Evaluating clinicians’ teaching performance
Boerebach, B.C.M.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
CHAPTER 6

EMBRACING THE COMPLEXITY OF CLINICIANS’ PERFORMANCE EVALUATIONS: A NECESSITY FOR VALID CHANGE ASSESSMENT

Benjamin C. M. Boerebach
Onyebuchi A. Arah
Maas Jan Heineman
Kiki M.J.M.H. Lombarts

Submitted
ABSTRACT

Evaluations of clinicians’ professional performance have become more entrenched in clinical practice globally. Evaluation systems and tools have been developed and implemented and factors that impact performance in response to evaluations have been studied. The validity and reliability of data yielded by such tools have been studied extensively. However, there is limited evidence on the validity and reliability of data used to assess performance change, often estimated from time-indexed measurements. Important methodological and statistical issues that can impact the assessment of performance change are often omitted or ignored by research and practice. In this paper, we aim to address these issues and show how they can impact the validity of performance change assessments, using empirical illustrations based on longitudinal data of clinicians’ teaching performance. We show how the characteristics of a measurement scale can affect the performance data yielded by such an evaluation tool and show how different summary statistics of the same data can lead to opposing conclusions regarding performance change. In addition, we discuss the impact that the context and setting can have on both performance scores and the importance of specific aspects of performance. Further, we show that current methods to estimate performance domain scores from measured performance items can be too simplistic. Finally, this paper concludes that assessments of performance change are complex, situation and topic specific, and context dependent. Critically appraising the complexity of performance evaluations and relevant data is crucial for valid assessments of change.
INTRODUCTION

Health care reform, increased patient-awareness, enhanced expectations regarding the quality of care and the rising health care expenses are among factors that enhanced the need for transparency and monitoring of performance indicators in healthcare.\textsuperscript{1-3} One of the results is that clinicians are nowadays involved in multiple initiatives to chart their professional performance.\textsuperscript{4,5} For this purpose, evaluation systems have been developed to evaluate clinicians’ performance when interacting with patients, colleagues, other healthcare professionals and trainees.\textsuperscript{6-9} Research on evaluation tools used to gather performance data found that clinicians’ performance can be captured periodically with sufficient statistical validity and reliability.\textsuperscript{6,9-11} Recently, Rosen and Pronovost (2014) stated that performance evaluation tools often fail to address the context, culture and setting of performance, limiting their value for practice.\textsuperscript{12} Furthermore, the simple linear estimations that are often used to calculate performance scores do not address the fact that not every item is equally important to obtain successful results in most healthcare procedures.\textsuperscript{13} These and other well substantiated arguments lead to the conclusion that performance assessment is a complex process and that only more advanced evaluation tools will yield robust performance data.\textsuperscript{12,13}

Literature on the assessment of change in clinicians’ performance, that is, assessment of unaltered, improved or decreased performance, is still scarce. Assessment of change is much desired in practice because it can shed light on whether interventions were effective in improving or maintaining performance. Intuitively, when robust point-wise evaluations are complex and require advanced evaluation tools, robust assessments of change will be even more complex because of the added time component. Nevertheless, change is often simply computed linearly from two or more time-indexed evaluations of performance. Whether this method yields valid and reliable change assessments is still empirically understudied. In this paper, we address five factors that impact the complexity of performance evaluations and subsequent assessment of change. First, we discuss how characteristics of the measurement scale can affect the performance
data yielded by an evaluation tool. Next, we show how different summary statistics of the same data can lead to opposing conclusions regarding performance change. Subsequently, we discuss how change at item or question level can translate to overall performance change and show that estimating change from two time-indexed measurements and assessing change retrospectively yields different results. Lastly, we discuss the importance of context in assessing performance change. We use empirical data from the System for Evaluation of Teaching Qualities (SETQ, textbox 1) measurements to provide empirical examples.

**Textbox 1 System for Evaluation of Teaching Qualities (SETQ) data**

The empirical data used in this study’s illustrative examples are teaching performance data gathered through the System for Evaluation of Teaching Qualities (SETQ). This system was developed to evaluate clinicians’ teaching performance in residency training. The SETQ gathers resident-evaluations of their clinician teachers using a tool (questionnaire) that was based on an extensive literature review and discussions with several stakeholders. The system was successfully implemented institution-wide, later nation-wide, and it is currently being implemented in several countries across Europe. The validity and reliability of the SETQ tool is published extensively. Across settings, the SETQ tools had satisfactory psychometric properties. Later, a follow-up study using confirmatory psychometric techniques updated the validity evidence of the SETQ tool and confirmed the high validity and reliability level of the SETQ data. The core tool comprises 20 items that can be categorized into five subscales labelled as **learning climate**, **professional attitude towards residents**, **communication of learning goals**, **evaluation of residents**, and **feedback**. Each subscale contains 3 to 5 items, rated on a 5-point Likert scale (1 = “strongly disagree”, 2 = “disagree”, 3 = “neutral”, 4 = “agree”, 5 = “strongly agree”). In addition to numerical items, residents are asked to provide clinician teachers’ strengths and suggestions for improvement through narrative descriptions. Because the SETQ has proven satisfactory validity for point-wise evaluations, but change assessment using the SETQ has not been studied, the SETQ is an excellent tool to illustrate examples in current paper. Below is a scheme of the evaluation process that is repeated annually.
MEASUREMENT SCALE CHARACTERISTICS

Professional performance assessments often rely on reports, if not perceptions, of observers about performance and can sometimes be subjective. Evaluators may find it hard to categorize performance as “unsatisfactory versus satisfactory” or “poor versus good” and nuances are often desired. In these kinds of evaluations, performance may not be appropriately rated dichotomously or by a limited number of categories. It is therefore surprising that the most widely used metric in performance evaluation tools is a 5-point Likert scale, although 4- and 7-point scales are also common. The validity and reliability of data yielded by a measurement scale depend on several characteristics and can differ across populations and tools. The tendency to use shorter measurement scales (such as 4-, 5- or 7-point Likert scales) instead of longer scales is based on evidence showing such shorter scales have higher reliability (inter-rater, test-retest) and validity (convergent/divergent) compared to longer scales. However, the evidence supporting the use of shorter scales is dated and was mainly based on educational tests that included questions that were knowledge-based (there were “true” and “false” answers and the “true” answer was known by the test developers). A more recent study that explored the reliability and validity of subjective data suggests that longer scales may be more valid and reliable.

The sensitivity of a scale to rate performance appropriately on a performance continuum is important and not always appropriately reflected in the statistical reliability and validity of scale. Research on the sensitivity of measurement scales in professional performance assessments is scarce. For illustrative purpose, we address two issues regarding sensitivity of measurement scales, namely the ability to detect small performance changes and the ability to detect change when the performance distributions are skewed.

Ability to detect small performance changes

Assessing performance change requires a measurement scale that is sensitive enough to capture the changes of effect sizes that can be expected. In performance evaluations of students, performance changes can be large even over limited time, because the learning curve of students is usually
steep. For student evaluations, even shorter scales that are less sensitive to small performance changes may capture changes for most of the students appropriately. In contrast, performance changes in experienced professionals are expected to be small (though relevant), since professionals tend to build up and, sometimes, top their education and clinical experience. For professional performance evaluations, it is thus even more important that measurement scales have the ability to capture even small changes in performance. The ability of a short Likert scale to measure small performance changes is largely unknown, although some questions about this have been posed before.\cite{14,15} A review about measuring quality of life, for example, found that studies using longer scales (10- or 15-point Likert scale) reported effect sizes that were up to twice as large compared to studies using a 5-point Likert scale.\cite{14} The authors argued that these differences were largely caused by the limited sensitivity of the 5-point scale to detect the relatively small changes in quality of life that were observed in most studies.\cite{14} One study in the context of medical education found that the length of a rating scale impacted the percentage of students rated as “having potential problems” or “performs above expectations”, also favoring the use of longer scales.\cite{18} To the best of our knowledge, more research on these issues in medical education has not been conducted. Intuitively, we may understand that short scales (such as a 5-point Likert scale) are less suitable when changes in performance are small. This topic deserves further study and experimenting with longer rating scales should be encouraged, especially since studies explored that raters experience no difficulties with longer rating scales.\cite{15,18}

*Ability to detect change in skewed distributions*

In addition to the ability of measurement scales to detect small performance changes, scales have to provide clinicians room to change their performance scores. Any scale with a pre-specified maximum and minimum has physical boundaries that can impair the measurement of positive or negative change. Clinicians who are able to improve their performance in practice should be able to improve their performance scores on an evaluation tool as well. Most of the professional performance evaluation tools published describe positively skewed evaluation scores, with average scores often around or above 4.0
on a 5-point scale. This implies that a large part of the evaluated clinicians scored well above 4.0 and for them, the 5-point scale simply allows very little positive change in future evaluations. It is unclear whether the limited ability to change performance on the measurement scale corresponds with a limited ability to change actual performance.

To explore if the positive boundary of the 5-point Likert scale impairs the ability to detect performance improvement, we tested this using empirical data on clinicians’ teaching performance. We created two subgroups of clinician teachers; the first subgroup of clinicians scored above 4.0 on their initial performance measurement and had therefore limited room to change scores positively. The second subgroup scored below 4.0 on their initial measurement and had sufficient room to improve their scores. All clinicians were rated by at least six residents on two subsequent occasions, assuring high measurement reliability of both measurements. For both groups, the average performance change on all performance items (difference between average score of initial and subsequent year) was calculated. To obtain a “global indication of clinicians’ performance change” we also asked residents who worked with a clinician for over a year to indicate whether clinicians changed their teaching performance in the last year (possible answers were: 1= worsened, 2= same, 3= somewhat improved, 4= strongly improved). The average performance change was calculated for both subgroups (table 1).

Table 1 indicates that the average scores of the subgroup of clinicians who scored below 4.0 on the initial measurement were enhanced by 0.11 point. In contrast, the subgroup of teachers who scored above 4.0 on their initial measurement dropped in performance score by 0.09 point. Further, residents indicated comparable global performance change for both subgroups (global improvement). The global improvement was expected for the subgroup below 4.0 (who enhanced their performance scores), but was surprising for the subgroup above 4.0, who dropped in performance score.
Table 1: average item and global performance change for clinician teachers scoring above 4.0 and below 4.0 on an initial measurement

<table>
<thead>
<tr>
<th></th>
<th>Initial measurement below 4.0</th>
<th>Initial measurement above 4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of clinicians included</td>
<td>932</td>
<td>500</td>
</tr>
<tr>
<td>average performance score change on all performance items (0 = no change)</td>
<td>+0.11</td>
<td>-0.09</td>
</tr>
<tr>
<td>Average global indication of performance change (0 = no change)</td>
<td>+0.25</td>
<td>+0.27</td>
</tr>
</tbody>
</table>

* this item has been rescaled (originally a score of 2 corresponded to “unchanged performance”)

These results can be seen as an indication that the upper-limit of the 5-point scale actually impairs the ability to capture positive performance change in the high scoring clinicians. When handling skewed performance data, like in most performance evaluation systems, extreme scores (that are close to the positive or negative boundary) will not only apply to a few outliers, but applies to a large proportion of the evaluated clinicians. Future evaluations should aim to limit ceiling effects (positively skewed distributions) by either altering the measurement scale or formulate the performance items differently.

SUMMARIZING CHANGE IN PERFORMANCE SCORES

In the previous section we showed the impact that measurement scale characteristics can have on performance and change assessments. In reporting performance data to clinicians, the data often need to be summarized. There are several summary statistics that can be used for a Likert scale. By far the most commonly used statistics are median and mean differences. Although some argue that mean scores are not appropriate to summarize ordinal data, they are frequently used. We will now illustrate the use of median and mean scores in computing change indicators for two clinician teachers.
Ten residents evaluated two clinicians (clinician A and B) in our study on the performance item “regularly provides trainees with constructive feedback” for two subsequent years (table 2). For clinician A, the median score suggests unchanged performance while the mean score suggests that clinician A improved his performance considerably. For clinician B the median score suggests improvement for, while his mean score suggests unchanged performance (table 2). These simple examples show that two commonly used summary statistics can result in the different conclusions regarding performance change. Because the median and mean may not always provide the information that informs clinicians, researchers and policy makers most appropriately, additional statistics are necessary. In many performance domains, the proportion of cases in which a clinician performed substandard or satisfactory will be valued more than the median or mean performance. If we consider the examples of clinician A and B again, we may assume that for the evaluated performance item a score below 3, i.e. 1 or 2 on the 5-point scale, is considered unsatisfactory/substandard and a score of 4 or 5 is considered satisfactory. A score of 3 could, perhaps, be considered neutral. Expanding the performance metrics illustrated in the previous paragraph with the proportion of residents that evaluated clinicians’ performance substandard and the proportion who evaluated the performance as satisfactory provides another look at the performance change of clinician A and B (table 2).

As shown in table 2, the four change statistics will lead to completely different conclusions regarding performance improvement for clinician B. For clinician A, most of the statistics will lead to the conclusion that this clinician improved his performance, however, the magnitude of performance change (which is important as an effect size measure) differs considerable between the statistics. The kind of information that is desired by clinicians, researchers, directors or policy makers will determine which (combination of) statistics will provide the most appropriate summary of clinicians’ performance.
Table 2 performance scores and summary statistics for clinician A and B

<table>
<thead>
<tr>
<th></th>
<th>Clinician A</th>
<th>Clinician B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Year 1</td>
<td>Year 2</td>
</tr>
<tr>
<td>Raw perfor-</td>
<td>[1, 2, 2, 3, 3, 4, 4, 5]</td>
<td>[2, 3, 3, 3, 3, 3, 4, 4, 5, 5]</td>
</tr>
<tr>
<td>mance scores</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Summarizing metric**

<table>
<thead>
<tr>
<th></th>
<th>Clinician A</th>
<th>Clinician B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median perfor-</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>mance score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean perfor-</td>
<td>2.9</td>
<td>3.5</td>
</tr>
<tr>
<td>mance score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion of</td>
<td>40%</td>
<td>10%</td>
</tr>
<tr>
<td>substandard</td>
<td></td>
<td></td>
</tr>
<tr>
<td>scores (&lt;3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion of</td>
<td>30%</td>
<td>40%</td>
</tr>
<tr>
<td>satisfactory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>scores (&gt;3)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**PERFORMANCE CHANGE IN ITEM, SCALE AND OVERALL PERFORMANCE SCORES**

Performance evaluation tools evaluate different domains of clinicians’ professional performance. Performance domains, such as professionalism, communication, interpersonal relationships, management, and teaching, are usually captured by one or more performance scales of an evaluation tool. These performance scales are measured by multiple performance items. Since only the items of an evaluation tool are measured (domains and scales are latent), and only the items contain concrete behaviors, actions or attitudes, it can be argued that the actual performance change happens at item level. Most evaluation tools contain many items so it may take some time to assess performance and performance change for all items. Therefore, scales, domains or overall performance are often the levels of performance assessment. However, estimating domain or overall performance scores can be challenging.
By far the most applied method to obtain performance domain scores is averaging the performance items that belong to a certain performance domain.\textsuperscript{12} This method assumes that all performance items are of equal importance for a performance domain, an assumption that seems implausible in practice. In previous studies students, residents, coworkers and patients were able to prioritize the performance items that belonged to a domain easily.\textsuperscript{13,26,27} Therefore, calculating domain scores by averaging all items seems an approach that is too simplistic.\textsuperscript{12} Likewise, not all performance domains may be equally important for clinicians’ overall professional performance.\textsuperscript{28,29} Thus, simply averaging all performance domains to obtain an overall performance score is simplistic too. This highlights that the information practitioners desire cannot always be obtained by domain or overall performance scores. Carefully looking at item data and weighting the importance of the items for a specific domain is crucial for appropriate interpretation of performance evaluations.

**ESTIMATING CHANGE OR ASSESSING CHANGE RETROSPECTIVELY**

Apart from calculating change from two time-indexed performance evaluations, it is also possible to assess clinicians’ performance change retrospectively, by asking colleagues they worked with over a certain time period to indicate if a clinician improved his/her performance. As shown in table 1, these two methods yield different change results for high and lower scoring clinicians. Even more, the correlation between the scores yielded by these two methods is low, even for the group who scored below 4.0 (table 3).
Table 3 correlation between the average performance change (calculated from two time-indexed performance evaluations) and global performance change assessed retrospectively, separately for clinician teachers scoring above 4.0 and below 4.0 on an initial measurement

<table>
<thead>
<tr>
<th>Initial measurement below 4.0</th>
<th>Initial measurement above 4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of clinicians included</td>
<td>932</td>
</tr>
<tr>
<td>Pearson parametric correlation coefficient*</td>
<td>0.23 (p&lt;0.01)</td>
</tr>
<tr>
<td>Kendall rank order correlation coefficient*</td>
<td>0.16 (p&lt;0.01)</td>
</tr>
</tbody>
</table>

* the correlations are of comparable size when only the residents that evaluated the same clinician twice were included.

One of the explanations for this surprising finding may be that changes on certain performance items are more important than changes on other performance items. Therefore, the average change over all items may be a poor indicator for the perceived global performance change. This again highlights that the averaging method to calculate overall performance scores is probably too simplistic to assess performance and performance change appropriately. Previous studies on performance measurement at one point in time showed that global indications of performance had high overlap with average overall performance scores (correlations of 0.50 up to 0.98). The current analysis shows that this does not account for estimates of performance change.

PERFORMANCE WITHIN THE CONTEXT

A final issue to address is the role of context in assessing performance. Differences between specialties, departments, hospitals or countries can result in different perceptions of performance. For example, in hospitals with a focus on teaching, the professional infrastructure and learning climate can facilitate clinician teachers to excel in teaching, leading to overall high teaching scores among most teachers. In hospitals with limited teaching facilities
it can be harder for clinician teachers to obtain high teaching performance scores. Therefore it is important to interpret and assess performance within its context, i.e. assess how far item, scale and overall scores of clinicians are from their group means or medians. To exemplify this, we displayed the scores of all clinician teachers of three obstetrics and gynecology (OB-GYN) departments at different teaching hospitals, who were evaluated on the performance domain communication (figure 1). If we consider a clinician with a performance score of 3.5, he would be among the lowest scoring clinicians in department 1, somewhat above average in department 2, and among the top scoring clinicians in department 3. Irrespective of the factors that may have caused the differences between these three departments (which is another interesting topic), this example shows the importance of considering performance scores within their context.
In addition to performance scores being context dependent, the importance of performance items or domains may also differ between settings. For example, doctor-patient communication is very important to primary care clinicians but less important to pathologists. In centers with many trainees, clinicians will have fewer direct patient contacts due to their role as trainers. Consequently, the importance of clinicians’ interpersonal behaviors towards patients will be proportionate to the amount of patient interactions they have.

**CONCLUSION**

Critically appraising the complexity of performance evaluation data is crucial when using performance evaluations and assessing change. Future performance evaluation tools and checklists should consider measurement scale length carefully and methods to handle skewed data should be used. Depending on the context and performance domain evaluated, the most appropriate (combination of) summary statistics should be chosen. Finally, the interpretation of evaluations should embrace the fact that not all performance items equally contribute to the evaluated performance domain and that the context may impact the importance of and scores on performance evaluations.
REFERENCES


