Journal of Evaluation in Clinical Practice ISSN 1365-2753

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

Saskia M. Tuijn MsC,[1] Huub van den Bergh PhD,[2,3] Paul Robben PhD[4,5] and Frans Janssens PhD[6]

[1]Consultant, Project Manager, O&I, IGZ, Utrecht, The Netherlands
[2]Professor, Modern Languages Department, Faculty of Humanities, University of Utrecht, Utrecht, The Netherlands
[3]Professor, Teacher Training Institute, University of Amsterdam, The Netherlands
[4]Professor, Effectivity of Regulation, Institute for Policy and Management in Health Care, Erasmus University Rotterdam, Rotterdam, The Netherlands
[5]Advisor, Dutch Health Care Inspectorate, The Netherlands
[6]Professor, Management and Organization of Education, Faculty of Behavioral Sciences, University of Twente, Enschede, The Netherlands

Keywords
evaluation, health services research

Correspondence
Ms Saskia Maria Tuijn
O&I
IGZ
St Jacobsstraat 16
Utrecht 3511 BS
The Netherlands
E-mail: s.tuijn@IGZ.nl

Accepted for publication: 27 March 2014

doi:10.1111/jep.12136

Abstract

Rationale, aims and objectives We examined the effect of two interventions on both the reliability and validity of regulatory judgments: adjusting the regulatory instrument and attending a consensus meeting.

Method We adjusted the regulatory instrument. With a randomized controlled trial (RCT) we examined the effect of the adjustments we made to the instrument. In the consensus meeting inspectors discussed cases and had to reach consensus about the order of the cases. We used a before and after case study to assess the effect of the consensus meeting. We compared the judgments assigned in the RCT with the unadjusted instrument with the judgments assigned with the unadjusted instrument after the consensus meeting. Moreover we explored the effect of increasing the number of inspectors per regulatory visit based on the estimates of the two interventions.

Results The consensus meeting improved the agreement between inspectors; the variance between inspectors was smallest (0.03) and the reliability coefficient was highest (0.59). Validity is assessed by examining the relation between the assigned judgments and the corporate standard and expressed by a correlation coefficient. This coefficient was highest after the consensus meeting (0.48). Adjustment of the instrument did not increase reliability and validity coefficients.

Conclusions Participating in a consensus meeting improved reliability and validity. Increasing the number of inspectors resulted in both higher reliability and validity values. Organizing consensus meetings and increasing the number of inspectors per regulatory visit seem to be valuable interventions for improving regulatory judgments.

Introduction

Government regulation of health care aims to monitor and minimize risks in health care and to simultaneously stimulate the quality of care. Internationally, the effects of regulation on the quality of public services have been discussed extensively and sometimes criticized [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. Inspectors make judgments and decide whether health care organizations have to improve quality. Both the credibility and authority of enforcement agencies will be hampered when the judgments are not reliable or valid. Moreover, scientific publications on this subject make it possible to exchange knowledge and learn internationally.

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 standardized 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 organizations they regulate, a process described as 'detection' [17]. For this purpose, the IGZ uses a combination of three methods. First, 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. Third, 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 that 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. The policy of IGZ is that the specific health care sectors represented within IGZ develop their own regulatory instruments. To gain insight in the instrument used for regulation of nursing home we will describe this instrument.

Instrument for regulation of nursing home care in the Netherlands

The instrument for regulation of nursing home care that 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 4-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. Consequently, inspectors are not satisfied with the instrument as it is.

Aim of the study

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 inter-rater 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 methods of training vary, but all focus on the identification on sources of disagreement. The outcomes of the studies on health care professionals seem relevant for health care inspectors because the characteristics of health care inspectors resemble those of health care professionals: inspectors are professionals as well, and also have health care backgrounds. However, there are differences as well: inspectors assess organizations instead of patients, using instruments such as written criteria or standards instead of highly technical instruments such as computed tomography scans [27]. Our main research questions concern the effects of:

1 adjustment of the regulatory instrument on the reliability and validity of regulatory judgments.

2 attending a consensus meeting on the reliability and validity of regulatory judgments.

Furthermore, based on the results the effect of increasing the number of inspectors on the reliability and validity was estimated.

We performed an experimental study to answer our research questions.

Methods

Adjusting the regulatory instrument

We organized 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

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.

very difficult [24]. Second, 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 selected two adjustments:

1 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.

2 We made checking the aspects of risk mandatory.

After the expert meeting, we used a randomized controlled trial to examine the effect of the adjustments. We randomly assigned the inspectors ($n = 25$) to group 1 and group 2 (see Figure 1). The inspectors in group 1 used the unadjusted instrument, and the inspectors in group 2 used the adjusted instrument. 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. To increase response among the inspectors, we sent two reminders for both the first and second measurements. In the end, nine inspectors used the adjusted instrument and 15 inspectors used the unadjusted instrument. To examine the effect of the adjustments, we compared the judgments assigned with the unadjusted instrument to the judgments assigned with the adjusted instrument.

Participating in a consensus meeting

To gain insight into the effect of participating in a consensus meeting, we used a before and after design (see Figure 1). Four weeks after the randomized clinical trial we organized a consensus meeting for the inspectors ($n = 25$) 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.

Of the 25 inspectors, 15 inspectors attended the consensus meeting (60%) and all of these 15 inspectors (100%) examined the 16 cases after the consensus meeting with the unadjusted instrument within 6 weeks (group 3). These cases were very similar to the cases used in the randomized controlled trial described above, but not completely identical to prevent learning effects. This second round of review was conducted after a significant period of time had elapsed following the randomized controlled trial (6 weeks); this was performed to prevent recollection, which would have introduced bias into the review process [29]. We compared the judgments of group 3 with the judgments of group 1 in the randomized controlled trail described above.

Case descriptions

In both the randomized controlled trial and the before and after study, inspectors examined cases and assigned regulatory judgments to the situations described in these cases. The cases concerned two criteria: 16 of the cases described the criterion 'pressure ulcers' and 16 cases described 'professionalism of the staff'. 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 cover only a small part of the spectrum; they visit institutions that perform relatively less good on the indicators. 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.

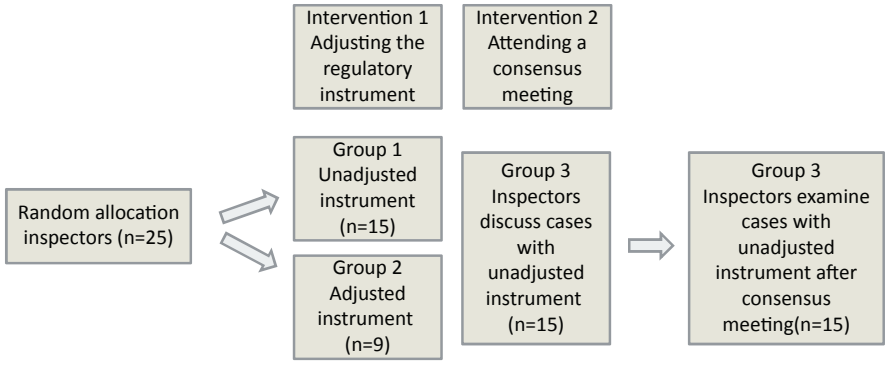

Figure 1 Research design of the study.

Box 1 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.

In this experiment, we tried to simulate this situation in an optimal way 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 4-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 1 presents an example of such a case.

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 minimized the possibility of discussing the cases simultaneously. The inspectors who dropped out withdrew themselves from the study despite the reminders we sent.

Increasing the number of inspectors per regulatory visit

The effect of the manipulation can be described by the agreement between inspectors (reliability) and by the correspondence between the assigned judgments and the corporate standard (validity). For both the effect of increasing the number of inspectors was estimated.

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.

Statistical analysis

The data collected in the experiments are hierarchical in nature, 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 generalizable over both inspectors and cases. Therefore, we need to estimate three components of variance: the variance between cases, the variance between inspectors (the extent to which inspectors differ in their judgments on a case about nursing home care) and the interaction between inspectors and cases that 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 that is represented by the reliability coefficient (rho). To calculate rho we used the formula [30] presented in Box 2.

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

Box 2 Formula used to calculate rho

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

Box 3 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 organization.
3 The level of palpability of the criterion is important. Inspectors experience the criterion 'pressure ulcers' as concrete 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 unequivocally 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 organization in terms of the number of beds is not part of the instrument's criteria, nor is it explained how inspectors can account for an organization'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. Inspectors' frames of reference vary, and play a role in judging an organization.
8 Some inspectors focus on details, whereas 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, whereas 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.

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.

Results

In this section we will first describe the outcomes of the identification of sources of variation in the consensus meeting. Next, we will explain the results of the experiments.

The results of the consensus meeting are presented in Box 3.

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 [1,3–6,9,12], whereas 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 organization'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.

Effects of the interventions are presented. The fit of the models is presented in Table 2.

In the first model (the equal reliability model) neither variance was allowed to differ between conditions. In the second model (the different error model), the error variance was allowed to differ between conditions. If this model fits best, the reliability varies as a result of error variance. In the third model (the different variance and error model), the variance between inspectors was allowed as well. If this model fits best, the reliability varies as a result of variance between inspectors as well.

The results indicate that −2LL of model 1 is significantly higher compared with the −2LL of model 2 [$\chi^2 = 75.24$; degrees of freedom (d.f.) = 5; $P < 0.0001$]. Yet, the third model fits better to the observed data as model 2 ($\chi^2 = 74.85$; d.f. = 6; $P < 0.0001$). Hence, the variance between inspectors as well as the error variance differs between measurements. The estimated variances of model 3 give insight into the effect of the interventions on reliability (Table 3) and validity (Table 4). In Tables 4 and 5 the effects of the adjustment of the instrument and the effect of the consensus meeting are presented. The calculated effects represent the values of the judgment of one inspector assigned with the unadjusted instrument, assigned with the adjusted instrument and assigned after the consensus meeting.

Table 2 Outcomes of the comparison of the three models used to represent our data

	–2 log likelihood	Comparison			
		Model	χ^2	d.f.	*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				

d.f., degrees of freedom.

Table 3 The effect of adjusting the instrument and a consensus meeting on inter-rater 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); 0.47
Adjusted	3.27 (2.82; 3.72)	0.22 (26)	0.22 (26)	0.41 (48); 0.48
Consensus	3.81 (3.48; 4.14)	0.26 (37)	0.03 (4)	0.41 (59); 0.59
	Cases on pressure ulcers			
Unadjusted	2.51 (2.18; 2.84)	0.61 (62)	0.02 (2)	0.35 (35); 0.35
Adjusted	2.93 (2.53; 3.34)	0.39 (45)	0.14 (16)	0.35 (40); 0.40
Consensus	2.99 (2.63; 3.30)	0.24 (38)	0.05 (8)	0.35 (54); 0.54

% case, percentage of variance explained by cases; % error, percentage of variance explained by error; % inspector, percentage of variance explained by inspectors; CI, 80% confidence intervals; rho, mean reliability when one inspector examines a case; S^2_{case}, variance of cases; S^2_{error}, variance of inspectors and cases; $S^2_{inspector}$, variance of inspectors.

Table 4 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

% case, percentage of variance explained by cases; % error, percentage of variance explained by error; % inspector, percentage of variance explained by inspectors; CI, 80% confidence intervals; rho, mean reliability when one inspector examines a case; S^2_{case}, variance of cases; S^2_{error}, variance of inspectors and cases; $S^2_{inspector}$, variance of inspectors.

Table 3 shows that if one rater rates 'professionalism', the reliability is estimated as 0.47 if the unadjusted instrument is used. When one inspector assigns a judgment with the adjusted instrument, the reliability increases to 0.48. When one inspector assigns a judgment after the consensus meeting, the reliability has increased to 0.59. Table 3 shows 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 4 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%). Inspector variance was relatively small after the consensus meeting (0.03) compared with both the pretest 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 pretest 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 (Fig. 2).

Figure 2a shows that when the number of inspectors increases, reliability increases as well. The increase of rho varies between conditions. Figure 2b 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. Figure 2a,b show that the effect of the increase of inspectors declines after three inspectors.

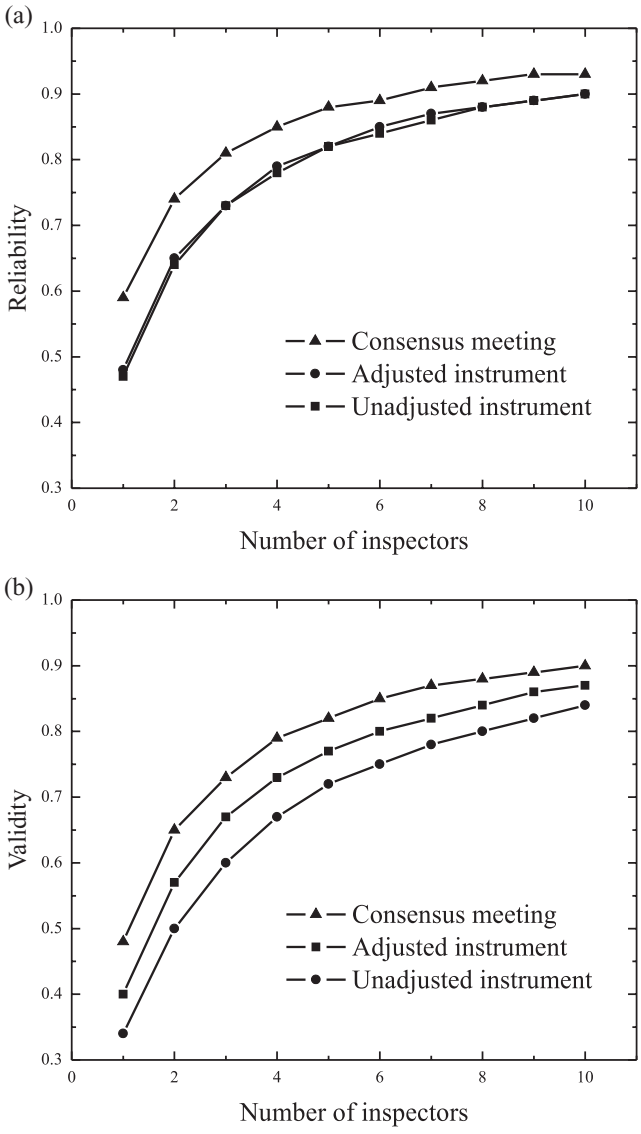

Figure 2 (a, b) The effect of increasing the number of inspectors on the reliability and validity of regulatory judgments.

Discussion

In this study our research questions concerned the effects of:
1 adjustment of the regulatory instrument on the reliability and validity of regulatory judgments.
2 attending a consensus meeting on the reliability and validity of regulatory judgments.

Furthermore, we estimated the effect of increasing the number of inspectors on the reliability and validity of regulatory judgments based on the results of the experiment.

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. 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. The results indicate that participating in a consensus meeting and increasing the number of inspectors per examined case improved reliability and validity.

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]. 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.

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 characterized 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 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.

Strengths and limitations

This study has several strengths. First, to our knowledge, this is the first study to investigate the effect of interventions on inter-rater 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 6 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 minimized 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 inter-rater 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 organization withdrew from the study.

Implications

Despite the methodological limitations of our study, it does have some implications and offers the possibility of providing opportunities for further improvement of regulatory judgments. 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 organizations 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.

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 (e.g. 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.

Acknowledgements

We gratefully acknowledge Jenneke van Veen (IGZ), Anja Jonkers (IGZ programme director) and all the study participants. Without

their support and cooperation, it would not have been possible to perform this study. We would like to thank Harold Block (IGZ) for his skilful, quick processing of the cases in the web-based technique.

References

1. Brennan, T. A. (1998) The role of regulation in quality improvement. *The Milbank Quarterly*, 76 (4), 709–731.
2. Walshe, K. (1999) Improvement through inspection? The development of the new Commission for Health Improvement in England and Wales. *Quality in Health Care*, 8, 191–196.
3. Walshe, K. (2002) The rise of regulation in the NHS. *British Medical Journal*, 324, 967–970.
4. Bevan, G. & Hood, C. (2004) Targets, inspections, and transparency. *British Medical Journal*, 324, 967–970.
5. Ham, C. (2005) From targets to standards: but not just yet. The challenge will be for ministers not to interfere in a regulated service. *British Medical Journal*, 330, 106–107.
6. Bevan, G. & Hood, C. (2006) Have targets improved performance in the English NHS? *British Medical Journal*, 332, 419–422.
7. Bevan, G. & Hood, C. (2006) What's measured is what matters: targets and gaming in the English public health care system. *British Medical Journal*, 84, 517–538.
8. Bevan, G. (2009) Have targets done more harm than good in the English NHS? *British Medical Journal*, 338, a3129.
9. Gubb, J. (2009) Have targets done more harm than good in the English NHS? *British Medical Journal*, 338, a3130.
10. Ham, C. (2010) Improving the performance of the English NHS. Systems of care are needed to build on progress to date. *BMJ (Clinical Research Ed.)*, 340, c1776.
11. Day, P. & Klein, R. (1987) The regulation of nursing homes. *The Milbank Quarterly*, 65 (3), 303–347.
12. Hutter, B. M. (1989) Variations in regulatory enforcement styles. *Law and Policy*, 2, 153–174.
13. Braithwaite, J., Makkai, T. & Braithwaite, V. (2007) Regulating Aged Care. Ritualism and the New Pyramid. Cheltenham: Edward Elgar.
14. Greenfield, D., Braithwaite, J. & Pawsey, M. (2008) Healthcare accreditation surveyor styles typology. *International Journal of Health Care Quality Assurance*, 21, 435–443.
15. Mascini, P. & van Wijk, E. (2008) Responsive regulation at the Dutch Food and Consumer Product Safety Authority: an empirical assessment of assumptions underlying the theory. *Regulation & Governance*, 3, 27–47.
16. IGZ (2011) Policy Plan 2012–2015. For justified trust in appropriate and safe care II [Meerjaren beleidsplan 2012–2015. Voor gerechtvaardigd vertrouwen in verantwoorde zorg], (in Dutch).
17. Walshe, K. (2003) Regulating Healthcare: a Prescription for Improvement?, pp. 34. Berkshire: Open University Press.
18. IGZ (2007) Policy plan 2008–2011. For justified trust in appropriate and safe care [Meerjaren Beleidsplan 2008–2011. Voor gerechtvaardigd vertrouwen in verantwoorde zorg], (in Dutch).
19. Luthi, J. C., McClellan, W. M., Flanders, W. D., Pitts, S. & Burnd-Hand, B. (2002) Quality of health care surveillance systems: review and implementation in the Swiss setting. *Swiss Medical Weekly*, 132, 461–469.
20. Kollberg, B., Elg, M. & Lindmark, J. (2005) Design and implementation of a performance measurement system in Swedish health care services: a multiple case study of 6 development teams. *Quality Management in Health Care*, 14, 95–111.
21. Pettersen, I. J. & Nyland, K. (2006) Management and control of public hospitals – the use of performance measures in Norwegian hospitals. A case study. *The International Journal of Health Planning and Management*, 21, 133–149.
22. Lugtenberg, M. & Westert, G. (2007) Quality of healthcare and information for citizens to select healthcare. An international study on initiatives, (in Dutch).
23. Tuijn, S. M., Janssens, F. J. G., Van den Bergh, H. & Robben, P. B. M. (2009) 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]. *Nederlands Tijdschrift Voor Geneeskunde*, 153 (8), 322–326, (in Dutch).
24. Tuijn, S. M., Van den Bergh, H., Robben, P. B. M. & Janssens, F. J. G. (2009) The relationship between standards and judgments in the regulation of health care [De relatie tussen normen en oordelen in het toezicht op de gezondheidszorg]. *Tijdschrift voor Gezondheidswetenschappen*, 6, 264–271, (in Dutch).
25. Feldt, L. S. & Brennan, R. L. (1989) Reliability. In Educational Measurement, 3rd edn (ed. R. L. Linn), pp. 105–146. New York: MacMillan Publishing Company.
26. Tuijn, S. M., Janssens, F. J. G., Robben, P. B. M. & 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.
27. Walshe, K. (2003) Regulating Healthcare. A Prescription for Improvement?, pp. 182. Berkshire: Open University Press.
28. Tuijn, S. M., Robben, P. B. M., Janssens, F. J. G. & Van den Bergh, H. (2011) Evaluating instruments for regulation of health care in the Netherlands. *Journal of Evaluation in Clinical Practice*, 17, 411–419.
29. Nunnally, J. C. (1978) Psychometric Theory. New York: McGraw-Hill.
30. Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972) The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley.
31. Heuvelmans, A. P. J. M. & Sanders, P. F. (1993) Interrater Reliability [Beoordelaarsbetrouwbaarheid]. In Psychometrics in Practice [Psychometrie in de praktijk] (eds T. J. H. M. Eggen & P. F. Sanders), pp. 468. Arnhem: CITO.

