WS, Willebrand D, Arends JW. Efforts to improve interobserver agreement in histopathological grading. J Clin Epidemiol 1995;48(7):869-873.

18 Hofer T, Bernstein SJ, DeMonner S, Hayward RA. Discussion between reviewers does not improve reliability of peer review of hospital quality. Med Care 2000;38:152-161.

19 Lujan ME, Chizen DR, Peppin AK, Kriegler S. Improving inter-observer variability in the evaluation of ultrasonographic features of polycystic ovaries. Reprod Biol Endocrinol. 2008;18:1-11.

20 Magnan MA, Maklebust J. The effect of Web-based Braden Scale training on the reliability and precision of Braden Scale pressure ulcer risk assessments. J Wound Ostomy Continence Nurs 2008;35:199-212.

21 Mikami Y, Manabe T, Epstein JI, Shiraishi T, Furusato M, Tsuzuki T, et al. Accuracy of Gleason grading by practicing pathologists and the impact of education on improving agreement. Hum. Pathol 2003;34:658-665.

22 Molyneux PD, Miller DH, Filippi M, Yousry TA, Radü EW, Adèr HJ, et al. Visual analysis of serial T2-weighted MRI in multiple sclerosis: intra- and interobserver reproducibility. Neuroradiology 1999;41(12):882-888.

23 Tsuda H, Akiyama F, Kurosumi M, Sakamoto G, Watanabe T. The efficacy and limitations of repeated slide conferences for improving interobserver agreement when judging nuclear atypia of breast cancer. The Japan National Surgical Adjuvant Study of Breast Cancer (NSAS-BC) Pathology Section. Jpn J Clin Oncol 1999;29(2):68-73.

24 Zhang GG, Singh B, Lee W. Improvement of Agreement in TCM Diagnosis Among TCM Practitioners for Persons with the Conventional Diagnosis of Rheumatoid Arthritis: Effect of Training. Journal of Alternative and Complementary Medicine 2008;14(4):381-387.

25 Aviv RI, Shelef I, Malam S, Chakraborty S, Sahlas DJ, Tomlinson G, et al. Early stroke detection and extent: impact of experience and the role of computed tomography angiography source images. Clinical Radiology 2007;62:447-452.

26 Brunetti A, Tedeschi G, Costanzo di A, Covelli EM, Aloj L, Bonavita S, et al. White matter lesion detection in multiple sclerosis: improved interobserver concordance with multispectral MRI display. J Neurol 1997;244:586-590.

27 Cook C, Braga-Baiak A, Pietrobon R, Shah A, Neto AC, Barros de N. Observer agreement of spine stenosis on magnetic resonance imaging analysis of patients with cervical spine myelopathy. J Manipulative Physiol Ther. 2008;31:271-276.

28 Cordonnier C, Potter GM, Jackson CA, Doubal F, Keir ', Sudlow CL, et al. Improving interrater agreement about brain microbleeds: development of the Brain Observer MicroBleed Scale (BOMBS). Stroke 2009;40:94-99.

29 Craig WL, Dirschl DR. Effects of binary decision making on the classification of fractures of the ankle. J Orthop Trauma 1998;12:280-283.

30 Dirschl DR, Adams GL. A Critical Assessment of Factors Influencing Reliability in the Classification of Fractures, Using Fractures of the Tibial Plafond as a Model. Journal of orthopaedic trauma 1997;11(7):471-476.

31 Doornberg J, Lindenhovius A, Kloen P, Dijk van NC, Zurakowski D, Ring D.
Two and Three-Dimensional Computed Tomography for the Classification and Management of Distal Humeral Fractures. Evaluation of Reliability and Diagnostic Accuracy. J Bone Joint Surg Am. 2006(88):1795-1801.

32 Fabbri A, Marchesini G, Morselli-Labate AM, Ruggeri S, Fallani M, Melandri R, et al.
Comprehensive drug screening in decision making of patients attending the emergency department for suspected drug overdose. Emerg Med J. 2003;20(1):25-28.

33 Raine-Fenning NJ, Campbell BK, Clewes JS, Johnson IR. The interobserver reliability of ovarian volume measurement is improved with three-dimensional ultrasound, but dependent upon technique. Ultrasound in medicine and biology 2003;29(12):1685-1690.

34 Hansen KE, Binkley N, Blank RD, Krueger DC, Christian RC, Malone DG, et al. An atlas improves interobserver agreement regarding application of the ISCD vertebral body exclusion criteria. J Clin Densitom 2007;10(4):359-364.

35 Ioffe OB, Sagae S, Moritani S, Dahmoush L, Chen TT, Silverberg SG. Proposal of a new scoring scheme for the diagnosis of noninvasive endocervical glandular lesions. Am J Surg Pathol 2003;27(4):452-460.

36 Jiang Y, Nishikawa RM, Schmidt RA, Toledano AY, Doi K. Potential of computed-aided diagnosis to reduce variability in radiologists' interpretation of mammograms depicting microcalcifications. Radiology 2001;220(3):787-794.

37 Klaes R, Benner A, Friedrich T, Ridder R, Herrington S, Jenkins D, et al. P16INK4a immunohistochemistry improves interobserver agreement in the diagnosis of cervical intraepithelial neoplasia. Am J Surg Pathol 2002;26(11):1389-1399.

38 Koo H, Leveridge M, Thompson C, Zdero R, Bhandari M, Kreder HJ, et al. Interobserver reliability of the young-burgess and tile classification systems for fractures of the pelvic ring. J Orthop Tr 2008;22:379-384.

39 Koudstaal PJ, Gijn van J, Staal A, Duivenvoorden HJ, Gerritsma JG, Kraaijeveld CL. Diagnosis of transient ischemic attacks: improvement of interobserver agreement by a check-list in ordinary language. Stroke 1986;17(4):723-728.

40 Longnecker DS, Adsay NV, Fernandez-del CC, Hruban RH, Kasugai T, Klimstra DS, et al. Histopathological diagnosis of pancreatic intraepithelial neoplasia and intraductal papillary-mucinous neoplasms: interobserver agreement. Pancreas 2005;31(4):344-359.

41 Martens JW, Versnel H, Dejonckere PH. The effect of visible speech in the perceptual rating of pathological voices. Arch Otolaryngol Head Neck Surg 2007;133(2):178-185.

42 Meisamy S, Bolan PJ, Baker EH, Pollema MG. Adding in vivo quantitative 1H MR spectroscopy to improve diagnostic accuracy of breast MR imaging: preliminary results of observer performance study at 4.0 T. Radiology 2005;236(2):465-475.

43 Nemes N, Geleijnse ML, Krenning BJ, Soliman OII, Anwar AM, Vletter WB, et al. Usefulness of Ultrasound Contrast Agent to Improve Image Quality During Real-time Three-Dimensional Stress Echocardiography. The American Journal of Cardiology 2007(99):275-278.

44 Nemes A, Leung KY, Burken van G, Stralen van M, Bosch JG, Soliman OI, et al. Side-by-side viewing of anatomically aligned left ventricular segments in three-dimensional stress echocardiography. Echocardiography 2008;26:189-195.

45 Petrisor BA, Bhandari M, Orr RD, Mandel S, Kwok DC, Schemitsch EH. Improving reliability in the classification of fractures of the acetabulum. Arch Orthop Trauma Surg 2003;123(5):228-233.

46 Redheuil AB, Kachenoura N, Laporte R, Azarine A, Lyon X, Jolivet O, et al. Interobserver variability in assessing segmental function can be reduced by combining visual analysis of CMR cine sequences with corresponding parametric images of myocardial contraction. J Cardiovasc Magn Reson 2007;9(6):863-872.

47 Rizvi K, Deboisblanc BP, Truwit JD, Dhillon G, Arroliga A, Fuchs BD, et al. Effect of airway pressure display on interobserver agreement in the assessment of vascular pressures in patients with acute lung injury and acute respiratory distress syndrome. Crit Care Med 2005;33(1):98-103.

48 Robertson RL, Robson CD, Zurakowski D, Antiles S, Strauss K, Mulkern RV. CT versus MR in neonatal brain imaging at term. Pediatr. Radiol 2003;33:242-249.

49 Sadik M, Suurkula M, Höglund P, Järund A, Edenbrandt L. Improved Classifications of Planar Whole-Body Bone Scans Using a Computer-Assisted Diagnosis System: A Multicenter, Multiple-Reader, Multiple-Case Study. J Nucl Med 2009;17:368-375.

50 Sayed K, Korourian S, Ellison DA, Kozlowski K, Talley L, Horn HV, et al. Diagnosing cervical biopsies in adolescents: the use of p16 immunohistochemistry to improve reliability and reproducibility. J Low Genit Tract Dis. 2007;11(3):141-146.

51 Sjödén GO, Movin T, Aspelin P, Güntner P, Shalabi A. 3D-radiographic analysis does not improve the Neer and AO classifications of proximal humeral fractures. Acta Orthop Scand. 1999;70(4):325-328.

52 Snidow JJ, Aisen AM, Harris VJ, Trerotola SO, Johnson MS, Sawchuk AP, et al. Iliac artery MR angiography: comparison of three-dimensional gadolinium-enhanced and two-dimensional time-of-flight techniques. Radiology 1995;196:371-378.

53 Song KS, Jang EC, Jung HJ, Kim KW, Yu H. Observer variability in the evaluation of multiple lumbar stenosis by routine MR--myelography and MRI. J Spinal Disord Tech 2008;21:569-74.

54 I Leng Tan LI, Schijndel van RA, Fazeka F, Filippi M.
Improved interobserver agreement for visual detection of active T2 lesions on serial MR scans in multiple sclerosis using image registration. J Neurol 2001(248):789-794.

55 Taylor M, Hipp JA, Gertzbein SD, Gopinath S, Reitman CA. Observer agreement in assessing flexion-extension X-rays of the cervical spine, with and without the use of quantitative measurements of intervertebral motion. Spine J. 2007;7(6):654-658.

56 Straaten van EC, Scheltens P, Knol DL, Buchem van MA, Dijk van EJ. Operational definitions for the NINDS-AIREN criteria for vascular dementia: an interobserver study. Stroke 2003;34(8):1907-1912.

57 Vesprini D, Ung Y, Dinniwell R, Breen S, Cheung F, Grabarz D. Improving observer variability in target delineation for gastro-oesophageal cancer--the role of (18F)fluoro-2-deoxy-D-glucose positron emission tomography/computed tomography. Clin Oncol 2008;20:631-638.

58 Wackers FJ, Bodenheimer M, Fleiss JL, Brown M. Factors affecting uniformity in interpretation of planar thallium-201 imaging in a multicenter trial. The Multicenter Study on Silent Myocardial Ischemia (MSSMI) Thallium-201 Investigators. J Am Coll Cardiol. 1993;21(5):1064-1074.

59 Walimbe V, Garcia M, Lalude O, Thomas J, Shekhar R. Quantitative real-time 3-dimensional stress echocardiography: a preliminary investigation of feasibility and effectiveness. J Am Soc Echocardiogr 2007;20:13-22.

60 Zaglavara T, Norton M, Cumberledge B, Morris D, Irvine T, Cummins C, et al. Dobutamine stress echocardiography: Improved endocardial border definition and wall motion analysis with tissue harmonic imaging. J. Am. Soc. Echocardiogr 1999;12:706-713.

61 Carli P, De Giorgi V, Naldi L, Dosi G. Reliability and inter-observer agreement of dermoscopic diagnosis of melanoma and melanocytic naevi. Eur.J. Cancer Prev 1998;7:397-402.

62 Elizabeth I, Harris EI, Lewin DN, Wang HL, Lauwers GY, Srivastava A, et al. Lymphovascular Invasion in Colorectal Cancer. An Interobserver Variability Study. Am J Surg Pathol 2008;32(12):1816-1821.

63 Berg WA, D'Orsi CJ, Jackson VP, Bassett LW, Beam CA, Lewis RS, et al. Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography? Radiology 2002;224:871-880.

64 Chan KM, Yiu EM. The effect of anchors and training on the reliability of perceptual voice evaluation. J Speech Lang Hear Res 2002;45(1):111-126.

65 Duggan MA, Brashert P, Ostor A, Scurry J, Billson V, Kneafsey P, et al. The accuracy and interobserver reproducibility of endometrial dating. Pathology 2001;33(3):292-297.

66 Engelke C, Schmidt S, Bakai A, Auer F, Marten K. Computer-assisted detection of pulmonary embolism: performance evaluation in consensus with experienced and inexperienced chest radiologists. Eur Radiol 2008;18(2):298-307.

67 Griffiths DF, Melia J, McWilliam LJ, Ball RY, Grigor K, Harnden P, et al. A study of Gleason score interpretation in different groups of UK pathologists; techniques for improving reproducibility. Histopathology 2006;48(6):655-662.

68 Turner RR, Weaver DL, Cserni G, Lester SC, Hirsch K, Elashoff DA, et al. Nodal stage classification for breast carcinoma: improving interobserver reproducibility through standardized histologic criteria and image-based training. J Clin Oncol 2008;26(2):258-263.

69 Wilson JT, Hareendran A, Grant M, Baird T, Schulz UG, Muir KW, et al. Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified Rankin Scale. Stroke 2002;33(9):2243-2246.

70 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

71 Weiss DJ, Shanteau J. The vice of consensus and the virtue of consistency. New York: Cambridge University Press; 2004. p. 226.

Chapter 6

Experimental studies to improve the reliability and validity of regulatory judgments on health care in the Netherlands: a randomized controlled trial and before and after case-study

This chapter is published as:

Tuijn SM, Van den Bergh H, Robben PBM, Janssens FJG. Experimental studies to improve the reliability and validity of regulatory judgments on health care in the Netherlands: a randomized controlled trial and before and after case-study. *Journal of Evaluation in Clinical Practice (27March 2014)*; doi: 10.1111/jep.12136

6.1 Abstract

We investigated the effect of two interventions on the reliability and validity of the judgments of health care inspectors, and explored the effect of increasing the number of inspectors based on the estimates of the interventions. We expected that the interventions and increasing the number of inspectors would improve the reliability and validity.

A randomized controlled trial and before and after case study were used. One random group of inspectors examined 16 cases with the unadjusted regulatory instrument and the other group used the adjusted instrument. Next, all inspectors took part in a consensus meeting. Subsequently, the inspectors examined 16 cases to assess the consensus meeting's effect.

To account for the hierarchical structure of our data and to generalise the results, we analysed three models to see which model fit our data best (MLWin v2.19). The model that allowed both inspector variance and error variance to vary between conditions fit best. Inspector variance was smallest after the consensus meeting (0.03) and the reliability coefficient was highest after the consensus meeting (0.59). The effect of the consensus meeting on validity is similar: inspector variance was lowest after the consensus meeting for the case on professionalism (0.03). The correlation coefficient to express the correlation between the assigned judgment and the corporate judgment was highest after the consensus meeting (0.48). Increasing the number of inspectors resulted in both higher reliability and validity values.

We conclude that the adjusted instrument increased rather than improved the variance. Mandatory participation in a consensus meeting and increasing the number of inspectors (within the conditions of the experiment) improved reliability and validity of regulatory judgments. Organising consensus meetings as well as increasing the number of inspectors per regulatory visit seem to be valuable interventions for improving regulatory judgments.

6.2 Introduction

Government regulation of health care aims to monitor and minimise risks in health care and simultaneously stimulate the quality of care. Internationally, the effects of regulation on the quality of public services have been discussed extensively and sometimes criticised [1-10]. Scientific research on the effects of regulation is limited, and generally focuses on the effects of using quality indicators to improve performance[6,7] and the effects of enforcement or surveyor styles [11-15]. As a research area, studies on the reliability and validity of regulatory judgments are still scarce. Research and publication on this subject is of particular importance. As inspectors

make judgments and decide whether health care organisations have to improve quality. Both the credibility and authority of enforcement agencies will be hampered when the judgments are not reliable nor valid. Moreover, scientific publications on this subject make it possible to exchange knowledge and learn internationally.

6.3 Regulation of health care in the Netherlands

In the Netherlands, regulation of health care is performed by the Dutch Health Care Inspectorate (IGZ). The IGZ is an independent agency within the Ministry of Health, Welfare and Sport. The IGZ safeguards the quality of care and enforces over 20 laws, for example, the Care Institutions Quality Act [16]. The IGZ aims for standardised procedures and reliable and valid judgments to stimulate the quality of care and to justify its regulatory decisions and activities. Regulators need methods to measure and monitor the performance of the organisations they regulate, a process described as 'detection' [17]. For this purpose, the IGZ uses a combination of three methods. Firstly, the IGZ employs regulation in response to incidents, in the event of emergencies that indicate structural shortcomings in health care. The second method is theme-based regulation. This method focuses on specific issues in health care. Sometimes these issues requiring the regulator's attention are put forward by the minister or parliament. Thirdly, since 2002 the IGZ has been using risk-based supervision to assess the quality of health care by means of indicators [18].

As in countries like Australia, the United States, Switzerland, Sweden, and Norway, quality indicators were introduced in the Netherlands to monitor and stimulate the quality of health care [1,19-22]. In risk-based supervision, a framework for the quality of care and accompanying sets of quality indicators are drawn up in cooperation with representatives from the health care sector. Subsequently, risk-based supervision consists of three phases: First, the IGZ analyses the data collected with the indicators and selects institutions at risk. Next, inspectors visit the selected institutions which are obliged to cooperate. Inspectors are required to express their opinion of the examined care. When the quality of care does not meet the standards of IGZ, institutions have to draw up an improvement plan and are obliged to improve their care accordingly. If inspectors have any doubt on the improvement plan, inspectors can decide whether a follow-up visit is necessary. Finally, if the improvements are not satisfactory, the IGZ can impose administrative sanctions and initiate penal measures.

This study focuses on regulatory judgments assigned within the system of risk-based supervision of nursing home care in the Netherlands. In this system, inspectors visit a selection of health care institutions consisting mainly of institutions at risk. This selection means that the institutions visited do not vary widely with respect to the risk score on the indicators. This implies that the inspectors visit and examine institutions that cov-

er only a small part of the spectrum; they visit institutions which perform relatively less good on the indicators. Therefore, inspectors have to distinguish between institutions which perform not so good and not good. As a result, it is necessary to measure very accurately to be able to expose small differences between these institutions. This implies strict requirements of the regulatory instruments and the inspectors. However, earlier research on regulation of health care in the Netherlands shows that the reliability and validity of regulatory judgments can be improved [23,24].

In the scientific literature on reliability, the main approach to increasing reliability seems to involve increasing the number of observers and improving the instrument used [25]. Literature on interventions used by regulatory authorities to improve their judgments is still scarce. Fortunately, improving interrater reliability is an important part of other professions as well. Earlier research shows that empirical studies on interventions to improve reliability are an integral part of improving medical practice ,[26] and that the main approach of improving reliability as described in the literature can be complemented by two other interventions: training the users of diagnostic instruments, and the combination of improving the instrument and training the users [26]. The outcomes of these studies would also seem relevant for health care inspectors. After all, health care inspectors are professionals as well, and also have health care backgrounds. However, it can be questioned whether the outcomes of earlier research can be generalised to health care professionals to health care inspectors with no restrictions. Inspectors assess organisations instead of patients, using instruments like written criteria or standards [27]. Is it fair to assume that interventions that increase the reliability of health care professionals also increase the reliability of health care inspectors? And what is the effect of increasing the number of inspectors on the reliability and validity of regulatory judgments?

To answer these questions, we investigated the interventions that proved effective for health care professionals [26] and fit into the corporate culture of the IGZ. We studied the effect of two interventions on reliability and validity: adjusting a regulatory instrument and participating in a consensus meeting. Consequently, we explored the effect of increasing the number of inspectors on the reliability and validity of regulatory judgments based on the results of the interventions. We expect that the reliability and validity of regulatory judgments will improve as a result of the interventions; we also expect this will improve as a result of the increase of the number of inspectors examining similar cases. This study offers the possibility of providing opportunities for further professionalisation of health care regulation. Before explaining the method used in this study, we will elucidate the instruments used.

6.4 Instrument for regulation of nursing home care in the Netherlands

The instrument for regulation of nursing home care which is used since 2008 consists of standards, a framework (including criteria), and aspects of risk. The standards describe the desired situation in nursing home care. The framework defines which judgment applies in which situation according to the criteria. Check marks can be placed next to the aspects of risks to support the judgment. Because the standards describe the desired situation for a specific nursing home criterion, they are formulated positively. In contrast to the standards, the aspects of risk describe situations considered to be potential risks, and are formulated negatively. The criteria are examined by inspectors during regulatory visits, and judged on a four-point scale: 'no risk', 'slight risk', 'high risk', and 'very high risk'. This scale runs from positive to negative, with 'very high risk' being the most negative. Inspectors can, but are not required to, check relevant aspects of risk before they make their judgment. The number of aspects of risk differs per criteria. For example the criterion 'pressure ulcers' consists of eight aspects of risk (Table 1).

The meaning of the judgments is determined largely by the aspects of risk, because checking the aspects in essence determines the meaning of the judgment. As can be seen in Table 1, if one aspect of risk is checked for the criterion 'pressure ulcers', the judgment 'slight risk' is conceivable. The meaning of 'slight risk' depends on which aspect has been checked. This implies that 'slight risk' can have at least eight different meanings, because eight different aspects of risk can be checked for pressure ulcers. In addition, other arguments (both defined and non-defined) can also decide whether the judgment 'slight risk' applies. This implies that there can be endless variations to the meaning of 'slight risk' and therefore the meaning is unclear. This can hamper the validity of the judgment.

6.5 Methods

We studied the effect of two separate interventions on both the reliability and the validity of the regulatory judgments: adjusting the regulatory instrument and participating in a consensus meeting.

Table 1 The criterion 'pressure ulcers' from the regulatory instrument for nursing home regulation in the Netherlands in 2009.

IGZ standard: pressure ulcers	Aspects of risk	No risk	Slight risk	High risk	Very high risk
-Timely recognition of health risks. -The right balance between adequate technical operation and the client's wishes and preferences/is representative at least of the prevention and treatment of pressure ulcers. -Aids readily available, and their appropriate and safe use. -Staff members apply guidelines and protocols based on current knowledge according to professional, generally accepted standards that include at least the subject of pressure ulcers. For each subject, national and, if possible, multidisciplinary guidelines are used. For the prevention and treatment of pressure ulcers, these guidelines are: -'Pressure ulcers', second edition, CBO 2002: This guideline includes scientific results, views of professionals, and best practices for pressure ulcers. - 'Collaboration and logistics for pressure ulcers', Salode 2003: Tripartite multidisciplinary guideline (NVVA, Arcares, Sting, AVVV, NPCP): This guideline includes practical applications for the prevention and treatment of pressure ulcers in nursing homes, and describes the actual tasks of the different officials.	-The protocol does not meet the requirements. -The presence of pressure ulcers is not recorded. -Redness of the skin that does not disappear when pressure is applied is not observed in a structural way. -Effective preventive measures are not usable. -Education or testing of knowledge and skills is missing. -Individual agreements about the prevention or treatment of pressure ulcers are not recorded in the client files. -The diagnostics, treatment, and/or evaluation of pressure ulcers are not dealt with in a multidisciplinary fashion. -Conditions (like communication) that result in agreements not being kept.	-No aspects are checked. -Other arguments that indicate no risk.	-One aspect is checked. -Other arguments that indicate a slight risk.	-Preventive measures are not usable. -The protocol does not meet the requirements. -Two other aspects are checked. -Other arguments that indicate a high risk.	- Four or more aspects are checked. - Other arguments that indicate a very high risk.

6.6 The first intervention: adjusting the regulatory instrument

Inspectors are not satisfied with the instrument as it is. Therefore, we organised an expert meeting with four experienced inspectors for nursing home care regulation to make an inventory of the desired adjustments. This meeting focused on two of the instrument's criteria: 'pressure ulcers' and 'professionalism of the staff'. We have chosen these criteria for two reasons. First, earlier research showed that rating the 'pressure ulcer' criterion can be very difficult [24]. Secondly, the 'professionalism of the staff' criterion is a new one in the instrument, and turned out to be hard to judge [28]. The experts evaluated the criteria on three dimensions: the clarity of the definition of the aspects of risks, the extent to which the aspects of risk cover situations in nursing homes, and the scoring methodology. This resulted in an inventory of possible adjustments to the instrument. As a result of this expert meeting, we adjusted the regulatory instrument on two points. First, we formulated the description of the aspects of risk for pressure ulcers and professionalism of the staff positively rather than negatively. In this manner, both the description of the standard and the aspects are formulated positively. Second, we made checking the aspects of risk mandatory. In summary, the first intervention we studied concerned the effect of these two adjustments on the instrument.

6.7 The second intervention: participating in a consensus meeting

The second intervention was a consensus meeting for nursing home care inspectors to identify common sources of variation. Therefore, the inspectors had to reach consensus about the order of two sets of four cases, which had to ascend from 'no risk' to 'very high risk'. They classified four cases for the criterion 'pressure ulcers' and four cases for 'professionalism of the staff' in order of severity of risks using the unadjusted instrument. First, they read the cases for one criterion to make an individual judgment. Next, the inspectors had to reach consensus about the order of the cases. The cases were presented on large wheeled boards. In this way, the inspectors could easily gather around the cases, discuss them, and change their order. They were only allowed to change the order if there was consensus about how to replace a case. The inspectors had to state their arguments so that all participants joined in the discussion. They had to reach consensus within a time limit of 30 minutes per criterion. At the end of the session, one of the inspectors had to present the order of the cases and give the arguments that led them to decide on the order. The sources of variation were explained as well. Except for the time limit, no further instructions were given on how the inspectors were to reach consensus. Two of the researchers attended the consensus meeting, clarified the purpose of the meeting, and observed the participants without intervening. We videotaped these meetings. The outcomes are presented in Box 1.

Box 1 Results of the consensus meeting: sources of variation.

1. Some inspectors focus mainly on the aspects of risk presented in the instrument; others make tactical choices as well, and involve the context when they make a judgment.
2. Some inspectors think of a regulatory visit as an instantaneous sample; others think of it as part of the long-term developments of the health care organisation.
3. The level of palpability of the criterion is important. Inspectors experience the criterion 'pressure ulcers' as con-crete in contrast with the criterion 'professionalism of the staff'.
4. The validity of the instrument plays a part in how it is used. Inspectors do not agree whether it can be stated une-quivocally that a very high risk is present for the care delivered if a nursing home does not meet the standards for good care.
5. The size of the organisation in terms of the number of beds is not part of the instrument's criteria, nor is it ex-plained how inspectors can account for an organisation's size.
6. How the information in the instrument is formulated plays a part in the inspectors' judgments.
7. A regulatory judgment is not clinical, but is always based on the inspector's experience and knowledge. Inspec-tors' frames of ref-erence vary, and play a role in judging an organisation.
8. Some inspectors focus on details, while others focus on the main points.
9. Some possibilities for improving the instrument:
   - The instrument is too unstable in relation to the subjects and application.
   - The instrument is ambiguous and unclear on some points.
   - Does the subject of the instrument actually reveal risks in health care?
10. Some inspectors consider the instrument a decision-making aid, while others consider it to be an end in itself.
11. Some inspectors would like to start a regulatory report by explaining why some of the instrument's modules were either discussed or not discussed during the regulatory visit.
12. Some inspectors object to scoring 'no risk', and never assign this score.
13. Some inspectors prefer the strategy of building credits with an institution, and do not assign the score 'very high risk' for this reason. Other inspectors are convinced that the frame of reference is determined for the scores they assign. If in their experience a judgment has not had the foreseen effect, they assign scores in a different manner.
14. Sometimes inspectors choose not to write a report on the regulatory visit. Instead they give the institution the chance to improve the care. These inspectors think this strategy is more effective compared with assigning a lot of 'very high risk' scores.

Box 1 shows that the inspectors came up with many different arguments to reach consensus and identified different sources of variation. Some of the sources are focused mainly on the instrument (1,3,4,5,6,9,12), while others are more general (2,7,8,10,11,13,14). For example, whether a regulatory visit is an instantaneous sample or part of the health care organisation's long-term development is a more general point of difference. Choosing not to write a report as a strategy for letting the institutions improve on their own is a more personal type of variation.

Moreover, we calculated the effect of increasing the number of inspectors who examined the same cases on the reliability and validity of the regulatory judgments within the conditions of the experimental setting. We calculated reliability and validity when two inspectors examined the same cases, when three inspectors examined the same cases, and so on, up to a total of 10 inspectors. The values calculated represent the values that can be obtained when the requirements of the experimental setting are met. In this study, this implies that the inspectors do not talk with each other while they are examining the cases. In this study we used a randomized controlled trial and a before and after case study to examine the effect of the interventions (Table 2). We randomly assigned the inspectors (n=25) to Group 1 and Group 2 for the first measurement. During the first measurement, inspectors in both groups examined 16 identical cases within 6 weeks. Eight of the cases concerned pressure ulcers and eight cases concerned the professionalism of the nursing home staff. The inspectors in Group 1 used the unadjusted instrument, and the inspectors in Group 2 used the adjusted instrument.

Table 2 Research design of the study.

R Group 1: unadjusted instrument (O) N=15 inspectors	Consensus meeting (X) N=15 inspectors	Group 3: unadjusted instrument, after consensus meeting (O) N=15 inspectors
R Group 2: adjusted instrument (X) N=9 inspectors		

Legend
O: observation
X: treatment
R: random allocation of inspectors

For the second measurement, all inspectors who participated in the consensus meeting were assigned to Group 3. Although the 16 cases used in Group 3 were very similar to the cases used in Groups 1 and 2, to prevent learning effects they were not completely identical. Four weeks after the first measurement, the consensus meeting took place for all inspectors. After this meeting, the inspectors of Group 3 examined the cases with the unadjusted instrument within six weeks. This second round of review was conducted after a significant period of time had elapsed following the first measurement (six weeks); this was done to prevent recollection, which would have introduced bias into the review process [29].

To increase response among the inspectors, we sent two reminders for both the first and second measurements. In the end, 9 inspectors used the adjusted instrument and 15 inspectors used the unadjusted instrument. Of the 25 inspectors, 15 inspectors attended the consensus meeting (60%). Of the 15 inspectors in Group 3, 15 inspectors examined the cases (100%). The inspectors who dropped out withdrew themselves from the study despite the reminders we sent.

The cases concerned two criteria: 16 of the cases described the criterion ‘pressure ulcers’ and 16 cases described ‘professionalism of the staff’. Because in the system of risk-based supervision inspectors visit a selection of health care institutions at risk, the institutions visited do not vary widely with respect to the risk score on the indicators. This implies that the inspectors visit and examine institutions that cover only a small part of the spectrum. In this experiment, we tried to simulate this situation in an optimal way by using only cases that also represented just a small part of the spectrum: cases that corresponded to the scoring categories ‘slight risk’ and ‘high risk’. We developed the cases using descriptions of situations from regulatory reports of nursing home visits in 2008, and validated them. The best test for validity would compare the results of a measurement process with a ‘true score’ [29]. To develop such a gold standard, three former nursing home care inspectors rated all cases. These inspectors read the cases independently and assigned scores based on the four-point scale. They

did not always agree on all cases. These cases were discussed and rewritten to reach consensus on the level of risk. Box 2 presents an example of such a case.

Box 2 Case on pressure ulcers (representing 'high risk' according to the IGZ corporate standards)

The Sparrow is a nursing home with 25 beds. The atmosphere seemed a little cool. The new cluster manager, who started his job in March 2008, stated that the employees are involved in the delivered care. The IGZ noticed that the volume of the television in the shared living room was very loud. The Sparrow's financial position has improved recently: the deficit has been reduced. The numbers of reported falling incidents and medication errors have been stable for years. The IGZ finds this a conspicuous fact.
The Sparrow does not use a protocol for pressure ulcers. However, general instructions for coping with pressure ulcers are present. In the interview that took place, it was said that a digital protocol for pressure ulcers was being developed that will be available via intranet. The prevalence of pressure ulcers is measured within the scope of high quality and safe care in nursing homes. The outcomes of the measurement are not currently in use. Early signs that can indicate the presence or development of pressure ulcers are not recorded. The prevention of pressure ulcers takes place by purchasing preventive materials in the short-term, and by changing the lying position of residents at risk for pressure ulcers. The Sparrow does not facilitate education on the subject of pressure ulcers. Agreements were made about recording the treatment of pressure ulcers. These agreements were present in two of the four files examined. Although pressure ulcers are diagnosed by a multidisciplinary team, the agreements made are not always carried out.

The inspectors examined cases individually online using a web-based survey. This technology made it possible to prevent inspectors from returning to a previous case once they had judged it. In this manner, we attempted to make it harder for the inspectors to mutually compare cases and stimulate inspectors to rely more on the regulatory instrument. Moreover, with this technology we made sure that inspectors had to check the required parts of the study before they were able to go on to the next case. This was necessary to be able to examine the effect of the requirement to check aspects of risk. In addition, with this technology we attempted to reduce the chances of missing data. Although we presented the cases randomly to prevent effects of sequence, the order in which every observer examined them was similar. Because inspectors examined the cases at different locations, we minimised the possibility of discussing the cases simultaneously.

The data of this study are hierarchical, as ratings are nested both within inspectors and cases; randomly chosen ratings of the same inspector are more alike than randomly chosen ratings of randomly chosen inspectors. The same holds for the ratings of the same cases. The results of this study have to be generalisable over both inspectors and cases. Therefore we need to estimate three components of variance: the variance between cases, the variance of inspectors (the extent to which inspectors differ in their judgments on a case about nursing home care) and the interaction between inspectors and cases which is represented by error variance. Note that both error variance and inspector variance are indications of the reliability of the ratings.

We are interested in the three components of variance and the proportion between these components to be able to compare between the interventions. Moreover we were interested in the overall effect which is represented by the reliability coefficient (rho). To calculate rho we used the following formula [30]:

$$\rho = \frac{variance\ between\ cases}{variance\ between\ cases + \frac{variance\ between\ inspectors}{N_{inspectors}} + \frac{residual\ variance}{N_{inspectors} * N_{cases}}}$$

However, we were not interested only in the effect of the interventions on reliability, but on validity as well. Therefore, not only the variances but also the mean differences between the actual judgments and the corporate judgment (which corresponds with the IGZ's corporate standards) are relevant, as they indicate whether the actual judgments differ from the corporate judgments.

To examine the effect of the interventions on validity, we constructed a new variable that represented the gold standard in this study: corporate judgment. This is the judgment that was assigned by the four experts during the expert meeting when the cases were validated. This made it possible to compare the corporate judgments and the judgments assigned by the inspectors during the experiment, and we were able to examine the effect of the interventions on reliability as well as on validity at the same time. First, we analysed which model fit our data best. Second, we analysed the data to gain insight into the means and the proportion of variances of the judgments for the three conditions with respect to reliability. Third, we analysed the data to gain insight into the relationship between the corporate judgment and the actual judgment for the four conditions. Fourth, we calculated the effect of increasing the number of inspectors on reliability and validity.

6.8 Results

Results of the modelling are presented in Table 3. The results indicate that -2LL of Model 1 was higher compared with the -2LL of Model 2. The difference between Model 1 and Model 2 was 75 and a loss of five degrees of freedom. Yet, the -2LL of Model 2 was higher compared with Model 3. The difference between Model 2 and Model 3 was 74 and a loss of six degrees of freedom. The results indicate that Model 3 fits our data best ($p<0.001$). The estimated variances of Model 3 give insight into the effect of the interventions on reliability (Table 4) and validity (Table 5).

Table 3 Outcomes of the comparison of the three models used to represent our data.

		Comparison			
	-2 Log Likelihood	Model	X^2	Df	P
1. Equal reliability model	1268.73	Model 1 with Model 2	75.24	5	<0.0001
2. Different error model	1193.49	Model 2 with Model 3	74.85	6	<0.0001
3. Different variance and error model	1118.65				

Table 4 The effect of adjusting the instrument and a consensus meeting on interrater reliability for the three conditions.

	Mean (CI)	S^2_{error} (%)	$S^2_{inspector}$ (%)	S^2_{case} (%); rho
	Cases on professionalism			
Unadjusted	2.12 (1.75; 2.50)	0.39 (44)	0.08 (9)	0.41 (47); .47
Adjusted	3.27 (2.82; 3.72)	0.22 (26)	0.22 (26)	0.41 (48); .48
Consensus	3.81 (3.48; 4.14)	0.26 (37)	0.03 (4)	0.41 (59); .59
	Cases on pressure ulcers			
Unadjusted	2.51 (2.18; 2.84)	0.61 (62)	0.02 (2)	0.35 (35); .35
Adjusted	2.93 (2.53; 3.34)	0.39 (45)	0.14 (16)	0.35 (40); .40
Consensus	2.99 (2.63; 3.30)	0.24 (38)	0.05 (8)	0.35 (54); .54

Legend
$S^2_{inspector}$: variance between inspectors
S^2_{case}: variance between cases
S^2_{error}:variance between inspectors and cases
% error: percentage of variance explained by error
% insp: percentage of variance explained by inspectors
% case: percentage of variance explained by cases
CI: 80% confidence intervals
Rho: mean reliability when one inspector examines a case

Table 4 shows that for both the cases on professionalism and on pressure ulcers, the mean judgment assigned after the consensus meeting was higher (more stringent) compared with the other conditions. The error variance for cases on professionalism was relatively small when the adjusted instrument was used (0.22) and after the consensus meeting (0.26). This is also represented in the percentages of variance: the percentage of error variance was relatively small when the adjusted instrument was used (26%) and after the consensus meeting (37%) compared with the percentage of error variance when the unadjusted instrument was used (44%). Moreover, inspector variance was relatively small after the consensus meeting (0.03) compared with both the condition in which the unadjusted instrument was used (0.08) and the condition in which the adjusted instrument was used (0.22). This means that the mean differences between inspectors were relatively small after the consensus meeting. This is also depicted in the percentage of inspector variance, which explains the inspectors' part in the total amount of variance: after the consensus meeting, 4% of the total variance can be explained by inspectors for the case on professionalism. The reliability coefficient was also highest after the consensus meeting (0.59). The

percentage of variance explained by inspectors when the unadjusted instrument was used (9%) was relatively small compared with the percentage of variance explained by cases (47%) or error (44%). To be able to examine the effect of the interventions on the validity of the judgments, we calculated the mean difference between the judgments assigned by the inspectors and the corporate judgment. In table 5 the parameter estimates of model 3 are presented. For both the cases on professionalism and on pressure ulcers, the mean judgment assigned with the unadjusted instrument was lower (more lenient) compared with the corporate judgment, which was expressed by a negative mean difference. Inspectors who used the adjusted instrument and inspectors who participated in the consensus meeting assigned higher scores (were more stringent), which was expressed by a positive mean difference. The percentage of error differed between the conditions, but was relatively high when inspectors used the unadjusted instrument (59%) compared with the percentage of error after the consensus meeting (47%).

Table 5 The effect of adjusting the instrument and a consensus meeting on validity for the three conditions.

	Mean difference (CI)	S^2_{error} (%)	$S^2_{inspector}$ (%)	S^2_{case} (%); rho
	Cases on professionalism			
Unadjusted	-0.26 (-0.59; 0.07)	0.49 (59)	0.07 (8)	0.28 (34); 0.34
Adjusted	0.77 (0.38; 1.16)	0.21 (30)	0.21 (30)	0.28 (40); 0.40
Consensus	0.85 (0.57; 1.12)	0.27 (47)	0.03 (5)	0.28 (48); 0.48
	Cases on pressure ulcers			
Unadjusted	-0.06 (-0.22; 0.1)	0.38 (59)	0.04 (6)	0.23 (35); 0.35
Adjusted	0.43 (0.19; 0.67)	0.40 (53)	0.13 (17)	0.23 (30); 0.30
Consensus	0.37 (0.24; 0.5)	0.23 (46)	0.05 (9)	0.23 (45); 0.45

Legend
$S^2_{inspector}$: variance between inspectors
S^2_{case}: variance between cases
S^2_{error}: variance between inspectors and cases
% error: percentage of variance explained by error
% insp: percentage of variance explained by inspectors
% case: percentage of variance explained by cases
CI: 80% confidence intervals
Rho: mean reliability when one inspector examines a case

Inspector variance was relatively small after the consensus meeting (0.03) compared with both the pre-test when the unadjusted instrument was used (0.07) and when the adjusted instrument was used (0.21). This is also depicted in the percentages of inspector variance that explain the total amount of variance: after the consensus meeting the percentage of inspector variance was relatively small (5%) compared with the pre-test when the unadjusted instrument was used (8%) and the condition in which the adjusted instrument was used (30%). This might be explained by the fact that the adjusted instrument was new for the inspectors and they were not educated in the use of the new instrument. The correlation coefficient to express the correlation between the assigned judgment and the corporate judgment was highest after the consensus meeting (0.48). To be able to gain insight into the effect of increasing the number of inspectors on the reliability and validity of judgments for the different conditions, we calculated the reliability coefficient (rho) for different numbers of inspectors (Figure 1). Figure 1a shows that when the number of inspectors increases, reliability increases as well. The increase of rho varies between conditions. Figure 1b shows that when the number of inspectors increases, the correlation coefficient between the assigned judgment and the corporate judgments increases as well. The increase varies between conditions. The highest increase is effected on both reliability and validity in the condition after the consensus meeting. Figures 1a and 1b show that the effect of the increase of inspectors declines after three inspectors.

Figure 1A The effect of increasing the number of inspectors on the reliability of judgments

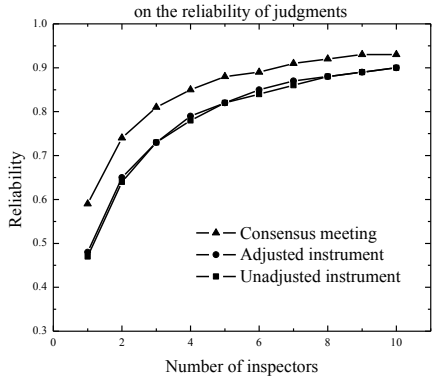


Figure 1B The effect of increasing the number of inspectors on the validity of judgments

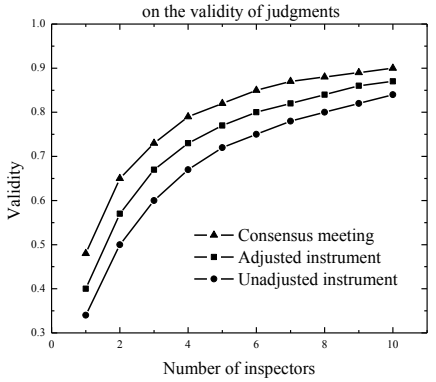


Figure 1 The effect of increasing the number of inspectors on the reliability and validity of regulatory judgments.

6.9 Discussion

In this study we examined the effects of a consensus meeting and adjusting the regulatory instrument on the reliability and validity of regulatory judgments. Based on the estimates of the variances we obtained from the results of the consensus meeting and adjusting the instrument, we explored the effect of increasing the number

of inspectors who examined similar cases. Because error variance was relatively small and rho was relatively large after the consensus meeting, we conclude that the consensus meeting results in more homogeneous judgments compared with adjusting the instrument. Moreover, when inspectors used the adjusted instrument, inspector variance was larger compared with the unadjusted instrument. This implies that inspectors who used the adjusted instrument are less mutually interchangeable when the adjusted instrument is used. The mean difference of the judgments was negative when the unadjusted instrument was used. This implies that when the unadjusted instrument is used, the assigned judgments are more lenient compared with the corporate judgments. Earlier research has confirmed this tendency towards false-positive judgments [28].

We conclude that the consensus meeting, adjusting the instrument, and increasing the number of inspectors per examined case, all influenced the mean judgments and components of variance. However, the results indicate that participating in a consensus meeting and increasing the number of inspectors per examined case improved reliability and validity. The calculations we made to explore the effect of the increase in the number of inspectors on reliability and validity presume that groups of inspectors assign scores to similar cases under the same condition as in our case study: they do not speak with each other about their scores when examining the cases. However, it seems unrealistic to expect that, when visiting in pairs or teams, inspectors will not discuss their observations with each other. Therefore, it seems reasonable to expect that, in actual practice (when inspectors do speak with each other about their scores), the increase in the reliability of the regulatory judgments will be higher. Although based on earlier research [26] we expected that adjusting the instrument would improve reliability and validity, we were not able to confirm this in our study. With respect to the reliability coefficients after the consensus meeting, we conclude that a relatively high percentage of variance of judgments is still not representative of variation between institutions. This might have to do with the system of risk-based supervision, which is characterised by visiting only a selection of institutions at risk that do not vary widely with respect to the risk score on the indicators. We simulated this by using cases that represented only ‘slight risk’ and ‘high risk’. The reliability coefficient after the consensus meeting indicates that examining cases that cover only a small part of the spectrum is very complex indeed.

This study has several strengths. First, to our knowledge, this is the first study to investigate the effect of interventions on interrater reliability and validity of health care inspectors. Second, in an experimental design in which cases are examined there is always a risk of recall effect and learning effect, which might affect the results. We attempted to limit these effects as much as possible by developing very similar but not completely identical cases for the first and second measurements, and planning six weeks between them. Third, the best test

for validity is to compare the results of a measurement process with a 'true score' [29]. To develop such a standard, we developed a proxy we referred to as the gold standard. We adjusted the cases in the experiment to approximate validity. Fourth, the inspectors examined cases individually online using a web-based survey. This technique made it possible to prevent inspectors from returning to a previous case once they had judged it. In this manner, we attempted to make it more difficult for inspectors to mutually compare individual cases and simultaneously stimulate inspectors to rely on the regulatory instrument. In addition, with this technique we attempted to reduce the chances of missing data. Because inspectors examined the cases at different locations, we minimised the possibility of discussing the cases simultaneously.

Limitations to the study should also be considered, because they may affect the results. First, although using cases to examine interrater reliability is very common, this might have affected the results. After all, no matter how well designed the cases are, they will never be completely identical to the complexity of reality. In this study we experienced quite a lot of resistance to the use of cases. Second, because study participants may have the tendency to concentrate particularly when they are aware they are participating in an experiment, the Hawthorne effect might be present. Third, in experimental designs it is recommended that every participant examine the cases in a random order, and this order differs among participants to prevent effects of sequence. In our study, although the cases were presented randomly, the order in which the cases appeared online did not vary among inspectors due to the web-based technique that was used. Moreover, the dropout rate in this experiment was not random. It is remarkable that despite the fact that management strongly supported this experiment, some members of the organisation withdrew from the study.

We think these results have important implications. Our results indicate that mandatory participation in a consensus meeting and increasing the number of inspectors per regulatory visit improves the reliability and validity of regulatory judgments. The results show that the way we adjusted the regulatory instrument did not improve reliability and validity. Almost all regulators use standards to state their expectations to other stakeholders in regulation, the most obvious being the organisations they regulate [27]. Training inspectors how to use instruments and bring about consensus on employing such standards is important for reliable and valid judgments. Maybe it was naïve to expect that adjusting the instrument without explicitly training the inspectors to use the new instrument would result in higher agreement or validity.

6.10 Future research

The results indicate that the reliability coefficient after the consensus meeting is still not yet optimal. The consensus meeting of our study was mainly focused on identifying sources of variation. Although this focus was an integral part of the consensus meetings studied earlier, the manner in which the outcomes of the consensus meetings were used differed among studies [26]. Because some of the sources of variation in this study were quite fundamental, it might be necessary to develop conventions to be able to further improve reliability. This seems a rational continuation for future research on this subject. In addition, we only examined two types of adjustments to the regulatory instrument in this study, without training the inspectors to use the new instrument. It could be valuable to investigate how other adjustments to regulatory instruments can accomplish improving the reliability and validity of the regulatory judgments in combination with training in using the instrument. Third, we examined the effect of increasing the number of inspectors in an experimental setting. This implies that the calculated values represent the reliability and validity of regulatory judgments when inspectors do not discuss their observations before they make a judgment. Because it seems unrealistic to expect that inspectors will meet these requirements in daily practice, it is worth examining the effect of increasing the number of inspectors during actual regulatory visits. Moreover, when inspectors visit institutions in pairs or teams, there is a risk of unwanted side effects. As a result of the dynamics in pairs of inspectors (for example, factors like dominance, seniority, status, and the ability to argue [31]), the agreement between pairs of inspectors or between inspectors of a regulatory region about a judgment can increase, but this does not necessarily imply that the judgment is valid. Therefore, future research on optimal conditions for inspectors to visit health care institutions in increasing numbers would be a valuable continuation of this study.

6.11 References

1 Brennan TA. The Role of Regulation in Quality Improvement. The Milbank Quarterly 1998;76(4):709-731.

2 Walshe K. Improvement through inspection? The development of the new Commission for Health Improvement in England and Wales. Qual. Health Care 1999(8):191-196.

3 Walshe K. The rise of regulation in the NHS. BMJ 2002;324:967-970.

4 Bevan G, Hood C. Targets, inspections, and transparency. British Medical Journal 2004;324:967-970.

5 Ham C. From targets to standards: but not just yet.The challenge will be for ministers not to interfere in a regulated service. BMJ 2005;330:106-107.

6 Bevan G., Hood C. Have targets improved performance in the English NHS? British Medical Journal 2006;332:419-422.

7 Bevan, G.,Hood, C. What's measured is what matters: targets and gaming in the english public health care system. BMJ 2006;84:517-538.

8 Bevan G. Have targets done more harm than good in the English NHS? BMJ 2009;338:a3129.

9 Gubb J. Have targets done more harm than good in the English NHS? BMJ 2009;338:a3130.

10 Ham C. Improving the performance of the English NHS. Systems of care are needed to build on progress to date. BMJ 2010;340:c1776.

11 Day P, Klein R. The regulation of nursing homes. The Milbank Quarterly 1987;65(3):303-347.

12 Hutter BM. Variations in regulatory enforcement styles. Law and Policy 1989;2:153-174.

13 Braithwaite J, Makkai T, Braithwaite V. Regulating Aged Care. Ritualism and the New Pyramid. Cheltenham, UK: Edward Elgar; 2007.

14 Greenfield D, Braithwaite J., Pawsey M. Healthcare accreditation surveyor styles typology. Int J of Health Care 2008;21:435-443.

15 Mascini P, Wijk van E. Responsive regulation at the Dutch Food and Consumer Product Safety Authority: an empirical assessment of assumptions underlying the theory. Regulation and Governance 2008;3:27-47.

16 IGZ. "Policy Plan 2012-2015. For justified trust in safe and appropriate care II" [Meerjaren beleidsplan 2012-2015. Voor gerechtvaardigd vertrouwen in verantwoorde zorg; in Dutch]. 2011.

17 Walshe K. Regulating Healthcare: A prescription for improvement? Philadelphia: Open University Press; 2003. p. 34.

18 IGZ. "Policy plan 2008-2011. For justified trust in safe and appropriate care" [Meerjaren Beleidsplan 2008-2011. Voor gerechtvaardigd vertrouwen in verantwoorde zorg; in Dutch]. 2007.

19 Luthi JC, McClellan WM, Flanders WD, Pitts S, Burnd-Hand B. Quality of health care surveillance systems: review and implementation in the Swiss setting. Swiss Med Wkly 2002;132:461-469.

20 Kollberg B, Elg M, Lindmark J. Design and Implementation of a Performance Measurement System in Swedish Health Care Services: A multiple case study of 6 development teams. Qual Manag Health Care 2005;14:95-111.

21 Pettersen IJ, Nyland K. Management and control of public hospitals – the use of performance measures in Norwegian hospitals. A case study. International Journal of Health Planning and Management 2006;21:133-149.

22 Lugtenberg M, Westert G. "Quality of health care and decision support-information for helping individuals to select health care. An international study on initiatives" [Kwaliteit van de gezondheidszorg en keuze-informatie voor burgers: een internationale verkenning van initiatieven; in Dutch]. 2007.

23 Tuijn SM, Janssens FJG, Van den Bergh H, Robben PBM. "Not all judgments are the same: a quantitative analysis of the interrater reliability of inspectors at the Dutch Health Care Inspectorate" [Het ene oordeel is het andere niet: Interinspecteursvariatie bij inspecteurs van de IGZ: een kwantitatieve analyse; in Dutch]. Nederlands Tijdschrift voor Geneeskunde 2009(8):322:326.

24 Tuijn SM, Van den Bergh H, Robben PBM, Janssens FJG. "The relationship between standards and judgments in the regulation of health care" [De relatie tussen normen en oordelen in het toezicht op de gezondheidszorg; in Dutch]. Tijdschrift voor Gezondheidswetenschappen 2009;6:264-271.

25 Feldt LS, Brennan RL. Reliability. In: R.L. Linn (Ed.), *Educational Measurement*. . Third ed. New York: Macmillan Publishing Company; 1989. p. 105-146.

26 Tuijn SM, Janssens FJG, Van den Bergh H, Robben PBM. Reducing interrater variability and improving health care: A meta-analytic review. JECP 2011;doi:10.1111/j.1365-2753.2011.01705.x:1-9.

27 Walshe K. Regulating Healthcare: A prescription for improvement? Philadelphia: Open University Press; 2003. p. 182.

28 Tuijn SM, Robben PBM, Janssens FJG, Van den Bergh H. Evaluating instruments for regulation of health care in the Netherlands. Journal of Evaluation in Clinical Practice 2011 DOI: 10.1111/j.1365-2753.2010.01431.x:411-419.

29 Nunnally JC. Psychometric Theory. New York: McGraw-Hill; 1978.

30 Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. The dependability of behavioral measurements : theory of generalizability for scores and profiles. New York: Wiley; 1972.

31 Heuvelmans APJM, Sanders PF. "Interrater reliability" [Beoordelaarsbetrouwbaarheid; in Dutch]. In: Eggen TJHM, Sanders PF, editors. "Psychometrics in practice" [Psychometrie in de praktijk; in Dutch] Arnhem: Cito; 1993. p. 468.

Chapter 7

General discussion

7.1 Overview

In this chapter we will answer the research questions, put forward some recommendations for policy and practice, discuss the methodological considerations, and make some suggestions for future research.

The main goal of this study was to identify and investigate possibilities for improving the reliability and validity of the regulatory judgments of IGZ inspectors by answering the following research questions:

1 Do IGZ inspectors systematically differ in the regulatory judgments they assign to similar health care institutions?
2 Do IGZ inspectors assign judgments to health care institutions that conform to the corporate standards and thus result in valid judgments?
3 Do the reliability and validity of the regulatory judgments of IGZ inspectors vary between two types of regulatory instruments?
4 Which interventions are effective for increasing the interrater reliability of professionals?
5 Which interventions are effective for increasing the reliability and validity of the regulatory judgments of IGZ inspectors?

The results of this study increased insight into factors that explain differences in the judgments of IGZ inspectors. We found that IGZ inspectors systematically differ in the regulatory judgments they assign to similar health care institutions (research question 1). We also found that IGZ inspectors systematically tend to assign judgments that are too positive compared with the IGZ corporate standards (research question 2).

We examined whether the reliability and validity of regulatory judgments varied between two types of instruments. We compared the lightly structured instrument (LSI) used for the regulation of hospital care with the highly structured instrument (HSI) used for the regulation of nursing home care. The results showed that with the LSI, the number of indicators discussed varied widely between inspectors. With the HSI, the average number of criteria discussed varied less, and the HSI criteria that were not discussed were generally the same ones. The results indicated problems with the reliability and validity of the judgments assigned with the HSI; however, reliability and validity could not be calculated with the LSI. The results showed that using an HSI is preferable to using an LSI (research question 3).

We performed a systematic meta-analytic review of the research literature to analyze the interventions professionals carry out to improve reliability. We found that three types of interventions could be defined: improving the diagnostic instrument, training the professional, and a combination of both. On average, although all types of interventions are effective, improving the diagnostic instrument seems to be the most effective; espe-

cially in the case of highly technical instruments, improvement has proven to be very effective. Because instrumental variables constitute a major source of error, improving the instrument is an important approach. However, our review offers solid arguments that can complement the literature and practice, with a focus on training the user of the instrument (research question 4).

We performed an experimental study to determine what kind of intervention would be effective for improving the reliability and validity of the regulatory judgments of IGZ inspectors. We set up a case study to examine the effect of participating in a consensus meeting and the effect of improving the regulatory instrument (research question 5).

The results showed that when an HSI was used, participating in a consensus meeting improved both the reliability and the validity of the regulatory judgments. Adjusting this instrument influenced but did not improve the reliability and validity of the judgments. This means that changing the instrument without training the inspectors in the use of the adjusted instrument does not improve the reliability and validity of the judgments. These outcomes emphasize the importance of the human factor in explaining variance between inspectors, and highlight the significance of training inspectors in the use of regulatory instruments. We calculated the effect of increasing the number of inspectors per case. As we expected, this increased the reliability and validity of the regulatory judgments. However, this increase can be expected when the conditions are similar to the conditions of the case study: when they examined the cases online, the inspectors were not allowed to talk to each other and discuss the cases or their scores. This was an important precondition for obtaining judgments that were as impartial as possible. Nevertheless, it seems unrealistic to expect that, when visiting in pairs or teams, inspectors will not discuss their observations and the argumentations for their judgments with each other. Therefore, the increase in reliability calculated under the conditions of the case study can be interpreted as the minimum increase in reliability. After all, it was shown that discussing cases and scores during the consensus meeting achieved an increase in reliability.

The same effect can be expected when inspectors visit in pairs or teams and discuss their observations and argumentations for scores. Yet, it is important to note that the expected increase in reliability when inspectors visit in pairs or teams is no guarantee of an increase in the validity of the judgments. When inspectors visit institutions in pairs or teams, there is a risk of unwanted side effects. Although agreement about a judgment between pairs of inspectors or between inspectors within a regulatory region can increase as a result of the dynamics in pairs of inspectors (for example, factors like dominance, seniority, status, and the ability to argue [1]), this does not necessarily imply that the judgment is valid. Visiting in pairs or teams at the same time is only one ex-

ample of using higher numbers of inspectors per regulatory visit. However, it is important to take into account what is best in actual practice, and what makes for a well-considered decision.

7.2 Implications for policy and practice within the professional context and the organizational context

In this study we investigated the reliability and validity of regulatory judgments of IGZ nursing home care inspectors. This increased our insight into the factors that influence the reliability and validity of the regulatory judgments of these inspectors, and have implications for practice. However, applying research results to practice is only possible when the field is ready to do so [2]. We will discuss the implications for practice within two contexts: the professional context and the organizational context. Moreover, the outcomes of our studies give some strategies for further professionalization of regulation. These will be discussed within these two contexts as well.

7.2.1 Implications within the professional context

In the literature, much has been written about how professionals learn. For example, Schön introduced the theory of reflection-in-action [3]. This theory implies that the professional uses his or her tacit knowledge in situations of uncertainty, instability, uniqueness, and value conflicts that he or she encounters in everyday practice (by means of knowing-in-action and reflection-in-action) [3]. Furthermore, the professional reflects on his or her own actions when he or she is not actively solving a problem in everyday practice (reflection-on-action) [4]. In Schön's theory, the professional acquires knowledge in an implicit manner in daily practice, while he or she learns in an explicit way by reflecting on daily practice [4]. These points of origin also apply to IGZ inspectors, because they encounter situations of uncertainty, instability, uniqueness, and value conflicts in everyday practice and use their tacit knowledge to deal with these situations. To be able to reflect on their actions, their interpretation of the regulatory observations, and the accompanying regulatory judgments, it is important that the inspectors share their experiences and ideas.

Facilitating such sharing of experiences and ideas is essential for several reasons. First, the IGZ inspectors have discretionary power, which is one of their characteristics. This gives them the opportunity to make their judgments on health care on their own, and can imply that they deviate from regulatory instruments or standards. This characteristic is essential for preventing regulation from becoming static. However, this discretionary power has to be used wisely so that a (regulatory) decision can be reproduced.

Second, inspectors in this system of risk-based supervision visit only a selection of institutions at risk, and to a large extent visit institutions that are identical in terms of risk criteria. This makes great demands not only on the regulatory instruments, but on the inspectors as well. After all, this selection means that inspectors have to distinguish between institutions that have characteristics similar to the standards developed by the IGZ for evaluating health care institutions. To be able to make reliable and valid judgments, it is important that IGZ inspectors have the skills to deal with both the gap between the instrument and reality and with discretionary power. Therefore, training in these competencies is essential. The importance of this is emphasized by the outcomes of our systematic review and case study (Chapters 5 and 6). Education is also important because the longer it has been since they completed their training programs, the more the professionals individualize their decision processes [5,6]. Consequently, compulsory continuous education is important for professionals regardless of discipline [6].

The outcomes of the case study (Chapter 6) show that adjusting the regulatory instrument alone does not result in higher reliability of the judgments of IGZ inspectors. With respect to the inspectors' discretionary power they should use, we would recommend that they use the instruments they have at their disposal. And when necessary, they can deviate from the instrument. However, they should be able to account for any regulatory decision. This means that there should be grounds for the regulatory judgments, they should be reproducible, and explanations should be given for any deviations. Thorough training in the use of regulatory instruments (including dealing with the gap between the instrument and reality) may contribute to higher reliability. Organizing a consensus meeting with compulsory attendance proved to be an effective method for improving the reliability and validity of the inspectors' regulatory judgments. Moreover, this meeting proved successful in launching a discussion on sources of variation. The reliability of the judgments can be further improved by increasing the number of inspectors per regulatory visit.

Interventions to improve the reliability and validity of IGZ inspectors' regulatory judgments may contribute to the further professionalization of the regulation of health care. These interventions should focus on monitoring and improving the reliability and validity of the judgments. As part of one of the interventions, inspectors are required to participate in one or more consensus meetings to improve the reliability of the regulatory judgments. Inspectors from all of the regulatory regions are represented in these consensus meetings. Examining validated cases should be part of the meetings. In addition, the possibility of allowing inspectors to visit institutions in pairs or in teams could be considered, while at the same time taking the implications for the institutions into account.

Along with facilitating reflection-on-action by organizing consensus meetings, the outcomes of our study indicate that the level of structure of regulatory instruments and how they are used play an important part in arriving at reliable and valid judgments. The expert meeting we organized to elicit critical remarks on the regulatory instrument resulted in remarks that, although they differed in focus, were all of a methodological nature. Some concerned the correspondence between the aspects of risk and the IGZ standards, while others focused on the need to place a check mark the aspects of risk or the lack of a weighting scheme in the instrument. The instrument's validity was discussed as well: some inspectors focus mainly on the aspects of risk presented in the instrument, while others also make tactical choices and involve the context when they arrive at a judgment. Furthermore, some inspectors appeared to be dissatisfied with the semantic labeling of the four categories of judgments, and stated that they never assign the score "no risk" or "very high risk." When inspectors include standards other than those described in the instrument to arrive at their judgment, it becomes hard to arrive at reliable and valid judgments. With risk-based supervision, inspectors visit only a selection of health care institutions that are considered to be at risk, which makes it necessary to make very accurate measurements to be able to reveal small differences between these institutions. Whether regulatory instruments that use a four-point scale are sufficiently accurate for this complex task could be a matter of debate [7]. We believe that consensus among the inspectors on the standards and the semantic labeling of the categories of judgments is a valuable complement to this approach. After all, the inspectors' compliance with the regulatory instrument would seem to be a precondition for successful implementation.

When an evaluation of indicators or standards indicates that they do not appear to be valid or do not distinguish between health care institutions, we recommend that they should be replaced. By continuously monitoring the regulatory standards and criteria, it is possible to fine-tune the instruments when necessary. Moreover, continuous education in the use of the regulatory instruments may prevent inspectors from excessively individualizing their regulatory decision process. Implementing an accreditation program that requires inspectors to obtain a minimum number of accreditation points every year may facilitate compliance with the instruments and promote uniform decision making.

In addition, when combining both political and methodological requirements, we would recommend making use of methodological counseling for the development of every regulatory instrument [8]. The use of HSIs optimizes the likelihood that similar institutions will be examined in the same way. Furthermore, using HSIs makes it possible to evaluate and improve the interrater reliability and validity of regulatory judgments.

7.2.2 Implications within the organizational context

Continuous improvement implies continuous transformation, a characteristic of a learning organization. A learning organization is defined as one that facilitates the learning of its members and continuously transforms itself [9]. Learning organizations develop as a result of the pressures facing modern organizations, and this enables them to remain competitive in the business environment [10]. Governmental regulatory authorities continuously have to adapt to new circumstances and must also be able to deal with the new requirements of today's society. According to Senge, a learning organization has five main characteristics: systems thinking, personal mastery, mental models, a shared vision, and team learning [11]. Using the outcomes of our study, we will argue why the characteristics of the learning organization offer opportunities for the further professionalization of regulatory authorities.

The concept of the learning organization is a theoretical framework that permits people to examine an organization as enclosed components within the system as a whole [11]. This framework was developed to make full patterns clearer and to help us see how to change them effectively [11]. Learning organizations use this method of thinking when they study their company and have information systems that measure the performance of the entire organization and its different components [12]. The monitoring and improvement of the reliability and validity of the judgments can be considered to be a component of the performance of the IGZ. The monitoring and, when necessary, improvement of the regulatory judgments can be realized by means of an information system that offers the possibility of measuring performance in controlled circumstances: online case studies facilitated by web-based surveys. To be able to decide whether the reliability and validity are acceptable, performance criteria (such as a standard for interrater agreement and validity of regulatory judgments) must be developed. In many studies on the reliability of health care professionals, kappa values are used to express the level of agreement. The interpretation of these values is often done according to the guidelines proposed by Landis and Koch [13]. As stated in these guidelines, kappa values of 0.00-0.20 indicate slight agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, and 0.61-0.80 substantial agreement. Values over 0.80 are considered almost perfect agreement. Insight into the target standards of other regulatory authorities can be helpful for determining a standard for the IGZ. For example, the Dutch Inspectorate of Education has developed a program to monitor the reliability and validity of its regulatory judgments, which includes target standards [14]. However, the method of choice for expressing a level of agreement and validity depends on the methodological design as well.

In a learning organization, personal mastery refers to the commitment by an individual member of the organization to the process of learning [11]. If a person does not want to learn, trying to educate this person will be a mission impossible. The willingness to learn is an important precondition for successful education, especially because members of an organization have assumptions, or what are referred to as mental models [11]. For some purposes it can be necessary to aim for consistent mental models between members of an organization. This might imply that some members of an organization have to adjust their mental models. Without a willingness to learn, this becomes very difficult. For example, in this study, the outcomes of the consensus meeting showed that the frame of reference – which can be referred to as a mental model – differs between inspectors: Some inspectors consider a regulatory visit to be a snapshot, while others consider it to be one moment in long-term developments. To develop towards becoming a learning organization, these models have to be discussed [11]. Furthermore, it is important to develop a shared vision that motivates staff members to learn as they create a common identity that provides focus [11]. To improve the reliability and validity of regulatory judgments, the aim should be to achieve uniformity of these models and a consistent vision among inspectors.

A learning organization has been described as the sum of individual learning, but there must be mechanisms for transmitting individual learning so that it can become organizational learning, or what is referred to as team learning [15]. Therefore, it is necessary for individual members of the organization to participate in dialogue and discussion [10]. In creating a learning environment it is essential to encourage an open culture [16] that promotes asking questions and encourages trust [10]. In our study, discussing sources of variation in a consensus meeting proved to be an effective intervention for improving the reliability and validity of regulatory judgments. A consensus meeting organized by an independent chairperson who keeps an eye on the processes within the group seems to create optimal preconditions for an open atmosphere and trust, and therefore for team learning. The benefit of team learning is that staff members grow more quickly, and the problem-solving capacity of the organization is improved through better access to knowledge and expertise [10,16].

Learning organizations have structures that facilitate team learning which feature boundary crossing and openness [12]. For the IGZ, these structures can be further developed to stimulate interchange of knowledge within and between programs, and to develop further towards becoming a learning organization. This is characterized by excellent knowledge management structures, and allows creation, acquisition, dissemination, and implementation of the acquired knowledge in the organization [15] – and equally important, to maintain the acquired knowledge in the organization. Helping to establish academic collaborative centers, as the IGZ did in

2011, is an example of working towards bringing the worlds of research and regulatory practice closer together, and facilitating team learning within a strong knowledge infrastructure [17,18].

The outcomes of our study shows that reliability and validity issues are a cause for concern in the regulation of health care, and that organizing consensus meetings is an effective intervention for improving reliability and validity. The implementation of monitoring and improving regulatory judgments should be incorporated into the internal processes of regulatory authorities. The monitoring and improvement of the reliability and validity of the regulatory judgments should be part of the internal quality system, and should have a company-wide design. This implies that it would affect all health care sectors represented within the regulatory authority. A performance standard for the reliability and validity of regulatory judgments will aid both the regulatory process and the process of monitoring the reliability and validity of regulatory judgments. Inspectors should only go for regulatory visits if the performance standard for reliability and validity (which has been determined in advance) has been met. The coordination of the monitoring and improvement of regulatory judgments should be closely related to the regulatory process. Experience with data management, methodological expertise, and expertise in the development of regulatory instruments are necessary requirements for developing a system for continuous monitoring and improvement of judgments. Close cooperation between the experts on methodology, data structure, web-based surveys, and regulatory education is also a premise for success.

7.3 Methodological considerations

As with all research, our study was characterized by methodological strengths and limitations. The methodological strengths will be discussed first.

7.4 Methodological strengths

The Hawthorne effect is notorious for its effect on research outcomes. This effect was first reported when different methods of increasing productivity in the Western Electrical Company's Hawthorne Works were examined in the 1920s and 30s. These studies found that no matter what change was introduced to working conditions, they all led to increased productivity. For example, increasing or reducing the lighting in the production area being studied had similar results [19]. This effect has been defined as "an increase in worker productivity produced by the psychological stimulus of being singled out and made to feel important" [20]. This definition was then broadened to include treatment response rather than productivity [21]. In our study we adapted the

definition to include reliability and validity: an increase in the reliability and validity of regulatory judgments produced by the psychological stimulus of being singled out and made to feel important.

Because of the bias that can be caused by the Hawthorne effect, it can be difficult to explain outcomes of research. The systematic variance between inspectors in our study was shown by analyzing judgments assigned in actual regulatory visits as well as judgments assigned in the case study. Because the judgments that we examined were assigned during actual regulatory visits, inspectors were not aware that these judgments would be used for research, and it is unlikely that the Hawthorne effect had an affect on the judgments and outcomes of our analysis. The systematic variance we found is not likely to have been biased. Moreover, the ecological validity of these data is good.

However, in the case study in which we examined the effect of two interventions on the reliability and validity of regulatory judgments, inspectors were aware that their judgments were going to be used for research purposes. Therefore, the Hawthorne effect might have affected the results. Even so, if this effect was indeed present, it was present during both the pretest and the posttest, and in all conditions of this experiment: because in both the pre-test and the post-test inspectors were aware they were participating in an experiment and that their judgments were going to be examined. Therefore, it seems unlikely that the effect we found as a result of the consensus meeting was a result of the Hawthorne effect.

In the case study, we used an experimental design to examine the cases. The validity of cases is important. The best test for validity compares the results of a measurement process with a "gold standard" [22]. To develop such a standard, three inspectors validated the cases. In addition, we attempted to limit the recall effect and the learning effect as much as possible by developing very similar but not completely identical cases for the first and second measurements, and by planning six weeks between the first and second measurements.

Furthermore, inspectors examined the cases independently using an online web-based survey. This technique made it possible to prevent inspectors from returning to an earlier case once they had judged it. In this manner, we attempted to minimize the possibility of comparing cases, and to encourage inspectors to rely on the instrument as much as possible. Because inspectors examined the cases at different locations, we minimized the possibility of discussing the cases among themselves.

7.5 Methodological limitations

In all chapters, we treated ordinal data like discrete data. In other words, we converted the semantic categories "absent," "present," "operational," and "fulfilled" and the categories "no risk," "small risk," "high risk," and

"very high risk" into a numeric score from one to four, which resulted in the creation of an interval scale. It was difficult to interpret the calculated numeric means because it did not fit into one of the four semantic categories. However, the method we used to transform ordinal data into discrete data is one frequently used for analyzing ordinal data, and generally results in only very minor distortions [23].

In our study we examined both the judgments of inspectors assigned to nursing homes during actual regulatory visits and judgments assigned to cases. To be able to compare the judgments that were assigned during on-site visits, we had to statistically correct for characteristics of nursing homes later on to homogenize the nursing homes as much as possible. By performing this statistical correction, we attempted to approach the desired situation in similar nursing homes as much as possible. However, we are aware that this correction only approximates reality at best.

Moreover, we examined the reliability and validity of regulatory judgments assigned with an LSI and with an HSI in relation to accountability in the Netherlands. Both instruments are used in different health care sectors: the LSI is used for the regulation of hospital care, and the HSI is used for the regulation of nursing home care. It would have been better to compare two types of instruments within the same health care sector. However, in the regulation of health care in the Netherlands, the regulatory instruments vary per health care sector, which meant it was not possible to compare two types of instruments within the same sector. In addition, the instrument used in the period 2005/2006 was still under construction while the inspectors were using it during their regulatory visits. Despite the fact that only very minor changes were made and the criteria used for regulation remained unchanged, these minor changes could have affected the results.

In the systematic review of the literature on interventions to improve interrater reliability, we performed a meta-analysis on the data. We pooled the data that differed in design and setting, and this is unusual. Because improving interrater variability applies to a wide variety of medical and paramedical decision-making settings, the inclusion of a broad range of studies in this review adds to the validity of the study, as we described a phenomenon that is present in all medical and paramedical professions. Although this approach was innovative, it was, nevertheless, necessary for investigating general interrater variability.

Furthermore, the systematic review was based on a sensitive search strategy, and we believe it to be unlikely that any study we may have overlooked would have changed our conclusion. However, we acknowledge that this review is subject to selection bias, because studies with negative results are published less often.

In the case study we performed to examine the effects of two interventions on interrater reliability, we used judgments assigned to validated cases. We are aware that the use of cases, no matter how well developed,

differs from on-site visits. Even though the use of cases or vignettes to examine interrater reliability is very common, this might have affected the results. After all, no matter how well designed the cases may be, they can never completely duplicate the complexity of reality.

7.6 Recommendations for future research

Would it be legitimate to presume that if regulatory judgments were completely reliable and valid, regulation would be utterly effective? Reliable and valid judgments are indeed very important requirements. However, it is worth asking whether additional factors might provide valuable complements to these preconditions. To start with, the outcomes of this study give reason to assume that reporting on regulatory visits is an important factor.

In this study, we only examined the effect of two types of adjustments to the regulatory instrument. It seems fair to assume that out of all of the possible adjustments that can be made, some of them will result in better agreement and higher validity. It might be debatable whether a four-point scale used in risk-based supervision is optimal for assigning regulatory judgments to institutions that resemble each other to a large extent. To examine this, research needs to be conducted on the effect of a scale with more than four categories on the reliability and validity of judgments.

Future research on the optimal conditions for visits to health care institutions by larger numbers of inspectors would be a valuable continuation of this study. This could give more insight into the effect of visiting in pairs or teams, and the possible side effects of these arrangements.

The outcomes of our case study suggest that discussing the considerations for arriving at a judgment and also the sources of variation results in a change in the cognitive process that underlies decision making. Although we did not examine the change that was made, it would seem that parts of the mental maps were modified in the black box of the cognitive considerations that underlie any kind of decision making, and that this resulted in higher reliability and validity [24]. In this study, we did not examine the effect of the intervention on the cognitive process that underlies decision making. We focused on examining its effect on the reliability and validity of regulatory judgments. However, to better understand the mechanism of transformation, this would be worth investigating. We also investigated the effect of a single consensus meeting on the reliability and validity of judgments. Earlier research on the effect of more than one consensus meeting showed that the reliability of health care professionals continued to improve after every meeting [25]. It could be valuable to examine whether this effect can be generalized to inspectors as well. Furthermore, we focused on the effect of a consensus meeting in this study, and did not convert the outcomes of the consensus meeting into conventions or guidelines

that had to be employed. To gain further insight into effective interventions, it could be valuable to study the effect of the employment of conventions that were the result of a consensus meeting. In addition, we could increase our understanding of effective interventions by studying more types of interventions, for example, peer review. This type of review is often used by professionals, and it could be helpful to examine its effect on the reliability and validity of judgments.

In this study, we examined the regulatory judgments within the IGZ's system of risk-based supervision. However, the monitoring and improvement of regulatory judgments can also be applied to other forms of regulation, such as theme-based regulation, regulation in response to calamities, or government regulation that employs monetary fines. It is important to examine the reliability and validity of regulatory judgments within other regulatory systems as well.

Would regulation be perfect if the indicators were completely reliable and valid, if there were no systematic differences in the definitions of indicators and if the collection and coding of data were fully homogenous across institutions, if the regulatory judgments were completely reliable and valid, and if the regulatory reports were always written and conformed to the corporate standards? Of course, this is what is aimed for. However, regulation is an interactive process. This implies that the face-to-face feedback on the regulatory findings after the regulatory visit can be of importance as well. Earlier research has shown that even small differences in the formulation of questions can have significant effects in interviews [26]. As a result of the interactive character of regulation, the way in which regulatory judgments are dealt with depends in part on the institutions themselves. When criteria have risk scores that are high or very high, the IGZ requires measures to be formulated to improve health care on the criterion or criteria concerned. Actually putting these intentions down on paper is the responsibility of the institutions themselves.

7.7 References

1 Heuvelmans APJM, Sanders PF. "Interrater reliability" [Beoordelaarsbetrouwbaarheid; in Dutch]. In: Eggen TJHM, Sanders PF, editors. "Psychometrics in practice" [Psychometrie in de praktijk; in Dutch] Arnhem: Cito; 1993. p. 468.

2 Ashton RH, Asthon AH. Judgment and decision making research in accounting and auditing. Cambridge: Cambridge University Press; 1995. p. 23.

3 Schön DA. The reflective practitioner. How professionals think in action. London: Maurice Temple Smith Ltd; 1983.

4 Hutschemaekers GJM, Plochg T. "Research on professionals" [Onderzoek naar professionals; in Dutch]. In: Plochg T, Juttmann RE, Klazinga NS, Mackenbach JP, editors. "Manual for health care research [Handboek gezondheidszorgonderzoek; in Dutch]. Houten: Bohn Stafleu van Loghum; 2007. p. 263.

5 Molyneux PD, Miller DH, Filippi M, Yousry TA, Radü EW, Adèr HJ, et al. Visual analysis of serial T2-weighted MRI in multiple sclerosis: intra- and interobserver reproducibility. Neuroradiology 1999;41(12):882-888.

6 Zhang GG, Singh B, Lee W. Improvement of Agreement in TCM Diagnosis Among TCM Practitioners for Persons with the Conventional Diagnosis of Rheumatoid Arthritis: Effect of Training. Journal of Alternative and Complementary Medicine 2008;14(4):381-387.

7 Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. 3rd ed. New York: Oxford University Press; 2003. p. 38.

8 Walshe K. Regulating Healthcare: A prescription for improvement? Berkshire: Open University Press; 2003. p. 191.

9 Pedler M, Burgogyne J, Boydell T. The Learning Company: A strategy for sustainable development. London: McGraw-Hill.; 1997.

10 O'Keeffe T. Organizational Learning: a new perspective. Journal of European Industrial Training 2002;26:130-141.

11 Senge PM. The Fifth Discipline. London: Century Business; 1990.

12 Argyris C. On Organizational Learning. Oxford: Blackwell Publishing; 1999.

13 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

14 Janssens FJG. "From research to evaluation. The methodology of the Dutch Inspectorate of Education" [Van onderzoek naar evaluatie. De methodologie van de Onderwijsinspectie; in Dutch]. 1997.

15 Wang CL, Ahmed PK. Organizational learning: a critical review. The Learning Organization 2003;10:8-17.

16 McHugh D, Groves D, Alker A. Managing learning: what do we learn from a learning organization? The Learning Organization 1998;5:209-220.

17 Health Council of the Netherlands "Towards evidence based regulation. Research on the effects of regulation by the Dutch Health Care Inspectorate" [Op weg naar evidence based toezicht. Het onderzoek naar effecten van toezicht door de Inspectie van de Gezondheidszorg; in Dutch]. 2011;ISBN: 978-90-5549-838-3.

18 Scientific Council for Government Policy (WRR). " Overseeing the Public Interest. Towards a broader perspective on government regulation" [Toezien op Publieke Belangen. Naar een verruimd perspectief op rijkstoezicht; in Dutch]. 2013.

19 Mayo E. The human problems of an industrial civilization. New York: Macmillan; 1933.

20 Franke RH, Kaul JD. The Hawthorne experiments: First statistical interpretation. Am Sociol Rev 1978;43:623-643.

21 McCarney R, Warner J, Iliffe S, van Haselen R, Griffin M, Fisher P. The Hawthorne Effect: a randomised controlled trial. BMC Medical Research Methodology 2007:1-8.

22 Nunnally JC. Psychometric Theory. New York: McGraw-Hill; 1978.

23 Bergh van den H, Zwarts M, Peter-Sips M. "Quality of the educational learning process". [Kwaliteit van het onderwijsleerproces; in Dutch]. Tijdschrift voor Onderwijsresearch 2000;25(1/2):20-39.

24 Argyris C, Schön D. Theory in practice. Increasing Professional Effectiveness. Oxford, England: Jossey-Bass; 1974.

25 Tsuda H, Akiyama F, Kurosumi M, Sakamoto G, Watanabe T. The efficacy and limitations of repeated slide conferences for improving interobserver agreement when judging nuclear atypia of breast cancer. The Japan National Surgical Adjuvant Study of Breast Cancer (NSAS-BC) Pathology Section. Jpn J Clin Oncol 1999;29(2):68-73.

26 Tuijn SM, Janssens FJG, Van den Bergh H, Robben PBM. "Not all judgments are the same: a quantitative analysis of the interrater reliability of inspectors at the Dutch Health Care Inspectorate". [Het ene oordeel is het andere niet: Interinspecteursvariatie bij inspecteurs van de IGZ: een kwantitatieve analyse;[In Dutch]. Nederlands Tijdschrift voor Geneeskunde 2009(8):322:326.

Summary

This study examined the reliability and validity of regulatory judgments within the system of risk-based supervision. This research describes the correspondence between regulatory judgments, and provides insight into the extent to which health care inspectors assign similar judgments to similar situations (reliability) and whether these judgments correspond with the standards developed by the regulatory authority (the Dutch Health Care Inspectorate, IGZ) for its regulatory task (validity). This study examined which interventions are effective for improving the reliability and validity of regulatory judgments. Monitoring and improving the reliability and validity of the judgments can be considered to be a component of the overall performance of the IGZ.

8.1 General introduction

This dissertation starts by defining interrater reliability and validity of judgments. Reliable and valid judgments are important in the regulation of health care. Based on the judgments of their inspectors, the IGZ asks health care institutions to improve the quality of the care they deliver when necessary. If the improvements are not satisfactory, the IGZ can impose administrative sanctions and initiate penal measures. When regulatory judgments are not reliable, institutions with similar characteristics may be judged differently. When this happens, it is hard to explain why some institutions have to improve the quality of their care while others with similar characteristics, do not have to improve their quality of care.

However, it is not only the reliability of regulatory decisions that is important – it is equally important that these decisions be valid. When regulatory judgments are not valid, even though inspectors might all assign the same judgment to institutions with similar characteristics, this judgment will not correspond with the regulatory authority's corporate standards. In the case of false-positive judgments, there is the risk that institutions will not be asked to improve their care, while in fact this should have happened.

Interrater reliability has been discussed since the seventeenth century, and the subject is a common one in a variety of professions. The concept of observer error has been studied extensively in the fields of education, medicine, medical insurance science, penal regulation, and accounting and auditing.

The research questions of this thesis are:

1 Do IGZ inspectors systematically differ in the regulatory judgments they assign to similar health care institutions?

2 Do IGZ inspectors assign judgments to health care institutions that conform to the corporate standards and thus result in valid judgments?

3 Do the reliability and validity of the regulatory judgments of IGZ inspectors vary between two types of regulatory instruments?

4 Which interventions are effective for increasing the interrater reliability of professionals?

5 Which interventions are effective for increasing the reliability and validity of the regulatory judgments of IGZ inspectors?

8.2 Not all judgments are the same

This chapter describes the analysis of the interrater reliability of the regulatory judgments of nursing home care inspectors. These judgments were assigned to criteria for nursing home care in 2005/2006. The regulatory instrument consisted of criteria for examining the quality of care. These criteria were a combination of measurements of structure, processes, and outcomes. One of these criteria was "pressure ulcers." For this criterion, inspectors assessed whether the prevalence of pressure ulcers is recorded by the staff (process) as well as whether the staff has a protocol for pressure ulcers (structure). During regulatory visits, inspectors examined the quality of care using these criteria, and assigned scores to the criteria on a four-point scale: "absent," "present," "operational," or "fulfilled." The regulatory instrument describes exactly which judgment applies in which situation.

The results indicated that inspectors' regulatory judgments vary when examining institutions: institutions with similar characteristics with regard to health care indicators are judged differently. Moreover, inspec-

tors have to provide grounds for their judgments. The presence of grounds for the judgments seems to depend on both the individual inspector and the judgment assigned. Some inspectors provide grounds for their judgment while others do not. Moreover, compared with negative judgments, grounds are provided for positive judgments less often. Suboptimal interrater agreement is a cause for concern in the regulation of nursing home care. The next step in this research would be to gain insight into the level of stringency of the regulatory judgments. This could clarify whether the validity of the judgments could also be considered a source of variation.

8.3 The relationship between the employment of standards and judgments

This part of the dissertation describes the analysis of the validity of regulatory judgments on nursing home care. Judgments and the grounds for such judgments were selected for four criteria: “pressure ulcers”, “sufficient help with eating and drinking”, “continuous supervision in living rooms,” and “the extent of care needed.” We analyzed the extent to which the argumentations contained in the grounds for the judgments corresponded with the IGZ regulatory standards. We also studied the extent to which the actual judgments corresponded with the judgments that should have been assigned based on the arguments presented and the strict employment of the IGZ standards (corporate judgments). The results indicated that inspectors do not always formulate their judgments according to the corporate standards. About half of the analyzed judgments were too positive compared with the judgments that would have been assigned if the corporate standards had been strictly employed. Although the percentage of false-positive judgments depended on the criterion being judged, they were assigned by all inspectors.

These findings provide insight into the validity of the regulatory judgments: the correspondence between the judgments assigned and the corporate standards. The results indicated there are problems with both the reliability and the validity of these judgments. The type of regulatory instrument varies between health care sectors within the IGZ. The next step in this research will be to gain insight into the relationship between the types of regulatory instruments and the reliability and validity of regulatory judgments.

8.4 Not all instruments are the same

During this part of the research, we studied the reliability and validity of regulatory judgments assigned with two different types of regulatory instruments. Judgments assigned using a highly structured instrument (HSI) for the regulation of nursing home care were compared with the judgments assigned using a lightly structured instrument (LSI) for the regulation of hospital care in the Netherlands. An HSI consists of a non-variable set of

criteria that are examined and scored (judged) during every regulatory visit to a nursing home. An HSI describes exactly when a judgment should be assigned. An LSI consists of a permanent set of indicators. If an institution has a deviant score for one of these indicators (the indicator contains a warning signal), this indicator should be discussed during a regulatory visit. The LSI does not describe exactly when a judgment should be assigned.

The results showed that with the LSI, the number of indicators discussed varied widely between inspectors, and reliability and validity could not be calculated. Not enough data were available to compare institutions with similar characteristics. In contrast to the LSI, the average number of criteria discussed using the HSI varied less, and the criteria that were not discussed were generally the same ones. There was no relationship between the presence of a warning signal in an indicator and a discussion of that indicator during a regulatory visit: more indicators without signals were discussed compared with indicators with signals. Inspectors select the indicators to be discussed at their own discretion. With the HSI, all of the criteria are discussed during regulatory visits. The results indicated that although there are problems with the reliability and validity of the judgments assigned with the HSI, at least the same set of criteria is used to compare all of the institutions. The results indicated that using an HSI is preferable because it makes it possible to account for regulatory decisions.

The results showed that using an HSI has limitations as well. Because of this, an HSI does not seem to be the only solution for improving reliability and validity. How do other professionals improve their interrater reliability? To answer this question, a systematic review of the scientific literature was performed.

8.5 Improving interrater reliability: a meta-analytic review

According to the literature on reliability, the central approach for improving reliability seems to improve the quality of the instrument. A systematic review and meta-analysis was performed to find out whether additional training of the raters could be a valuable complement to this approach.

Because interrater variability occurs in a wide variety of professions, we searched medical and sociological databases. The interventions were categorized into three groups: training of professionals, improving the diagnostic instrument, and a combination of training and improving the instrument.

The results of our searches contained only articles about interventions for improving reliability among health care professionals. No empirical studies were found on interventions for increasing reliability among other professionals, such as judges, teachers, or inspectors. The results indicated that the effect of the three types of interventions is significant for the three groups of interventions. However, improving highly technical instruments (like ct-scans) has the largest effect on agreement. It could be concluded that although all types of inter-

ventions are effective, improving the instruments seems to be most effective, especially when it concerns highly technical instruments.

This review suggests solid arguments that can complement the literature and practice, with a focus on training the user of the instrument. To gain insight into whether these outcomes can be generalized to IGZ health care inspectors, the next step in this research was to perform an experimental case study.

8.6 Improving interrater reliability and validity: an experiment

We used a case study to investigate the effect of two interventions on the reliability and validity of judgments of nursing home care inspectors: adjustment of the regulatory instrument for the regulation of nursing home care and participation of inspectors in a consensus meeting. Moreover, we explored the effect of an increase in the number of inspectors on the reliability and validity of regulatory judgments.

A randomized controlled trial was used to examine the effect of the adjustment of the regulatory instrument. A before and after case study was used to examine the effect of the consensus meeting. Inspectors were randomly assigned to two groups, and they examined cases with either the adjusted or the unadjusted instrument. The instrument was adjusted in two ways. First, we formulated the description of the aspects of risk positively rather than negatively. As a result, the descriptions of both the standard and the aspects of risk were formulated positively. Second, we made it mandatory to check off the aspects of risk.

In a consensus meeting, professionals come together to discuss cases and try to reach consensus about a judgment. Inspectors discuss a set of cases that they have to rank from “no risk” to “high risk.” To examine the effect of a consensus meeting, all nursing home care inspectors attended one. The purpose of this meeting was to identify common sources of variation. Therefore, the inspectors had to reach consensus about the order of two sets of four cases. After the consensus meeting, the inspectors examined cases that were very similar to – but not completely identical to – those used in the pretest to prevent learning effects from the cases used previously.

The results showed that the reliability and validity of the judgments was highest after the consensus meeting. The results of increasing the number of inspectors indicated that this increases both the reliability and the validity of the regulatory judgments. These calculations presume that inspectors assigned scores under the same conditions as in the case study: Inspectors do not talk with each other about their scores when examining the cases. However, it seems unrealistic to expect that, when visiting in pairs or teams, inspectors will not discuss their observations with each other. Therefore, it seems reasonable to expect that there will be a greater increase in the reliability of the regulatory judgments in actual practice (when inspectors do talk with each other

about their scores). Whether this expected increase in reliability and validity can be unconditionally generalized to daily practice could be examined in the future.

8.7 General discussion

The results of this study showed that the level of structure of regulatory instruments and the use of these instruments are important factors in arriving at reliable and valid regulatory judgments. However, focusing only on the instrument would seem to be too narrow. Continuous education in the use of the regulatory instruments may prevent inspectors from excessively individualizing their regulatory decision process.

What are the implications of this study for daily regulatory practice? Improvements are possible in both the professional and the organizational context. In the "reflection-in-action" theory, the professional acquires knowledge in an implicit manner in daily practice. In the "reflection-on-action" theory he or she learns in an explicit way by reflecting on daily practice. To be able to reflect on their actions, their interpretation of the regulatory observations, and the accompanying regulatory judgments, it is important that the inspectors share their experiences and ideas. "Reflection-on-action" can be facilitated by organizing consensus meetings.

Continuous improvement implies constant transformation, which is a characteristic of learning organizations. The method of thinking used by learning organizations offers opportunities for the IGZ as well: monitoring and improving the reliability and validity of the judgments can be considered a characteristic of an organization that aims to develop itself continuously. Within a learning organization there must be mechanisms for transmitting individual learning so that it becomes organizational learning or what is known as team learning. The presence of structures that facilitate team learning that feature boundary crossing and openness are important characteristics of learning organizations. In the Netherlands this has given rise to organizations like academic collaborative centers, which aim to bring the worlds of research and regulatory practice closer together and facilitate team learning within a strong knowledge infrastructure.

Samenvatting

In dit proefschrift is de betrouwbaarheid en validiteit van oordelen in het toezicht op de gezondheidszorg binnen het systeem van risicogestuurd toezicht onderzocht. Het onderzoek beschrijft in welke mate inspecteurs hetzelfde oordeel toekennen in gelijke situaties (de betrouwbaarheid van de oordelen) en in hoeverre deze oordelen overeenkomen met de standaarden die de Inspectie voor de Gezondheidszorg (IGZ) heeft ontwikkeld voor haar toezicht (de validiteit van de oordelen). Onderzocht is welke interventies effectief zijn om zowel de betrouwbaarheid als de validiteit van inspecteursoordelen te verbeteren. Het monitoren en verbeteren van de betrouwbaarheid en validiteit van inspecteursoordelen is een belangrijke component van het toezicht door de IGZ.

9.1 Inleiding

Dit proefschrift begint met de introductie van de betekenis van de beoordelaarsbetrouwbaarheid en validiteit van oordelen. Betrouwbare en valide oordelen zijn van groot belang in het toezicht. Op basis van oordelen van inspecteurs, moeten zorginstellingen – als dat nodig blijkt - verbetermaatregelen nemen om de kwaliteit van hun zorg te verbeteren. Als deze verbeteringen niet passend zijn, kan de IGZ maatregelen treffen. Als oordelen in het toezicht niet betrouwbaar zijn, worden vergelijkbare instellingen, verschillend beoordeeld. Het is dan moeilijk te verantwoorden waarom sommige instellingen hun zorg moeten verbeteren terwijl andere instellingen met vergelijkbare zorg dat niet hoeven te doen. Onder vergelijkbare omstandigheden moeten gelijke oordelen gegeven worden. Al sinds de 17^{e} eeuw wordt er aandacht besteed aan beoordelaarsbetrouwbaarheid in verschillende

beroepen. Het concept van beoordelaarsbetrouwbaarheid is uitgebreid onderzocht in bijvoorbeeld het onderwijs (van de Nederlandse taal), de (verzekerings)geneeskunde, bij de rechtspraak en bij financiële controle.

Het is niet alleen belangrijk dat oordelen in het toezicht betrouwbaar zijn, ook de validiteit van oordelen is essentieel. Als oordelen niet valide zijn, kennen inspecteurs hetzelfde oordeel toe aan instellingen met gelijke kenmerken, maar komt dit oordeel niet overeen met de standaarden van de toezichthouder. In het geval van vals-positieve oordelen, wordt er vergeleken met de norm een relatief te positief oordeel gegeven en bestaat het risico dat instellingen geen verbetermaatregelen hoeven te nemen om hun zorg te verbeteren, terwijl ze dit eigenlijk wel hadden moeten doen.

De onderzoeksvragen die ten grondslag liggen aan dit proefschrift zijn de volgende:

1 Verschillen inspecteurs van IGZ systematisch in hun oordelen over instellingen met gelijke kenmerken?
2 Komen de oordelen over instellingen van IGZ-inspecteurs overeen met de standaarden die IGZ voor haar toezicht hanteert ?
3 Heeft het type toezichtsinstrument invloed op de beoordelaarsbetrouwbaarheid en validiteit van inspecteursoordelen?
4 Welke interventies zijn effectief om de beoordelaarsbetrouwbaarheid van professionals te vergroten?
5 Welke interventies zijn effectief om de betrouwbaarheid en validiteit van oordelen van IGZ-inspecteurs te vergroten?

9.2 Het ene oordeel is het andere niet

Hoofdstuk twee beschrijft de analyse van de betrouwbaarheid van inspecteursoordelen over criteria van zorg in verpleeghuizen. Deze oordelen zijn toegekend in de dagelijkse toezichtspraktijk in 2005/2006. Het toezichtsinstrument dat de oordeelsvorming ondersteunt bestaat uit criteria waarmee de kwaliteit van zorg onderzocht wordt. Deze criteria zijn een combinatie van metingen op structuur-, proces- en uitkomstniveau op een aantal onderwerpen die worden beschouwd als indicator voor kwalitatief goede en veilige zorg. Een van deze criteria is ‘ doorligwonden’ (decubitus). Bij dit criterium onderzoeken inspecteurs of de aanwezigheid van doorligwonden wordt geregistreerd door het personeel (proces) en of het personeel de beschikking heeft over een protocol voor de preventie en het behandelen van decubitus (structuur). Tijdens toezichtsbezoeken onderzoeken inspecteurs de kwaliteit van zorg op basis van deze criteria en oordelen over de zorg op basis van deze criteria. Zij oordelen op een vierpuntsschaal: ‘afwezig’, ‘aanwezig’, ‘operationeel’ en ‘geborgd’. Het toezichtsinstrument schrijft precies voor wanneer welk oordeel in welke situatie van toepassing is.

Inspecteursoordelen over de kwaliteit van zorginstellingen lopen uiteen als inspecteurs instellingen onderzoeken: vergelijkbare zorg in instellingen wordt niet altijd op gelijk wijze beoordeeld. Het gebruikte toezichtsinstrument vraagt van inspecteurs een onderbouwing van hun oordeel. De aanwezigheid van onderbouwingen bij de oordelen blijkt zowel af te hangen van de individuele inspecteur als van de aard van het gegeven oordeel, dat wil zeggen of het negatief of positief is. Sommige inspecteurs onderbouwen hun oordelen, terwijl anderen dat niet doen. Positieve oordelen worden minder vaak onderbouwd dan negatieve oordelen. De beoordelaarsbetrouwbaarheid is niet optimaal in het toezicht op de zorg in verpleeghuizen door IGZ.

Het vervolgonderzoek is gericht op het verkrijgen van inzicht in de mate van strengheid van oordelen. Dit geeft antwoord op de vraag of de validiteit van de oordelen een verklaring is voor de gevonden beoordelaarsverschillen.

9.3 De relatie tussen standaarden en oordelen

Dit deel van het onderzoek beschrijft de analyse van de validiteit van oordelen in het toezicht op de zorg in verpleeghuizen. Oordelen en de bijbehorende onderbouwingen over de volgende vier criteria zijn onderzocht: 'decubitus', 'voldoende hulp bij eten en drinken', ' continue toezicht in woonkamers' en 'de mate waarin zorg nodig is'. Geanalyseerd is in welke mate de onderbouwingen van de oordelen overeenkomen met de standaarden van de IGZ voor het toezicht op de zorg in verpleeghuizen. Nagegaan is in welke mate de feitelijke oordelen overeenkomen met oordelen die gegeven zouden moeten worden bij strikte toepassing van de IGZ-standaarden. Het onderzoek laat zien dat het oordeel van de inspecteurs niet altijd conform de IGZ-standaarden is. Vergelijking van de gegeven oordelen met de standaarden van de IGZ leert dat ongeveer de helft van de geanalyseerde oordelen te positief is. Het percentage vals-positieve oordelen hangt af van het criterium dat is beoordeeld, maar alle inspecteurs kennen in meer of mindere mate vals-positieve oordelen toe.

Deze bevindingen geven inzicht in de validiteit van de inspecteursoordelen: de mate van overeenkomst tussen de gegeven oordelen en de standaarden van de IGZ. Zowel de betrouwbaarheid als de validiteit van de oordelen is niet optimaal.

Het type toezichtsinstrument van de IGZ varieert per zorgveld. De volgende fase van het onderzoek is gericht op het verkrijgen van inzicht in de relatie tussen het type toezichtsinstrument en de betrouwbaarheid en validiteit van inspecteursoordelen.

9.4 Het ene instrument is het andere niet

De analyse van de betrouwbaarheid en validiteit van inspecteursoordelen die met twee verschillende type toezichtsinstrumenten zijn toegekend, staat centraal in dit deel van het onderzoek. De oordelen die toegekend zijn met een hoog-gestructureerd instrument (HSI) dat gebruikt wordt in het toezicht op zorg in verpleeghuizen zijn vergeleken met de oordelen die toegekend zijn met een laag-gestructureerd instrument (LSI) dat gebruikt wordt voor het toezicht op ziekenhuizen. Een HSI bestaat uit een vast aantal criteria dat bij elk toezichtsbezoek beoordeeld wordt. In het HSI is precies beschreven wanneer welk oordeel over de criteria van toepassing is. Een LSI bestaat uit een vast aantal criteria of zogenaamde indicatoren. Wanneer een instelling afwijkend scoort op een van deze indicatoren (een instelling scoort bijvoorbeeld opvallend goed of opvallend slecht, of er is sprake van een bepaalde trend in gegevens over meerdere jaren), dan is er sprake van een signaal op de betreffende indicator en dan moet deze indicator tijdens een toezichtsbezoek besproken worden. In het LSI is niet beschreven wanneer welk oordeel van toepassing is.

Het onderzoek toont aan dat het aantal indicatoren dat inspecteurs bespreken in een toezichtsbezoek bij ziekenhuizen erg uiteenloopt met een LSI. De betrouwbaarheid en validiteit van de inspecteursoordelen die toegekend zijn met een LSI kunnen hierdoor niet berekend worden. Er zijn onvoldoende gegevens om te kunnen vergelijken tussen instellingen met gelijke kenmerken. Het gemiddeld aantal criteria dat besproken wordt tijdens het toezichtsbezoek in verpleeghuizen met het HSI varieert veel minder. In tegenstelling tot het LSI, waarbij de niet-besproken indicatoren steeds verschillen, zijn de niet-besproken criteria bij het HSI over het algemeen steeds dezelfde. Dit betekent dat instellingen die beoordeeld worden met een HSI, met dezelfde set criteria onderzocht worden.

De analyse van de oordelen gegeven met een LSI laat ook zien dat er meer indicatoren zonder signaal besproken zijn dan indicatoren met signaal: inspecteurs kiezen de indicatoren die zij bespreken in een toezichtsbezoek op basis van hun individuele professionele inschatting en niet op basis van een signaal. Dit in contrast met het HSI: hiermee worden zo goed als alle criteria besproken in toezichtsbezoeken in verpleeghuizen. De resultaten laten problemen zien in de betrouwbaarheid en validiteit van de oordelen die toegekend zijn met het HSI, maar in elk geval worden met het HSI alle instellingen langs dezelfde meetlat gelegd. Het gebruik van een HSI heeft daarom de voorkeur boven het gebruik van een LSI. Hiermee is het beter mogelijk verantwoording af te leggen over beslissingen in het toezicht.

Hoewel een HSI de voorkeur geniet boven een LSI, kent ook het gebruik van een HSI beperkingen in de betrouwbaarheid en validiteit van oordelen. Het gebruik van een dergelijk instrument is mogelijk niet de eni-

ge oplossing om de betrouwbaarheid en validiteit van inspecteursoordelen te verbeteren. Hoe verbeteren andere professionals hun beoordelaarsbetrouwbaarheid? Om deze vraag te beantwoorden is een systematische literatuurstudie uitgevoerd.

9.5 Kan de overeenstemming tussen oordelen worden bevorderd: een meta-analytische review

In de literatuur over beoordelaarsbetrouwbaarheid staat de verbetering van de kwaliteit van het instrument centraal. Een systematische literatuurstudie en meta-analyse zijn uitgevoerd om te onderzoeken of additionele training van de beoordelaars een waardevolle aanvulling van deze benadering is.

Omdat beoordelaarsbetrouwbaarheid in veel verschillende soorten beroepen een rol speelt, werd literatuur in zowel medische als sociaal-wetenschappelijke databases gezocht. De interventies zijn in drie groepen gecategoriseerd: training van de professionals, verbeteren van het diagnostische instrument en een combinatie van training en het verbeteren van het diagnostische instrument. Er zijn uitsluitend artikelen over interventies om de beoordelaarsbetrouwbaarheid van (para)medische professionals te verbeteren gevonden. Er zijn geen empirische studies over interventies om de beoordelaarsbetrouwbaarheid van andere professionals zoals rechters, docenten of inspecteurs te verbeteren, gevonden.

Het effect van de drie soorten interventies (aanpassen van het instrument, training van beoordelaars en de combinatie van beiden) is significant. Het verbeteren van (technische) instrumenten heeft het grootste effect op de beoordelaarsbetrouwbaarheid, maar ook training vergroot de overeenstemming tussen beoordelaars. Twee van deze drie interventies zijn vervolgens onderzocht in een experimentele casusstudie onder IGZ-inspecteurs.

9.6 Kan de betrouwbaarheid en validiteit van oordelen worden bevorderd: een experiment

In een experimenteel opgezette casusstudie is het effect van twee interventies op de betrouwbaarheid en validiteit van inspecteursoordelen over zorg in verpleeghuizen onderzocht: aanpassing van het toezichtsinstrument en deelname van inspecteurs aan een consensusbijeenkomst. Ook is het effect nagegaan van het aantal oordelende inspecteurs op de betrouwbaarheid en validiteit van de oordelen.

Om het effect van het aanpassen van het toezichtsinstrument te onderzoeken, is een gerandomiseerd design met een controlegroep gebruikt. Hierbij is de toewijzing van de inspecteur aan één van de twee groepen aselect (door het lot) bepaald. De ene groep bespreekt en beoordeelt de casussen met het ongewijzigde instru-

ment (de controlegroep), de andere groep met het aangepaste instrument. Het instrument is aangepast op twee punten: de beschrijving van de risicoaspecten is positief geformuleerd in plaats van negatief. Hierdoor is zowel de beschrijving van de norm als de beschrijving van de aspecten positief geformuleerd. Daarnaast is het aanvinken van de risicoaspecten verplicht gemaakt. Het effect van de consensusbijeenkomst is onderzocht door een voor- en nameting uit te voeren en door de twee groepen met elkaar te vergelijken.

In de consensus bijeenkomst bespreken inspecteurs casuïstiek en proberen tot overeenstemming te komen over het oordeel. Inspecteurs bespreken een aantal criteria, dat zij op volgorde van laag risico tot hoog risico moeten rangschikken. Om het effect van de consensusbijeenkomst te onderzoeken, hebben alle inspecteurs van het toezicht op de verpleeghuiszorg deelgenomen aan deze bijeenkomst. Het doel ervan was om gemeenschappelijke bronnen van variatie in oordelen met elkaar te identificeren. Inspecteurs kregen de opdracht om consensus te bereiken over de volgorde van twee sets van vier casussen die zij van laag naar hoog risico moesten ordenen. Na de bijeenkomst onderzochten inspecteurs casussen die veel leken op de casussen van de voormeting, maar die niet precies hetzelfde waren om leereffecten van de vorige casussen te voorkomen.

Zowel de betrouwbaarheid als de validiteit van de inspecteursoordelen is het hoogst na de consensus bijeenkomst. De resultaten laten ook zien dat het vergroten van het aantal inspecteurs dat een casus beoordeelt, zowel de betrouwbaarheid als de validiteit van de oordelen doet toenemen. In deze casusstudie hebben inspecteurs niet met elkaar kunnen overleggen over hun oordelen. Dit is een gegeven geweest bij de analyse van het effect van het vergroten van het aantal inspecteurs dat een casus beoordeelt. Onder deze experimentele omstandigheden leidt het vergroten van het aantal inspecteurs tot een substantiële toename van zowel de betrouwbaarheid en validiteit van de oordelen. In de praktijk zullen inspecteurs, als zij in duo's of teams instellingen bezoeken, hun bevindingen en oordelen wel met elkaar bespreken. Het is redelijk te verwachten dat de toename van de betrouwbaarheid van de inspecteursoordelen hierdoor hoger zal zijn dan in de experimentele situatie. Echter, of deze verwachte toename in betrouwbaarheid en validiteit zonder meer gegeneraliseerd kan worden naar de praktijk, waarin inspecteurs wel (kunnen) overleggen over hun oordeel staat niet vast. Dit zou nader onderzocht kunnen worden.

9.7 Discussie

De uitkomsten van deze studies laten zien dat de structurering van beoordelingsinstrumenten en het gebruik van deze instrumenten een belangrijke rol spelen bij het realiseren van (meer) betrouwbare en valide inspecteursoor-

delen. Alleen focussen op het instrument lijkt echter te beperkt: continue scholing in het gebruik van toezichtsinstrumenten kan voorkomen dat inspecteurs hun beslissingsproces teveel individualiseren.

Wat zijn de implicaties van het onderzoek voor de praktijk van het toezicht? Zowel in de professionele context als de organisatiecontext zijn verbeteringen mogelijk. In de theorie van 'reflectie-in-actie' wordt ervan uit gegaan dat professionals in de dagelijkse praktijk kennis op een impliciete manier verwerven door reflectie op deze praktijk. Consensusbijeenkomsten gericht op oordeelsvorming stimuleren en kaderen deze reflectie door de uitwisseling van ervaringen en ideeën.

De werkwijze in lerende organisaties biedt ook kansen voor IGZ: het monitoren en verbeteren van de betrouwbaarheid en validiteit van de oordelen is een kenmerk van een organisatie die zich voordurend wil blijven ontwikkelen. In een lerende organisatie zijn er voorwaarden om individueel leren om te zetten in teamleren. In de Academische Werkplaats Toezicht worden de werelden van onderzoek en praktijk van het toezicht samengebracht. Dit stimuleert het teamleren binnen een sterke kennisstructuur.

Dankwoord

Een spannend promotieonderzoek uitvoeren over een interessant onderwerp. “Dat wordt een mooie tijd van verdiepen, denken, doen, analyseren, schrijven en leren”, dacht ik toen ik eraan begon. En dat was ook zo. Maar niet alleen was dit onderzoek inhoudelijk een uitdaging. Dit onderzoek bood me net zoveel mogelijkheden om me op persoonlijk vlak te ontwikkelen. Ik heb bij dit onderzoek op allerlei manieren hulp gekregen van heel veel mensen. Het is dan ook niet voor niets dat het proefschrift in de we-vorm is geschreven.

Toen het onderzoek van start ging was het onderwerp van dit proefschrift een gevoelig onderwerp voor de IGZ. Voor de inspecteurs en toezichtmedewerkers bij wie ik over de schouder mee keek naar hun oordelen, hun onderbouwingen, hun instrumenten. Maar ook voor de leidinggevenden was dit type onderzoek nieuw. Een open en transparante toezichthouder die durft te zeggen dat ze werkt aan haar toezicht om dit nog verder te professionaliseren. Een toezichthouder die dat niet alleen durft te zeggen, maar ook de daad bij het woord voegt door een uitgebreid onderzoek te laten uitvoeren waarbij over de uitkomsten gepubliceerd wordt. Dat vergt moed en een cultuuromslag. Daarom wil ik graag vele collega’s bedanken.

Zonder de medewerking en openheid van de inspecteurs en toezichtmedewerkers van programma 6 (toezicht op verpleeghuizen) en programma 4 (toezicht op ziekenhuizen) was dit onderzoek niet mogelijk geweest. Daarom wil ik alle collega’s uit programma 4 bedanken voor hun deelname aan mijn onderzoek en alle collega’s uit programma 6 bedanken voor hun deelname aan de casusstudie en de consensusbijeenkomst. Jullie bijdrage was essentieel voor het onderzoek.

Cruciaal voor de uitvoering van dit onderzoek was de steun van Inspecteur-generaal Gerrit van der Wal en plaats-vervangend Inspecteur-generaal Hans Janssen en vanaf 2013 van Inspecteur-generaal Ronnie van Diemen-Steenvoorde en directeur bedrijfsvoering Rob de Haan. Zonder jullie goedkeuring en support, was het niet mogelijk geweest dit onderzoek uit te voeren. Dank jullie voor jullie vertrouwen in het onderzoek en het belang dat jullie eraan hechtten.

Een speciaal woord van dank voor mijn leidinggevende Jeroen Geelhoed. Jeroen, zonder jou was het niet mogelijk geweest om aan dit onderzoek te werken. Ik ben je zeer erkentelijk voor de ruimte en mogelijkheid die je gaf en jouw vertrouwen in mij om dit onderzoek uit te voeren. Je bereidheid om altijd mee te denken, om kansen te creëren heb ik als zeer stimulerend ervaren. Jouw deur staat altijd open om even binnen te wandelen voor een praatje, een vraag of brainstorm. Graag wil ik je bedanken voor de mogelijkheid die je me geeft om me verder te ontwikkelen. Je brede visie, snelheid van denken, tomeloze energie, strategisch inzicht en gevoel voor ambitie maken dat ik veel kan leren. Je altijd aanwezige belangstelling waardeer ik ontzettend.

Een speciaal woord van dank voor mijn team van wijze heren, mijn promotoren: professor Paul Robben, professor Huub van den Bergh en professor Frans Janssens. Ik kan wel zeggen het een voorrecht was om met drie zulke bevlogen hoogleraren te mogen werken. Graag wil ik jullie alle drie zeer bedanken voor de fijne samenwerking de afgelopen jaren.

Beste Paul, je empatische stijl van begeleiden, je grote betrokkenheid, je snelle en stimulerende manier van feedback geven, maken het prettig om met jou te mogen werken. Jouw geduld en begrip lijken onuitputtelijk. Je vaardigheden om bruggen te bouwen en draagvlak te creëren zijn een ware gave en hebben me zeer geïnspireerd. Van je kennis over goed schrijven, over toezicht en bestuur, over onderzoek in een politieke omgeving heb ik veel geleerd. Graag wil ik je bedanken voor de ruimte die je gaf en geeft om altijd binnen te lopen met een vraag, mee te denken over een complex vraagstuk, even te brainstormen en je snelle reacties op de zovele stukken die ik stuurde. Altijd kom ik met een goed idee bij je vandaan. Je wist me altijd te motiveren als ik even niet wist hoe ik verder moest en hielp met relativeren als het tegenzat. Ik ben je zeer dankbaar voor de bijdrage die jij op vele fronten hebt geleverd. Het is een plezier om met jou te mogen samenwerken.

Beste Huub, jouw kracht om altijd in alles het positieve te zien vind ik bewonderenswaardig en inspireert me altijd opnieuw. Mogelijkheden zien, variabelen toevoegen waar nog wat extra kennis te vergaren valt, kansen creëren, alles is mogelijk. In het onderzoek. En in het leven. Dat waardeer ik ontzettend in jou. Onze gesprekken over het onderzoek en over alles behalve het onderzoek onder het genot van een verse kop koffie, stimuleren mij vaak tot nieuwe ideeën. Altijd kom ik geïnspireerd bij je vandaan. Je snelheid van denken in op-

lossingen, getallen, analyse en je kennis van statistische programma's en methodologie deden me vaak duizelen en maakten tegelijkertijd dat ik zoveel kon leren. Je vaardigheden om complexe materie uit te leggen en begrijpelijk te maken, is een ware gave. Het is een plezier om met jou te mogen samenwerken.

Beste Frans, jouw brede kennis over toezicht, bestuur en literatuur. Je directe manier van feedback geven, je enthousiasmerende mails met zo vaak weer een relevant artikel bijgevoegd, waren een bron van inspiratie en stimulatie. Juist op de momenten waarop het nodig was, stak jij een hart onder de riem, moedigde je me aan, wist met humor te relativeren, stuurde je relevante artikelen of gewoon een leuke foto. Je kunt ontzettend goed schrijven, verbinden, beargumenteren, nuanceren en relativeren. Graag wil ik je bedanken voor je grote betrokkenheid, je bereidheid om altijd mee te denken, je kunst om alternatieve routes aan te wijzen en soms even aan de rem te hangen. Ook toen je voor je werk een tijd in het buitenland zat, gaf je commentaar om mijn stukken, mocht ik je bellen en stuurde je me gevraagd of ongevraagd stukken ter verdieping. Je noemde dat "in de Tuijn werken" en ik vind het fantastisch dat je dat ook op de BES-eilanden deed. Ik dank je zeer voor je grote betrokkenheid en deskundige inbreng. Het is een plezier om met jou te mogen samenwerken.

Leden van de leescommissie, prof. dr. G. van der Wal, prof. dr. R.A. Bal, prof. dr. F.L. Leeuw, prof. dr. T.J.M. Sanders en prof. dr. G.A.M van den Bos; jullie wil ik bedanken voor de aandacht die jullie aan mijn proefschrift hebben geschonken.

Een aantal mensen in het bijzonder wil ik bedanken voor hun inzet bij dit onderzoek. Veel dank aan jou, Anja Jonkers. Zonder jouw betrokkenheid bij en steun voor het onderzoek was het lastig geworden om de soms moeilijk begaanbare paden te bewandelen. Dank je wel voor de tijd, moeite en energie die je hebt gestoken in het faciliteren van het onderzoek. Maar ook veel dank voor het vertrouwen dat je had in het onderzoek. Ook veel dank aan jou Marjo Ligthart. Je gaf me de mogelijkheid om de inspecteurs van programma vier te betrekken in het onderzoek wat veel waardevolle informatie heeft opgeleverd. Dank je wel voor het vertrouwen dat je had in het onderzoek.

Ook Jan van Wijngaarden, Josée Hansen en Joke de Vries wil ik bedanken. Jan, José en Joke, bedankt voor jullie vertrouwen in het onderzoek en het belang dat jullie eraan hechten.

Karen Kolenbrander graag wil ik je bedanken voor je bereidheid om als tweede observator mee te doen in het onderzoek naar de validiteit van inspecteurs oordelen. Je beoordeelden zonder morren 615 oordelen en beoordeelde ook bijbehorende onderbouwingen naast je drukke werkzaamheden. Ik ben je hiervoor dankbaar.

Marjolein Garretsen, jouw snelle, enthousiaste en deskundige afhandelingen van de tientallen literatuuraanvragen, waarbij je listige manieren bedacht om lastige stukken toch te pakken te krijgen, waren onmisbaar in dit onderzoek. Het was een plezier om met je samen te werken.

Ida Bream-de Ruiter, dank je wel voor je snelle en accurate afhandeling van de offerteverzoeken. Soms was er ineens tempo nodig en dat is bij jou in goede handen. Ik kon bij jou altijd terecht voor een vraag, een tip, een bakkie of gewoon een gezellige babbel. Het is fijn om met jou te mogen samenwerken.

Marijn Beelen, Fransien van ter Beek en Wouter van der Horst: Jullie journalistieke kennis en kunde waren van zeer grote waarde bij dit onderzoek. Jullie passie voor communicatie en jullie strategische aanpak, heb ik erg gewaardeerd. Ik vond het altijd weer een belevenis om jullie wereld van communicatie en journalistiek binnen te stappen en te horen wat jullie allemaal meemaakten en hoe jullie daarmee omgingen. Bedankt voor jullie hulp, support en jullie gezelligheid.

Harold Block, jouw kennis en kunde van web-based survey waren cruciaal in dit onderzoek. Jouw geduld als ik toch nog een keer “de puntjes op de i wilde zetten” leek nooit op te raken. Je bereidheid om mee te denken en slimme oplossingen te bedenken om technische obstakels te omzeilen heb ik erg gewaardeerd. Maar ook onze gezellige gesprekken waarin we samen lachten om kleine dingen, heb ik erg op prijs gesteld. Je bent een geweldige collega en het was super om met je samen te werken.

Richard Versteeg, dank je wel dat je vandaag jouw kwaliteiten als fotograaf wilt inzetten om deze bijzondere dag vast te leggen.

Jacqueline Caster, ook een woord van dank voor jou. Niet alleen voor alle keren dat je zonder mopperen vergaderruimtes boekte voor de overleggen die nodig waren voor dit onderzoek. Maar ook voor je altijd aanwezige warme persoonlijke belangstelling.

Ronald van Kessel, graag wil ik jou bedanken voor je technische hulp bij de video-opnames van de consensusbijeenkomsten. Altijd ben je bereid om te helpen. Met jouw instructies lukte het om goede opnames te maken die bruikbaar en nodig waren bij het onderzoek.

Jan van Berlo en Jan Slotema, veel dank voor jullie tijd en inspanning tijdens de validatie-ronde die nodig was om de casusstudie mogelijk te maken met gevalideerde casussen. Het waren leuke, gezellige en tegelijkertijd professionele en leerzame bijeenkomsten.

Marianne Bobeldijk, Anna van Beuge, Jolanda Peper en Joke Dalderup, dank voor jullie deelname aan de expertmeeting waarin we met elkaar inventariseerden welke aanpassingen aan het instrument gewenst waren. Jullie inbreng was onmisbaar voor de casusstudie!

Joke Dalderup, veel dank voor je bereidheid om met me te delen hoe het instrument voor de tweede fase van het gefaseerd toezicht voor het toezicht op de verpleeghuiszorg tot stand is gekomen. Een interessant proces waar ik veel van geleerd heb. Dank voor je kritische vragen en je bereidheid om met me te discussiëren over dit onderwerp.

Jannie Speksnijder, dank voor jouw kritische opmerkingen en mogelijkheid voor discussie. Ze hielpen me om het proces helder te krijgen en hier ook veel van te leren.

Rhinske Dhoeri-Plomp en Bep Corporaal, jullie wil ik graag bedanken voor jullie bereidheid om met me te delen hoe de Inspectie van het Onderwijs omgaat met het verschijnsel beoordelaarsverschillen. Het was altijd mogelijk om even te bellen of een afspraak te maken en dat heb ik erg op prijs gesteld.

Colleen Higgins, jou wil ik bedanken voor je hulp bij het schrijven in de Engelse taal. Daar waar ik dacht dat ik me aardig in het Engels had uitgedrukt, wist jij het mooier en beter te maken. Niet alleen jouw expertise van de Engelse taal, maar ook onze gezellige mailwisselingen heb ik erg gewaardeerd. Dank je wel daarvoor!

Ook wil ik graag mijn collega's van de afdeling Ontwikkeling en Innovatie bedanken voor de mogelijkheid die jullie gaven voor de korte brainstorms, jullie bijdrage aan de pilots, de gezelligheid. In het bijzonder wil ik Perry Koevoets bedanken: je hebt me altijd geholpen als ik weer de voetnoot-functie kwijt was, om een tabel of figuur mooier te krijgen en om me bij te staan met raad en daad. Voor mij onmisbare supportmomenten.

Tijmen, jou in het bijzonder wil ik graag bedanken. We waren 5 jaar kamergenoten en hebben altijd heel fijn contact gehad. Je bent een collega door dik en dun.

Ook jou Elise, wil ik graag in het bijzonder bedanken. We zijn gelijk begonnen met werken bij de IGZ. Niet alleen delen we nu ervaringen over ons werk bij IGZ, maar ook over het moederschap. Je bent altijd bereid om mee te denken, ongeacht het onderwerp. Je warme belangstelling voor hoe is het is, en het samen relativeren en samen even heerlijk lachen om "wat we nu weer hebben meegemaakt" maken je een heel fijne collega en een heel fijn mens. Ik hoop dat we nog lang zo fijn contact hebben.

Een speciaal woord van dank voor mijn twee paranimfen Marjon en Sandra. Dank jullie wel dat ik altijd wel even kon discussiëren over mijn onderzoek of mijn hart kon luchten. Jullie stonden me bij met raad en daad en met een grap als het even tegenzat of de schrijfdrempel hoog was en zijn twee heel fijne collega's die mij op jullie eigen manier op veel verschillende momenten en manieren hebben geholpen bij het afronden van dit onderzoek. Ik ben dan ook trots en gelukkig dat jullie vandaag aan mijn zijde staan.

Verder wil ik alle collega's van IGZ die (met regelmaat) vroegen hoe het met mijn onderzoek ging, die met kritische vragen kwamen, met aanvullende verklarende factoren, met complimenten of met een hart onder de riem, bedanken voor hun interesse.

Natuurlijk wil ik ook mijn lieve vrienden en vriendinnen bedanken. Lieve Simone en Remco, Sjors en Jennifer, Anette, Miriam, Janneke, Monique, Martine en Katja. Jullie zijn goud waard. Ik prijs me zeer gelukkig met zulke lieve vrienden om me heen. Ik ben blij dat ik met jullie vooral praatte over alles behalve dit onderzoek.

Tot slot wil ik me richten tot mijn familie. Mijn lieve ouders, mijn lieve pap en mam. Jullie zijn de meest fantastische ouders die ik me maar kan wensen. Jullie altijd aanwezige vertrouwen, belangstelling, jullie luisterend oor, jullie onvoorwaardelijke steun en liefde en jullie support om me te ontwikkelen. Altijd als ik jullie nodig heb, zijn jullie er. Dag en nacht staan jullie klaar voor mij en mijn gezin en het is nooit te veel. Geen reis is te ver. Geen storm te hard. Ik prijs me zo rijk en gelukkig met zulke geweldig lieve ouders en grootouders van m'n lieve jongens. Ik hoop nog lange tijd van jullie gezelligheid en liefde te mogen genieten.

Lieve Magda, lieve zus. Naast een fantastische moeder voor Fleur en Huib, ben je een ontzettend lieve zus en geweldige apotheker. Ik hoop nog heel lang te kunnen genieten van onze fijne gesprekken over de grote en kleine dingen in het leven en samen te genieten van gezellige koffiemomenten met onze ravottende kindjes om ons heen.

En dan mijn thuis.

Owen, mijn allerliefste kleine Owen. Je bent het kleinste mannetje in mijn leven, maar je hebt een net zo grote plek in mijn hart als mijn andere mannen. Het is fantastisch om te mogen meemaken hoe je je ontwikkelt naast je grote broer Aiden en te zien hoe je al zoveel wil. Je wil overal bij zijn, niets missen. Een levensgenieter in de dop! Elke dag geniet ik van de stapjes die je zet, je prachtige pretogen en grote glimlach. Elke dag koester ik het geluk dat ik jouw moeder mag zijn.

Aiden, mijn allerliefste lieve Aiden. Elke dag dat ik jouw lachende bruine ogen zie, je schaterlach hoor, je vrolijke geklets en je liedjes, prijs ik me de gelukkigste vrouw op de wereld. Je leergierige en eigenwijze inborst, je lust om de wereld te ontdekken, je humor, de vragen die je stelt. Elke dag koester ik het geluk dat ik jouw moeder mag zijn.

Mijn Mark, mijn lieve Mark. Vooral jij, bedankt voor je onvoorwaardelijke trouw, aan mij als levensgezel en aan mijn ambitie om mijn promotieonderzoek te voltooien. Je bent altijd bereid geweest mijn onder-

zoek te volgen en niet in de laatste plaats om jezelf opzij te zetten en me te helpen met figuren of data. Het is fantastisch om mijn leven met jou te delen en samen met jou de wereld te ontdekken en het avontuur van ouderschap te mogen beleven. Je bent een geweldige man voor mij en een fantastische vader van onze mannetjes Aiden en Owen. Met jou klopt alles, is het compleet.

Curriculum vitae

Saskia Tuijn was born on December 27, 1975 in Krommenie, the Netherlands. In 1994 she received her athenaeum diploma from the Bertrand Russell College in Krommenie (athenaeum is a form of pre-university education). She went on to study speech and language therapy at HU University of Applied Sciences Utrecht (Hogeschool Utrecht), where she received her bachelor's degree in 1998. After graduation, she worked as a speech and language therapist while studying for her Master of Arts degree in communication, and graduated cum laude from Utrecht University in 2003**.** She then worked as a researcher at the Netherlands Institute for Health Services Research (NIVEL)**.** In 2005, she went to work for the Dutch Health Care Inspectorate (IGZ), and started her PhD research in 2008. In 2010, she received the "Vide Publication Award" (Vide Publicatieprijs) for one of her publications. This dissertation is the result of the research she carried out from 2008 through 2013. Since 2012, she has been working at the IGZ on improving working processes, in particular the interrater reliability and validity of inspectors' regulatory judgments. Saskia lives in Houten with her partner Mark and their two sons, Aiden (2010) and Owen (2012).