The Influence of Assessor Training on Scoring and Feedback: A Qualitative Study of Assessor Perceptions

by

Kathryn Rae Hodwitz

A thesis submitted in conformity with the requirements for the degree of Master of Science
Institute for Health Policy, Management, and Evaluation
University of Toronto

© Copyright by Kathryn Rae Hodwitz 2018
The Influence of Assessor Training on Scoring and Feedback: A Qualitative Study of Assessor Perceptions

Kathryn Rae Hodwitz

Master of Science

Institute for Health Policy, Management, and Evaluation

University of Toronto

2018

Abstract

Assessor training is essential for defensible assessments of physician performance. This study explored the perceived influence of training and assessment tools on assessors’ conduct of assessments. Using a constructivist grounded theory approach, we interviewed 13 assessors about the effects of training and scoring tools on their conduct of assessments. Participants reported that training taught them about the potential variability of assessor judgement and prompted more reflective scoring. Using the new scoring tools reportedly led participants to sample performance data differently, choose ratings in accordance with the scoring tools, and provide more detailed feedback to assessed physicians. Participants did not perceive training or the tools to substantially affect their numeric ratings, but noted an increased confidence in scoring. Participants involved in developing the tools reported more pronounced effects of training and the tools. These findings support the continued study of assessor training programs as a method of enhancing assessor consistency.
# Table of Contents

## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>v</td>
<td></td>
</tr>
</tbody>
</table>

## List of Appendices

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>vi</td>
<td></td>
</tr>
</tbody>
</table>

## Chapter 1 Introduction

1. Workplace-Based Assessments ................................................................. 1
2. Assessor Training ......................................................................................... 1
3. Opportunities for Research ......................................................................... 4
4. Present Study ............................................................................................... 6
   4.1 Context...................................................................................................... 6
   4.2 WBA Development Process ....................................................................... 6
      4.2.1 Assessment Tool Development / Performance Dimension Training (PDT) ....... 7
      4.2.2 Frame-of-Reference Training (FORT) ......................................................... 9
      4.2.3 External Review of Assessment Tools ....................................................... 10
      4.2.4 Program Implementation and Evaluation .................................................. 10
   4.3 Focus of Present Study ............................................................................... 12

## Chapter 2 Methods

5. Epistemology .............................................................................................. 13
6. Methodology and Design .............................................................................. 13
7. Methods ...................................................................................................... 14
8. Sensitizing Concepts .................................................................................... 15
9. Analysis ....................................................................................................... 15
10. Reflexivity and Worldview ......................................................................... 17
Chapter 3 Results ................................................................................................................................. 19
11 Participants ........................................................................................................................................ 19
12 Findings ............................................................................................................................................. 19
   12.1 Perceived Effects of Training ......................................................................................................... 20
   12.2 Perceived Effects of the Tools ......................................................................................................... 25
   12.3 Differences in Perceived Effects based on Type of Training .......................................................... 31
   12.4 Perceptions of Improved Standardization ...................................................................................... 34
   12.5 Perceptions of Assessor Role .......................................................................................................... 36
13 Analytical Summary and Synthesis ..................................................................................................... 39
14 Theoretical Contribution ...................................................................................................................... 42

Chapter 4 Discussion ............................................................................................................................... 43
15 Discussion of Findings .......................................................................................................................... 43
   15.1 Perceived Effects of Training .......................................................................................................... 43
   15.2 Perceived Effects of the Tools .......................................................................................................... 46
   15.3 Perceptions of Standardization ....................................................................................................... 47
   15.4 Perceptions of Assessor Role .......................................................................................................... 49
16 Implications and Future Research ........................................................................................................ 50
17 Strengths and Limitations .................................................................................................................... 53
18 Conclusion ........................................................................................................................................... 55

References ............................................................................................................................................... 56

Appendix A: Sample Assessment Tools for Family Medicine ............................................................... 64
List of Tables

Table 1. Initial Interview Guide

Table 2. Supporting data for ‘Perceived Effects of Training’

Table 3. Supporting data for ‘Perceived Effects of the Tools’

Table 4. Supporting data for ‘Differences in Perceived Effects’

Table 5. Supporting data for ‘Perceptions of Improvement Standardization’

Table 6. Supporting data for ‘Perceptions of Assessor Role’
List of Appendices

Appendix A: Sample Assessment Tools for Family Medicine
In this thesis, I present a qualitative study exploring the perceived effects of training on assessors’ subsequent conduct of assessments. I focus on the perceptions of assessors who conduct workplace-based assessments of practicing physicians, exploring what assessors perceive to learn through training, how they implement what they have learned when conducting assessments, and how they utilize scoring tools to make evaluative judgements. This exploration sheds light on, and provides a foundation for further study of, if, how, and why training influences assessors’ behaviours and cognitive processes. The results contribute to an understanding of the effects of training on assessors’ scoring and feedback processes and provide insight into the development of future assessor training programs.

1 Workplace-Based Assessments

Workplace-based assessments (WBAs) are a method of evaluating physician performance. They assess how physicians perform in their own work settings, either directly through observation or retrospectively through reviews of administrative data, patient records, or interviews with colleagues and patients (1). They are used in both postgraduate training (2) and increasingly with physicians in practice (1,3,4) as a way to ensure and improve physician performance (5).

As a method of ensuring performance, WBAs can be used to evaluate whether or not competencies have been achieved or standards of practice have been adhered to. This is considered the summative component of WBAs, or “assessment of learning”. WBAs can also provide an opportunity for physicians to receive feedback about their strengths and weaknesses, promoting learning and encouraging practice improvement. This represents the formative aspect of assessments, or “assessment for learning” (6,7). The assessment of physicians within their own workplace therefore allows for an authentic evaluation of actual performance and the provision of individualized and contextualized feedback about that performance.

2 Assessor Training

Given the complexity of physician performance, WBAs necessitate expert judgement (5). Unlike tests of knowledge or discrete skills, WBAs require assessors to interpret complex behaviours in
unstandardized, real-world environments (5,8,9). In these circumstances, assessors need to use professional judgement to evaluate the nuances of performance and to account for a physician’s context, such as their work environment and patient population (3,9,10). They also must translate assessment findings into feedback, the key component of assessment for learning and practice improvement (5,11–13).

While medical educators acknowledge the inherent subjectivity of performance assessments (5,10,14), there is ongoing concern about the variability of assessor judgement and the potential unreliability of WBAs (15–17). Human judgement is subject to a number of limitations such as cognitive bias (18), stereotyping (19), self-referencing (assessors comparing students to themselves) (20), contrast effects (assessors comparing students to other recent assessments) (21) and cognitive load (22). Assessors also tend to vary in what they attend to during assessments and how they weight their observations, which can lead to vastly different assessments of performance (23). Concern about assessor reliability has spurred a growing body of research on training methods aimed at reducing unwanted variability among assessors. Two training methods that have shown promise are Performance Dimension Training and Frame-of-Reference Training.

Performance Dimension Training (PDT) is based on the premise that assessors judge performance as they observe it rather than afterwards when they provide ratings. Therefore, the aim is to train assessors to draw upon established performance dimensions during assessments by orienting them to these dimensions (e.g., rating scales) before they conduct assessments, or by including them in the development of the assessment tools (24). For example, assessors may collectively define the components of competence for specific domains and establish the scoring criteria that will differentiate performance (25).

Frame-of-Reference Training (FORT) builds on PDT by establishing a common set of standards that will be used to rate performance through an iterative process of practice and feedback (26). Assessors either develop behavioural examples for different levels of performance or are provided with them, then practice using these as standards to rate performance and receive feedback on how closely their ratings align with the agreed upon standards (24). This often involves assessors observing or conducting mock assessments in a group setting and then discussing how and why they each made their ratings; a facilitator highlights why ratings may
differ across assessors and helps to build a common frame of reference for assessors to use when rating performance based on the behavioural standards (25). Through practice and feedback in using the established frame of reference, it is intended that assessors calibrate their scoring to the appropriate standards and then use these common standards when assessing performance subsequently (26,27).

These complimentary training methods were developed and validated in the field of organizational psychology and have been found to significantly improve rating accuracy in this context (24,26). They have only recently been applied in medical education settings, specifically for faculty who conduct direct observation assessments of residents. Holmboe et al. (25) found that PDT and FORT increased the reliability of faculty’s ratings of videotaped resident-patient encounters, with improvements sustained eight months after training. Conversely, Cook et al. (28) found that these training methods did not improve accuracy or reliability of faculty ratings of either videotaped encounters or of actual resident-patient encounters. Both studies noted that faculty assessors valued the training process and found it increased their confidence in conducting assessments, despite their conflicting findings regarding assessor scoring.

In response to these contradictory findings, Kogan et al. (29) qualitatively explored faculty’s responses to PDT and FORT. Participants stated that they were willing and able to shift to a new frame of reference, reflected in the scoring tools. They also stated that the tools facilitated their observations, enhanced their feedback vocabulary, and served as valuable educational tools for learners. Kogan et al. also compared the responses of participants who were involved in developing the scoring tools with those who were given pre-defined scoring frameworks and found that those involved in inductively developing the tools responded most positively to training.

Collectively, the results of these studies suggest that PDT and FORT may be effective training methods for assessors in medical education settings, and that involvement in tool development may be particularly valuable for assessors, however, additional research is needed to further examine the effects of PDT and FORT in other settings.
3 Opportunities for Research

Assessor training using PDT and FORT is a relatively new area of study within medical education. Inconsistent findings about the effectiveness of these methods in this context suggest that more research is needed to examine if and how such training programs influence assessors’ ratings. Specifically, exploration is needed into what is learned in training and how this learning is transferred to practice. Roch et al. (26) suggest that FORT research should shift away from examining the effects of training on rating accuracy and towards understanding the extent to which what is learned in training transfers into the assessment context, focusing on the reactions, motivations, and perceptions of assessors. Qualitative methods are well suited to study perceptions and subjective meaning, and are designed to answer questions such as how or why training influences assessors’ ratings (30). A qualitative exploration of what assessors perceive to learn through training and how they perceive this to affect or not affect their conduct of assessments may provide insight into if, how, and why training influences assessors’ behaviours and cognitive processes.

Further research on assessor training is particularly needed in the context of WBAs of practicing physicians as previous studies have focused exclusively on the assessment of students by faculty through direct observation. The ways in which PDT and FORT may apply to other types of assessments or other contexts is unknown. Generalizability of the current literature to physician assessments is tenuous since the expectations and competencies of practicing physicians are different than that of students (31), thus the assessment tools, processes, and outcomes are also distinct. Additionally, faculty members typically have ongoing relationships with assessed students, whereas assessor of physicians often do not; these differing relationships play a role in how performance is evaluated and how feedback is provided and received (1,32–34). Given that WBAs are increasingly used in practice settings, and often include methods other than direct observation (1,31,35), the examination of assessors who evaluate practicing physicians through retrospective means such as chart reviews or interviews is a considerable gap in the literature.

Additionally, while medical education research has begun to examine the effects of training on assessor ratings, research has not explored how assessors integrate this training with their own professional judgement during WBAs. It is acknowledged that assessors must utilize subjective judgement during assessments, and that they should use structured tools and undergo training in
order to reliably and defensibly interpret physician performance (5,10,14,36); however, assessors’ experiences of integrating these external frames of reference with their own internal perspectives during assessments has not been studied. Exploring how assessors draw upon training and assessment tools while making subjective judgements, and the challenges they may face in doing this, may shed light on why assessor training programs have been inconsistently effective.

Understanding if, how and why training influence assessors is important for the development of effective training programs. It is also crucial for the validation of WBAs. All assessment programs must be continually validated to be defensible, which is to say that evidence must be collected to support the use of the assessment for its intended purpose (37). There are many sources of validity evidence, which must be prioritized based on the purpose and stakes of the assessment. In performance assessments, which rely on assessor judgement, the way that assessors conduct assessments, use assessment tools, and give ratings and feedback, is critical. The validity of performance assessments is typically related more to the assessors themselves than the tools they use (5), as it is assessors who interpret performance and assign scores.

Overall, assessors must use subjective judgement during performance assessments, and they must also produce fair and defensible evaluations of physician performance. Particularly when assessments result in summative decisions, or when judgements are provided by a single assessor, there is a need to promote assessor consistency and reduce idiosyncratic or arbitrary ratings. Given that assessors tend to have unique, and often ambiguous, frames of reference or internal criterion against which they make performance judgements (23,32,38), bringing these to light, building a common frame of reference through consensus, and discussing different assessment approaches through facilitated discussions seems promising for promoting assessor reliability. However, research in this area is mixed and contains a number of gaps; further exploration is needed into the perceived effects of training. Adding to our understanding of assessors’ scoring behaviours has clear implications for the development of effective training programs, the validation of WBAs, and a broader understanding of the effects of training on assessor cognition. The goal of this growing field of research is to enhance the quality of assessments in the workplace, thereby providing physicians with more meaningful feedback about their practices and, ultimately, improving the care they provide to patients.
4 Present Study

4.1 Context

The present study builds off previous research by studying the effects of PDT and FORT on assessors’ rating behaviour, with a focus on if and how assessors perceive training to have influenced their scoring. This study is part of a larger evaluation of a WBA program delivered by the College of Physicians and Surgeons of Ontario (CPSO).

The CPSO is the medical regulatory authority in Ontario, legislatively responsible for ensuring the competence and continuous improvement of physicians licensed in Ontario. As part of this mandate, the CPSO randomly selects physicians each year to undergo Peer Assessments, which are WBAs designed to ensure and promote practice improvement. These assessments are conducted by physician assessors practicing in the same specialty as the assessed physician and involve a review of patient records and an interview. Assessors complete reports about the physician’s practice which are then reviewed by a Quality Assurance Committee. This committee, comprised of physicians and public members, decide the outcome of the assessment, specifically whether formal follow up is needed (e.g., a reassessment).

The CPSO has administered Peer Assessments since the 1980s, but in 2012 began to redevelop the program to make it more educationally valuable for physicians. This coincides with the broader movement in medical education of emphasizing the formative aspect of assessments (7). The redevelopment of the Peer Assessment program has been described and published elsewhere (39). A summary of the redevelopment process and accompanying program evaluation are included below.

4.2 WBA Development Process

Redevelopment was approached from a constructivist perspective which recognizes the subjectivity inherent to performance assessments (9,40,41). The team acknowledged that assessors have unique perspectives from which they interpret performance such that variation among assessors represents meaningful variation rather than error (42). However, the perspective that assessors are “trainable” was also adopted, which emphasizes that scoring rubrics and training can be used to promote consistency in assessor judgement (42). While these perspectives are distinct, they can overlap. Particularly since CPSO assessors conduct assessments
independently rather than providing one of many ratings in a programmatic assessment model, it was important to both acknowledge subjectivity and build consistency in scoring for the reliability, defensibility, and fairness of the program.

Redevelopment was led by two researchers in the Research and Evaluation Department at the CPSO: one Senior Researcher with a background in cognitive neuroscience and one Research Associate with training in health services research and experience in program development and evaluation. This research team collaborated with peer assessors throughout the redevelopment process. Peer assessors are practicing physicians who were assessed themselves before becoming an assessor in the CPSO peer assessment program. Upon being assessed, they either volunteered or were recruited to become an assessor because they had an exemplary practice.

First, specialty-specific working groups of peer assessors were formed. These working groups of approximately six assessors each were involved in developing the first draft of the assessment tools for their discipline. All assessors within a given specialty then received an orientation to the tools and were given the opportunity to provide feedback and revise to the content. This comprised the ‘Performance Dimension’ portion of training. All assessors then underwent Frame of Reference training, either as one large group or, in cases where there were more than 30 assessors in a specialty, in multiple groups. The assessment tools underwent an external review by all Ontario physicians in a given specialty, as well as selected specialty organizations, before being implemented. Details of the program redevelopment and training processes are described below.

4.2.1 Assessment Tool Development / Performance Dimension Training (PDT)

The Peer Assessment framework and tools were inductively derived by assessors, who are both content experts and the end users of the tools. This was accomplished through two phases of PDT. The first phase involved a subset of six to eight assessors from each specialty developing all the assessment tools and processes that would support the new program. This occurred iteratively over approximately one year, with assessors convening monthly or bi-monthly, facilitated by members of the research team. Traditionally, PDT only involves the development of ratings scales, but PDT in this study included the development of a number of other assessment components.
First, assessors developed an assessment framework consisting of eight assessment domains that would be common across most specialties (some proceduralist specialties such as Anesthesiology, which do not follow a Family Medicine model, required a unique framework). Assessors then iteratively generated the elements of quality care that would comprise each assessment domain for their specialty. A three-point global rating scale was developed to rate performance with anchors related to the amount of quality improvement needed. Performance examples were then created for each rating across all assessment domains. Collectively, these comprised the scoring rubrics to be used when assessing physician performance. They were intended to align scoring with the program’s focus on quality improvement (43,44), to promote consistency (45), and to support the formative value of the assessments by transparently outlining high quality physician performance (46).

Assessors then developed the process for selecting patient records and the discussion themes for the physician interview component to standardize the collection of performance data. They helped create the assessment report template, comprised of a global rating and narrative comments for each domain. Narrative comments were emphasized to provide context for the scores and formative feedback for the assessed physician (47,48). With input from the QAC, each working group also created a sample assessment report for their specialty to serve as an example of the quality of narrative comments that should be included in the reports. Lastly, quality improvement resources were developed that define high quality care for specific patient presentations or therapeutic approaches. These are brief, at-a-glance documents designed to be a resource for assessors and educational materials for assessed physicians. A sample of the assessment tools developed are included in Appendix A.

The second phase of PDT involved all assessors in a given specialty (those who were involved in tool development as well as those who were not) receiving a two-hour orientation to the new assessment tools. Orientation sessions were led by the research team, with involvement from the assessors who developed the tools. Sessions included an overview of the purpose and goals of the redeveloped program, namely consistency in assessments, transparency in evaluation criteria, and increased focus on the formative purpose of the program. It also provided an orientation to the newly developed tools, including the standards defined in the scoring rubrics and the levels of performance that would constitute each scores on the rating scale. Afterwards, all assessors
could then review the materials on their own and provide revisions and feedback via an online survey platform, which were then incorporated into the tools.

PDT was divided into two phases to facilitate the tool development process. Given that some specialties have many assessors (over 100 in Family Medicine), working with a small subset to develop draft tools and then obtaining input and feedback from the remaining assessors was more manageable than trying to incorporate everyone’s input initially. The phased PDT approach meant that all assessors had the opportunity to contribute to the tools, but did not need to be involved in all stages of the process.

4.2.2 Frame-of-Reference Training (FORT)

Frame of Reference Training sessions were held with all assessors, by specialty. Sessions were approximately 5 hours in length. The number of assessors in each training session ranged from 5 to 30, depending on the size of the assessor group. Groups with more than 30 assessors (e.g., Family Medicine) were trained in multiple sessions. Assessors were convened in person when possible, but could join by webinar if needed. Training was led by one member of the research team, with support from a lead assessor in instances with large assessor groups.

An overview of the tools and processes was first provided to remind assessors how records should be selected and reviewed, how the interview should be conducted, how the scoring rubrics should be utilized to make ratings, and how the narrative reports should be written. These tools and processes were explicitly linked to the purpose of the program, specifically its formative focus on the quality improvement of assessed physicians.

Assessors then completed an exercise wherein they reviewed three to four simulated patient records and used the scoring rubrics to rate the quality of care represented in the records. Assessors rated each domain anonymously using clicker technology (i.e., interactive software that tallies and presents responses entered on clickers) and were then presented with the distribution of scores for all assessors. Upon seeing their level of consistency, assessors would then engage in a facilitated discussion about why they each made their ratings and have the opportunity to rate again. This process occurred iteratively until agreement was met, defined as 80% of assessors agreeing on a score. This training session enabled assessors to iteratively come to consensus on how performance should be scored and discuss what the collective expectations
should be for certain types of care (e.g., how frequently mental status exams should be documented within psychiatric practices). It also provided an opportunity to modify the tools for increased clarity.

The training provided in this study was based on FORT (practice and feedback in using the tools), but also included principles of consensus methodology, namely anonymity of rating, iterative rounds of scoring, controlled feedback, and calculation and presentation of the group’s response (49,50). By providing assessors with an opportunity to practice using the scoring rubrics, enabling anonymous scoring, presenting aggregated results, providing feedback about their rating, and allowing for discussion with other assessors, it was intended that assessors would learn how to use the scoring rubrics and calibrate their scoring to the rubrics through consensus. This was thought to be achieved once assessors were using the rubrics to choose ratings and 80% agreement in scores was obtained.

4.2.3 External Review of Assessment Tools

Once the tools were developed and piloted through training, they were reviewed externally by all physicians in Ontario as well as selected physician organizations to ensure stakeholders found the assessment criteria to be acceptable. First, all physicians in a given specialty were sent a link to an online survey where they could view the scoring rubrics and provide feedback about the clarity and appropriateness of the assessment criteria. Then, relevant physician specialty organizations (e.g., the Ontario Association of Cardiologists) were contacted to review and provide feedback about the tools. The results of the external review supported the acceptability of the tools by stakeholders (12,51) and provided information about further modifications to the tools for increased clarity and relevance, which were reviewed and incorporated by assessors as appropriate.

4.2.4 Program Implementation and Evaluation

The new program was implemented in early 2017. A comprehensive program evaluation is concurrently being conducted to examine the utility and impact of the program on stakeholders including CPSO staff, Quality Assurance Committee members (i.e., decision makers), assessors, and assessed physicians. The current focus of the evaluation is on understanding how the program operates; in this case, how assessors conduct assessments. This aim is of particular
interest given that assessor involvement in tool development and training was pivotal to the redevelopment initiative. Also, understanding if and how assessors make use of the training and tools provided to them is important for determining whether observed outcomes, such as physicians’ subsequent practice improvement, can be attributed to the activities and features of the program (52).

Program evaluation, in the context of assessments, can be considered a form of validation. It is the ongoing process whereby evidence is collected and appraised to demonstrate that an assessment program is achieving its intended purpose (43). Validity frameworks should be used to guide the validation process so that it is done systematically and comprehensively (36,53). Kane’s validity framework (37), for example, organizes validity evidence according to four inferences: scoring, generalization, extrapolation, and implications. Collecting evidence to support these inferences ensures that the way assessment performance is scored is accurate, that it is generalizable to all of a physician’s assessment performance, that it extrapolates to real world performance, and that it contributes to the intended outcome of the program.

Kane’s framework was used to guide the redevelopment of the Peer Assessment program and will be used to inform the ongoing evaluation and validation activities (39). An examination of how assessors provide scores and feedback to physicians will contribute to the broader program evaluation and provide important validity evidence for scoring (the accuracy, reliability, appropriateness, and fairness of assessment scores) and implications (the impact of the program and assessment decisions on stakeholders) (36,37). Specifically, understanding how assessors use scoring tools and draw upon training principles during assessments will offer insight into whether the scores stemming from assessments are reliable and fair. Assessors’ perceptions of how they score and provide feedback will also indicate whether assessment decisions will be defensible and whether the assessed physicians will find educational value in the program. This, in turn, will shed light on whether the program is achieving its intended purpose of ensuring and improving physician performance, supporting defensible decision making by the QAC and promoting physician practice improvement. While validation is an iterative and ongoing process, and requires multiple sources of evidence, the perspectives of assessors as they relate to scoring and implications contribute valuable information about the effectiveness of training, the conduct of assessments, and the potential impact of the program.
4.3 Focus of Present Study

The purpose of the present study is to explore assessors’ experiences of being involved in PDT and FORT and conducting assessments of practicing physicians subsequently. We will not examine rating accuracy, but rather assessors’ perceptions of what they learned in training and how this, and the newly developed tools, influenced the way they conduct assessments, make evaluative judgements, and provide scores and feedback to assessed physicians. The overarching research question guiding this study is: What is the perceived influence of training and consensus-derived assessment tools on assessors’ conduct of workplace-based assessments?

This study will add to the literature by offering an examination of what assessors perceive to learn from training and how they apply this to WBAs of practicing physicians. It will also contribute to the validity argument supporting the CPSO’s assessment program by providing evidence about assessors’ scoring behaviours and the potential implications of the program. As WBAs are increasingly used to promote physician practice improvement, it is critical that assessment programs are comprehensively and transparently validated for this purpose. Examining the assessors who conduct WBAs is an essential part of this endeavour.
5 Epistemology

In alignment with the epistemological orientation of the development process, this study was conducted from a constructivist perspective, which assumes that there are multiple legitimate interpretations of the same event (40). Research within the constructivist paradigm focuses on the subjective meaning that participants assign to events and seeks to explore individual perspectives about experiences. As such, we sought to explore the potential plurality of experiences that assessors may have of being involved in the redevelopment process, undergoing training, and drawing upon (or not drawing upon) these experiences while conducting assessments. We did not intend to determine the “truth” of whether involvement in tool development and training “works” or does not “work”, but rather to develop an understanding of how assessors uniquely respond to training and whether and how they integrate consensus-derived scoring materials with their own professional judgement during assessments.

6 Methodology and Design

We employed a constructivist grounded theory methodology (54,55). Grounded theory seeks to inductively generate theory from data through an iterative and simultaneous process of data generation and analysis. As data are collected, it is analyzed through constant comparisons of emerging themes with new data to refine themes and direct ongoing data collection. Sampling is purposive and theoretical such that as the emergent theory evolves, data is sought out to confirm, refute, or fill gaps in the evidence (56).

Classic grounded theory is considered post-positivist in orientation as it follows a structured analytical process that relies on the assumption that a neutral researcher “finds” themes that exist within and emerge from the data (56). Constructivist grounded theory differs from this by following a more flexible analytical process that acknowledges and supports the interpretive nature of qualitative inquiry (55). From this perspective, it is acknowledged that themes do not emerge on their own, rather they are generated through a process of interpretation that the researcher is fundamentally a part of.
Following a constructivist grounded theory methodology, we aimed to iteratively generate a theory about how training affects assessors, specifically what assessors perceive to have learned in training and what effects they perceive this to have on their subsequent rating behaviours.

7 Methods

We conducted semi-structured interviews with CPSO assessors who had been involved in PDT (phase 1 or 2) and FORT, and who had conducted at least one assessment in the new program. The new program was implemented incrementally by specialty, therefore at the time of this study, only the following specialties had been implemented: Family Medicine, Walk-In / Urgent Care, Psychiatry, and Medical Psychotherapy. We used purposive sampling to identify potential participants from each of these specialties and from varying levels of involvement in the training process (i.e., PDT phase 1, PDT phase 2, and FORT).

We contacted prospective participants by e-mail with an invitation to participate in a one-hour telephone interview about their experience with training and the perceived effects of training on their conduct of assessments. The interviewer obtained verbal consent at the beginning of each interview to both participate in the study and to have the interview recorded. Interviews were audio recorded via the teleconference service and transcribed verbatim. We conducted interviews until theoretical saturation was reached. The initial interview guide is included in Table 1. We obtained ethical approval for this study from the University of Toronto’s Ethics Review Board.

Table 1. Initial Interview Guide

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Please describe your involvement in Peer Assessment Redesign.</td>
</tr>
<tr>
<td></td>
<td>a. When were you first introduced to the new program?</td>
</tr>
<tr>
<td></td>
<td>b. Were you a member of the working group who developed the tools?</td>
</tr>
<tr>
<td>2.</td>
<td>Please describe your experience undergoing consensus-building training.</td>
</tr>
<tr>
<td></td>
<td>a. What did you learn through the process? Did anything surprise you?</td>
</tr>
<tr>
<td></td>
<td>b. Did undergoing training change the way you thought about conducting assessments?</td>
</tr>
<tr>
<td>3.</td>
<td>After undergoing this training, how did you approach your most recent assessment?</td>
</tr>
<tr>
<td></td>
<td>a. Did you do anything differently than you used to?</td>
</tr>
<tr>
<td>4.</td>
<td>How did you use the new assessment materials (scoring rubrics, educational materials) during the assessment?</td>
</tr>
<tr>
<td></td>
<td>a. When did you use them? (e.g., before, during, and/or after the assessment)</td>
</tr>
</tbody>
</table>
b. Did you find that they influenced the way you thought about rating the physician?

c. Did they influence the way you provided feedback to the physician? How?

5. Peer assessments rely on the professional judgement of assessors. Can you tell me about how you tried to integrate the use of these new assessment materials with your own perspectives and judgement when evaluating the physician’s practice? What challenges did you face?

6. Overall, what kind of influence, if any, do you think the consensus-building process will have on the way you conduct assessments?

7. Overall, what kind of influence, if any, do you think the new assessment materials will have on the way you conduct assessments?

8. Is there anything else I didn’t ask you that you’d like to tell me about?

8 Sensitizing Concepts

We approached this study with the sensitizing concepts (54) of assessor subjectivity in WBAs (9,40,41), assessor “calibration” through training (10,26), and the dual notion of assessors as both “trainable” and “meaningfully idiosyncratic” (42). We also entered into the study with an understanding of contemporary validity theory and assessors’ role in contributing to scoring and implications evidence (36). These sensitizing concepts informed the design of this study and the development of the interview questions; however, we attempted to “bracket” these concepts during data generation and initial phases of data analysis to remain open to the data. This deliberate act of remaining open was employed to create space for the experiences of assessors to come through as we generated themes inductively, rather than our fitting the data into preconceived notions or frameworks. Once we analyzed all the data inductively, we then deliberately returned to our sensitizing concepts to explore the data in the broader context of assessor training and to generate a theory about how PDT and FORT influence assessor cognitions and behaviours. These initial concepts therefore helped us to situate our study within the established literature and to provide a theoretical contribution to that knowledge base.

9 Analysis

Given that constructivist grounded theory is characterized by concurrent and reflective data generation and analysis, the primary author (KH) iteratively conducted and analyzed interviews so that subsequent interviews could be tailored to explore emergent themes. While attempting to remain open to the data, she first read through each interview to get a sense of the story of each
participant. During this phase, KH employed a narrative approach to analysis which involved reading through the text without fragmenting it and attending to how participants constructed their stories (57); overarching themes about the experiences of training were tracked through individual interview memos. KH then coded each interview inductively using a constructivist grounded theory approach (58), which involved fragmenting the data into meaningful sections, labeling them, and comparing them to each other and to emerging themes.

By using both a narrative and coding approach to analysis, KH intended to gain a more comprehensive understanding of the data. Coding approaches in grounded theory can sometimes lead to data reduction which may result in lists of themes that are presented as “truths”. Narrative approaches, on the other hand, can facilitate analyses that are more interpretive and expansive. Utilizing both approaches on the same data enabled a more comprehensive interpretation of the data (57).

After the first five interviews had been inductively coded, KH used deductive analysis to organize the themes into categories based on the research questions. Through this process, KH created an analytic memo to summarize themes that answered the research questions, as well as other emergent themes, and to highlight where further data was needed to confirm trends or explore emerging ideas. This memo, with supporting data excerpts, was then reviewed by two other members of the research team (AK and RB) and discussed as a group. KH modified the interview guide to hone in on initial trends before conducting subsequent interviews.

This process of interviewing, analyzing, memoing, and discussing as a group continued in two more iterations, after 6 more interviews, then after the remaining 2 interviews. This iterative process allowed analysis to inform data generation, and vice versa, and helped us judge when saturation of themes was reached. The use of multiple analysts during this process enhanced the credibility of the results by allowing for multiple perspectives on the data and reducing the possibility of data elements or trends being overlooked.

After all the interviews had been analyzed, KH examined the results comprehensively to explore the overall stories within the data and to create an overarching analytic memo summarizing all the interviews. KH then re-read each interview to ensure that the memo accurately represented the participants’ experiences and perspectives. KH also explored and compared trends in the data.
based on the participants’ involvement in training, specifically any differences in the responses of participants involved in PDT and those involved only in FORT.

Lastly, we returned to our sensitizing concepts to theorize the results, link the findings with the broader literature, and provide a generalizable story of the findings. Theorizing essentially “triangulates” the data with relevant extant theories (59), facilitating the transferability of the findings to other settings and contributing to a more comprehensive understanding of the phenomenon. As such, we reviewed relevant literature and provided conceptual and practical recommendations for the development of future assessor training programs and suggestions for future research endeavours.

10 Reflexivity and Worldview

Throughout analysis, the researchers were mindful of their pre-existing perspectives and knowledge when examining the data. This reflexivity, or awareness of one’s own positionality, is an integral component of qualitative analysis and contributes to the credibility and trustworthiness of the results (60). While one can never eliminate one’s own perspectives, nor find benefit in doing so, researchers should be aware that they carry with them assumptions, worldviews, and knowledge bases that will influence their interpretation of data. Being cognizant of how these perspectives shape one’s analysis, and attempting to view data from alternative perspectives, enhances the richness and rigour of the analysis.

In this study, the researchers’ constructivist worldviews influenced their approach to the research. Specifically, their worldviews shaped the formation of the research question (i.e., how assessors uniquely respond to training), the design of the interview questions (i.e., open-ended questions to prompt subjective, experiential responses), and their analysis of the data (i.e., exploring the plurality of assessor experiences and the meaning that assessors individually attributed to their experiences). Additionally, the researchers’ knowledge of WBAs, assessor training, assessor cognition theories, and validity frameworks, as well as their professional roles and backgrounds, contributed to how they approached analysis. Most notably, the primary analyst (KH) is an employee of the CPSO and is involved in both the redevelopment and evaluation of the Peer Assessment program. Her nuanced understanding of the program and the redevelopment process provided a valuable perspective during data generation and analysis, as she was able to probe appropriately during interviews and interpret the data within context.
However, her role also represented an area of caution during inductive analysis given that such invested involvement could lead to assumptions about how training “should” affect assessors. To acknowledge this, KH listened to participants’ views about the training process without assuming that the training should have affected them in any particular way. She also attempted to use the language employed by participants during interpretation and memoing so as to not inadvertently misrepresent their perspectives. Involving AK and RB in the analysis also helped to ensure that data was interpreted carefully and comprehensively, and that data elements were not overlooked.
Chapter 3
Results

11 Participants

We recruited 13 assessors to participate in an interview. Six participants were involved in tool development (PDT phase 1) and training (FORT) and the remaining seven were involved in an orientation to the tools (PDT phase 2) and FORT. A number of specialty areas were represented: seven participants practice and assess in Family Medicine, two in Walk-In / Urgent Care, three in Psychiatry, and one in Medical Psychotherapy. Five of the participants were men and eight were women.

Participants had completed their first assessments post-training between June and October of 2017. Interviews were conducted between October and December of 2017 and were up to 60 minutes in length.

12 Findings

Participants reported training to be a valuable learning experience that taught them about the potential variability of assessor judgement. This, in turn, reportedly prompted them to be more reflective about scoring and feedback, motivated them to utilize the scoring rubrics, and caused some to adjust their expectation to align with other assessors. The use of the new tools reportedly led participants to sample and record performance data differently, to choose or confirm their ratings in accordance with the scoring rubrics, and to provide more detailed and specific feedback in their reports. Participants did not perceive training or the tools to substantially affect their ratings, but noted an increased confidence in scoring due to the consensus-derived tools.

Our analyses indicated differences in perceived effects based on whether or not participants were involved in developing the tools, with those involved reporting more pronounced effects. However, almost all participants noted that the new program will be more standardized and consistent than the previous program. Participants also conveyed their perceptions of their role in assessments, supporting the education of physicians and supporting decision making but not being decision makers.
Detailed descriptions of these findings with relevant supporting data are included below according to the following five categories:

1. Perceived Effects of Training
2. Perceived Effects of the Tools
3. Differences in Perceived Effects based on Type of Training
4. Perceptions of Improved Standardization
5. Perceptions of Assessor Role

An analytical summary follows, outlining our synthesis of these findings.

12.1 Perceived Effects of Training

Most prominently, participants noted that through FORT, they realized how variable assessor judgement can be. Almost every participant stated that they were interested and surprised to learn that other assessors had different perspectives than them about performance; that they could interpret and weight information differently and ultimately rate the same performance in a different way. Relatedly, participants stated that they learned where their perspectives were the same or different than other assessors. Learning that one was an outlier seemed to be particularly surprising and enlightening. Participants also reported learning how to use the scoring rubrics to make ratings and learning from each other about different practice environments, clinical contexts, and therapeutic approaches.

By gaining an awareness that other valid perspectives and frames of reference exist, some participants became more mindful of their own scoring and began to think about their judgements in relation to others. During assessments, these participants reported stepping back to think about how other assessors might rate the same performance. This reportedly highlighted the importance of using the structured tools, with some noting their motivation to utilize the scoring rubrics when making ratings. Some participants reported altering their expectations upon realizing that their way of rating certain elements of performance did not fall within the majority of assessor ratings, becoming more lenient or more stringent. Others, when learning that their way of scoring performance did fall within the majority, noted an increased confidence in their scoring. Some participants also altered their approach to feedback. Given that there is variability in how assessors interpret performance, they realized that there may also be variability in how
assessed physicians receive feedback; they noted attempting to gauge physicians’ receptivity and matching their feedback approach accordingly.

Learning about different practice environments also appeared to affect participants’ conduct of assessments. This reportedly increased one participant’s ability to account for physicians’ contexts in assessments and prompted another to learn more about different therapies in order to be better equipped for assessing physicians in practice settings unlike their own.

While learning about assessor variability seemed to have a positive impact on almost all participants, it caused two to question the program’s rigour. The realization that assessors have inconsistent perspectives and judgements caused one to be concerned about the delivery of assessments and motivated another to emphasize the educational aspects of the program, seemingly to lower the stakes of the program from summative to formative. The data to support these findings are displayed in Table 2.

Table 2. Supporting data for ‘Perceived Effects of Training’

<table>
<thead>
<tr>
<th>What assessors reported learning through training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessor variability (realizing assessors judge performance differently than each other)</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
and instinctively want to go for the middle ground. So, it was interesting to see how much, I guess, just how varied people’s attitudes and approaches were.” – P8

“I think it opens your eyes to you can see how different people will think differently… even though there are certain standards, everybody has a different approach sometimes to it. So, it was good to hear everybody’s different points of view.” – P11

“I guess the thing that surprised me is how different assessors can look at one chart and have a completely different opinion. You know, there were some outliers, I think that surprised me.” – P9

| How perspectives compared to others (realizing how one’s judgements compared to other assessors’ judgements) | “It actually had a huge impact… Because, even with all the years of peer assessing that I’ve done and in being involved in the redevelopment, and being involved in making one of the charts, and being involved in one of the QIRs, it was still useful to see where my peers, my colleagues, where we differed in things… And it was really interesting to see that there were one or two spots that I was an outlier, actually. And that kind of really stuck with me.” – P2

“I realised that, actually, my colleagues, quite in general, were much more lenient than I was… That was certainly one thing that really impacted me. That whole process was actually really, really helpful.” – P2

“It was kind of interesting because we had a variety of different assessors with different kind of viewpoints and different levels of experience. And so it was interesting finding out that, oh, maybe I’m being a little harsh in certain areas, or, oh, maybe I’m too lenient in certain areas.” – P4

“I think, a lot of the time, I ended up being with whatever point was the majority… But there were some instances where I was quite off a couple of times. The majority of the people thought… things were okay that I thought were quite mediocre. And that surprised me.” – P8

| How to use scoring rubrics | “It also focused our attention on the scoring rubrics, because we would learn, sometimes people were just still using their gut instinct rather than going through the, using the scoring rubrics that we developed… It reminded all of us to use the scoring rubrics, not just do our gut feel as we had done for many years as assessors.” – P1 |
“After I went through the handbook and then had these two conference meetings where there were case reviews [FORT], I thought that was extremely helpful… because I understood exactly what they were looking for in each section.” – P9

**Other practice contexts**

“That kind of tension, that kind of difference in perspectives is very helpful. So, the academic gets to understand the community and vice-versa… Remote doctors versus all the bells and whistles in Toronto or whatever, too. Very important to understand the world. The contextual set-up for the reports is very important, in my opinion… the environment in which the physician being assessed is working” – P1

“I’m just more aware of the fact that my colleagues come from a different place and the peers I’m assessing are coming from a different place… it’s just allowed me to get a bit more of a global perspective on patients’ backgrounds and the expectations within that clinical setting.” – P3

**Perceived effects of training on subsequent assessments**

**Became more mindful about scoring**

“The influence it will have on me conducting assessments going forward is being reflective, is being more cognizant of people thinking in different ways… I am more sensitised to the fact that some people, when presented with the same set of facts, will interpret it somewhat differently than I am. So, I think it has helped me be more understanding to people’s differences and the way that they’ll interpret things.” – P1

“When I did do the first assessment with the new system for them, I remember being quite thoughtful as to, okay, let’s make sure that is how they would have approached it. And I can remember sort of thinking that. So, instead of just sort of this is what I think as I went through that review with that physician, I remember thinking, hmm, would the other people in the spectrum of family physicians have found this a problem or not. And I think it did colour it for me a bit.” – P4

**Motivated to use scoring tools**

“It let other people know, wow, I just assumed other people would interpret it the way am, and I realised that no, not everybody interprets it the same way. So, it reinforced the importance of standardisation, of following the rules that are set down in the handbook” – P1

“I can’t say it’s changed my approach in any drastic way, I’m probably just a little more careful with trying to utilize the scoring rubric because we’ve had some
| Changed expectations | “I think I’m more hawkish. Not much more, but I think one of the things that the consensus-building process showed me is that there are physicians, colleagues, who are more exacting than I am. I have become a little more demanding in terms of what I want to see in records.” – P10

> “And sometimes, perhaps my expectations were a little bit higher coming from an area of clinical acuity if you will. So, perhaps my expectations were a bit more, in some ways, and so it was helpful for me to hear the discussions of my colleagues and perhaps to tone down some of my expectations.” – P3 |
| Increased confidence | “It was just a good opportunity to reassure myself that the majority of things I was doing were in alignment with what most of my peers were doing.” – P2

> “I didn’t find I was an outlier anywhere, I was sort of middle of the road with most of the general opinions, but there were some people who would rate that sample case worse than it was, and then some doctors would rate it much better than it was… it made me feel confident in how I assess.” – P9 |
| Changed approach to feedback | “It has just sensitised me to the way that people interpret things. And I try to, myself, understand what’s the best way to give feedback to this person. Some people are very businesslike, they just want the facts. Some you need to do the gentle, here, well I really liked what you did here. However, you’ve got to pave the way for it. Yeah. So, it has influenced me to some degree, the way I’m conducting, giving the feedback component of the assessment.” – P1

> “When I’m giving them feedback, I sound them out about whether or not they’re interested in the feedback or not. And, if they’re interested, I go forward with it and we talk about whatever… if it doesn’t resonate, then I would mention it. If it does resonate, then we would really get into it. You know what I mean, what could you do about it? I think I’m a little bit different in the interview.” – P8 |
| Motivated to learn about | “I know that there are different types of therapies, some of which I don’t choose to include in my own practice but I learned some ideas from people who did practice |
other therapies | those types of therapies. And that certainly helps me when I’m doing an assessment on candidates who are using them. It also made me look into some of the techniques a little more closely so I’d better understand them.” – P7

Detracted from perceived rigour of program | “I think what it brought home to me is, I thought about, really, the person being assessed, and how dependent it is on your assessor. And I guess it hit home even more for me that, really, the assessment should have a very strong collegial component to it and a very strong educational component. Because, if it’s so varied in opinions, that I could walk in say something was substandard and not what I would expect from a first year resident and someone else would walk in and say, oh that’s fine… It really made me want to always be more educational, and make sure that the person getting assessed got something out of it.” – P8

“What was really interesting was to see that physicians didn’t have a clear interpretation of what a 1 is or a 2… I thought that was the most interesting part and a little disturbing actually when I walked out… I walked away with being a little fearful of what was to come as opposed to being really bold and walking out of that group feeling empowered.” – P12

12.2 Perceived Effects of the Tools

Using the new assessment tools appeared to influence participants’ sampling of patient records, recording of performance data, and delivery of written and verbal feedback. Participants stated that the orientation of the new tools around global assessment domains caused them to select and review patient records more comprehensively, looking at more of each record (e.g., multiple visits rather than one visit). This, in turn, reportedly encouraged these participants to look at patterns of care rather than single patient encounters and to think more globally about performance. One noted that doing so likely will allow for a more representative assessment of a physician’s overall performance.

With the changed tools, many participants stated that they are now organizing information differently, by assessment domain, and are also more detailed and specific in their narrative feedback since they now need to report on eight assessment domains rather than provide one overall comment. They noted that the content is similar but the way they record it has changed.
They cited the sample reports as helpful for guiding report writing as they demonstrate how to organize information and how much narrative detail to include.

Participants stated that while they have changed their sampling of performance data, organization of information, and level of detail of written feedback, their scoring has remained mostly the same. They noted that they referred to the scoring rubrics to help them choose or confirm scores, but that they do not perceive differences in their scoring overall. Participants did note feeling more confident in their scoring and feedback now that they have consensus-derived materials to refer to, specifically the scoring rubrics and quality improvement resources. They stated that they feel better equipped and more comfortable providing feedback since they can refer to something tangible that has perceived credibility.

It appeared that assessors used the scoring rubrics but also drew upon their professional judgement when rating performance, finding their own way of integrating the two. Some stated that they made their judgements and then checked them against the rubrics, while others reportedly went “back and forth” between one’s notes and the rubrics to make final ratings. Participants noted that clinical judgement is needed to distinguish between different scores on the scale and that a physician’s context needs to be taken into account when doing so. One participant stated that the scoring rubrics provide a useful reference of the consensus that was achieved among assessors, but that the tool is not “prescriptive”; assessor judgement is still required.

Although many participants stated that the new assessment tools facilitated their conduct of assessments, some found difficulty choosing which domain to record information under and sometimes felt that they had faulted a physician twice for one issue. Some found the three-point scale to be limiting; since there is a wide range of performance that could fall under each score, some had difficulty choosing a rating. Some participants also noted that it took them longer to conduct the assessment and write the report in the new format. The data to support these findings are displayed in Table 3.

Table 3. Supporting data for ‘Perceived Effects of the Tools’

<table>
<thead>
<tr>
<th>Sampling of performance</th>
<th>“It makes you have to look at a bigger sample of charts on one patient… now, you</th>
</tr>
</thead>
<tbody>
<tr>
<td>More</td>
<td></td>
</tr>
</tbody>
</table>

Although many participants stated that the new assessment tools facilitated their conduct of assessments, some found difficulty choosing which domain to record information under and sometimes felt that they had faulted a physician twice for one issue. Some found the three-point scale to be limiting; since there is a wide range of performance that could fall under each score, some had difficulty choosing a rating. Some participants also noted that it took them longer to conduct the assessment and write the report in the new format. The data to support these findings are displayed in Table 3.

Table 3. Supporting data for ‘Perceived Effects of the Tools’

<table>
<thead>
<tr>
<th>Sampling of performance</th>
<th>“It makes you have to look at a bigger sample of charts on one patient… now, you</th>
</tr>
</thead>
<tbody>
<tr>
<td>More</td>
<td></td>
</tr>
</tbody>
</table>
comprehensive and structured sampling
really have to, in order to look at the follow-up the monitoring, the documentation of continuity of care, you have to really look at more than one visit, you have to look at a greater sample, so it’s probably a better representation of how the physician practices.” – P9

“I settled on going through, actually in reviewing the chart in a different way. I would go through the chart first… I would skim the history and then I started going through the list of things that were expected from the rubrics and checking them off one by one, which gave less flexibility but that may be a good thing because that’s likely to help eliminate assessor bias.” – P7

“I think the new templates that we’re using has made me change a couple of things… It allows me to look more for patterns of care than it does specifically looking at just specific entries… And I’ve found that I’m able to review many, many encounters and still fairly concisely summarise that chart in that new framework that we’re using. It seems to be a lot easier for me to do that now.” – P2

**Recording of Performance Data**

| Organized by domain | “I don’t think I’m really writing about things that are different, but… it’s just being organized perhaps in a better way.” – P13
|                     | “It’s definitely structure. And you’re definitely taking your thoughts and putting it together in a different way for them.” – P4
|                     | “The contents are basically the same, it’s the recording of them that is different.” – P7

| Increased detail and specificity | “It was similar to the old, to the other ones, just looks a little different… It’s more specific. We’re isolating every category a little bit more… things are being teased out, we would have a little bit more feedback, more specific feedback in different areas.” – P9
|                                | “Now that comments need to be made about each of the specific content areas. So, I suppose that that’s what’s different, rather than just providing sort of a gestalt of the file itself” – P6
|                                | “I think that I’m more specific because I want the physician to walk away with the
answer to the question of, well, what do you think I need to do differently, and what do you think I’m doing right. Yes, I’m being more specific.” – P10

<table>
<thead>
<tr>
<th>Score</th>
<th>Interviewee</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scoring and Feedback</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rubrics helped choose or confirm scores</td>
<td>“I referenced it a couple of times during the assessment, just to make sure that what I thought was probably moderate, little to no improvement or, sort of moderate improvement and not a concern, actually was reflected in the template as far as their bullet points for the elements of quality evaluation criteria.” – P5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“I found them very useful in terms of thinking about, okay, what do I really want to say here, and how is this reflected in this score. Is it the minor thing, or is it a major thing that could be potentially life-threatening.” – P10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“When it came to the gray areas of, does this physician just need a little topping up of the clinical, or the documentation, or is this really a serious problem that needs to be identified as such so that, between a two and a three if you will, I think the development of the scoring rubric there with some of the clinical scenarios was helpful.” – P3</td>
<td></td>
</tr>
<tr>
<td>Sample reports helped guide report writing</td>
<td>“The sample was very, very helpful. It gives you a baseline to model and it also allows you to see what’s put in the categories, in the interpretations.” – P12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“So having the model report, when I wrote mine, I had it right beside me and I modelled it on that… Now I know how to be more succinct, yet to get information across to the QAC in a most succinct and useful fashion.” – P1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“I liked when they gave us… exemplars where there are practice assessment reports showing ones with moderate concerns and ones with significant concerns and what they’re actually looking for.” – P5</td>
<td></td>
</tr>
<tr>
<td>Numeric scoring unchanged</td>
<td>“And the new format helps break it down into domains. And I can actually make relevant comments in one domain that isn’t necessarily specifically related to just that single encounter. But rather, I can comment on how that domain has manifested itself throughout the chart.” – P2</td>
<td></td>
</tr>
</tbody>
</table>

*Interviewer: Has that affected the way that you score physicians? The actual ratings that you give and the feedback that you provide?*
| **Verbal feedback facilitated by tools** | “Probably not really, but I imagine that it... makes it so much easier for them to understand exactly why my recommendations say what they do.” – P2  

“Would the same physician be getting the same rating? Yes, yes they would, but maybe because things are being teased out we would have a little bit more feedback, more specific feedback in different areas.” – P9 |

| **Increased confidence in scoring and feedback** | “And the QIRs... they actually make it really a lot less, for lack of a better word, it’s a little bit less cringy when I come across someone who’s really missing the point on some educational areas that they need help on. It’s really helpful to be able to go to a QIR and say... let me photocopy this for you here at your office, and let’s review it together. And it allows me to sort of not put myself in the place of saying, hey, I’m your taskmaster, I’m telling you what you need to be doing right. It gives me a little bit more credibility to say, oh look, here’s a third-party information sheet that, we can look at this together and go through it... So, it’s allowed me to, hopefully, come off in a much less authoritarian sort of way... It just seems a little bit more neutral to be able to come at that, from that point.” – P2  

“You can point to this. These tools were developed by a consensus process... And these are things that a group, through a consensus process, generally feels are what should be done... it’s not just me verbalising. I can actually point to things here.” – P1  

“The feedback I give is more meaningful, I think, now. It’s more robust and I believe it’s better appreciated.” – P1 |

| **Increased confidence in scoring and feedback** | “I feel better equipped, and secondly, I feel empowered to be able to supply the assessed physician with something that is an authorised point of view. One of the anxieties that I had before was, am I advising my colleagues in ways that... would not serve them well, or that someone at the College would take exception... You don’t want to be misleading your colleagues, and now I feel more comfortable with that.” – P10  

“I feel that I am backed up more here. I was confident in my feedback before but let’s say handwriting, which is a common concern... I can point to the legibility scale here. And we have a good scale on what is legible and illegible. And I think...” – P10 |
it depersonalises the feedback I give to the physician.” – P1
“I feel much more confident now using their phrases or lining up as well the different areas of the chart in a more narrative straightforward fashion.” – P5

| Integration of rubrics and judgement | “Once you get into doing assessments, you have a really good instinct of what is a good clinical practice and what is not so that is extremely helpful. I think that previous experience is helpful to be able to see the diversity of clinicians and really what the expectation is… The rating score [rubric] is definitely helpful, but also the experience of doing a lot of assessments is helpful. I don’t think you can isolate the two… I think experience is helpful to be able to interpret a rating score as well.” – P12
“I mean you’re looking at a trend, but if you have a clinician that does document let’s say, for example, physicals, but they may not document to the T of the rating tool. It’s a clinical judgement whether they’re a 1 or a 2.” – P12
“In a way it’s a very comprehensive assessment expectation… but people, when they do the assessment, have to really put into context the place where someone is working.” – P13
“It’s not really completely prescriptive but it gives a good framework on which to fall back and sometimes the actual scoring is debatable and then going back to looking at the way we had structured the scoring rubrics… is a very good reference point to remind us how we achieved that consensus.” – P3
“You go back and forth and sort of make a stab at it with your initial impression and then re-look at it [the scoring rubrics] afterwards when you’re doing the final review of the document. And I think I did switch back and forth a number of times… because it’s still subjective even though you’ve got your criteria there.” – P4

<table>
<thead>
<tr>
<th>Challenges with Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domains</td>
</tr>
</tbody>
</table>
| “I think the report rating is more challenging because things can be put in more than one area, and as a physician, you have to decide.” – P9
“I had trouble deciding where to put it and it turns out that if I put it in both that’s the same as double-dipping with this scoring rubric… The person would be
penalized twice where I didn’t.’” – P7

“One of the things I didn’t appreciate was that… I should only mention that sort of thing once in one place, so that the physician doesn’t get penalised twice.” – P10

| Three-point scale | “For myself, the more categories you have across a board from excellent to really terrible, is more sort of useful than making it smaller and clumping it more… I think it sort of takes away from the subtlety of it a bit.” – P4

“It’s not a very easy grading system really. It would be, I think, a lot easier if you would have a five-point grading system… in a way, because the groupings are so small… a lot of people end up being in the 2 zone. And that’s a huge spectrum of people. And the quality difference of people you would have to grade to is quite large.” – P8

| Increased time | “The number of charts that I was able to assess during the actual interview was probably about half, 50% of what I normally would be able to get through in a usual amount of time which is a couple of hours.” – P6

### 12.3 Differences in Perceived Effects based on Type of Training

The perceived effects of training and the new assessment tools appeared to differ based on whether or not participants were involved in developing the tools. Those involved in tool development (PDT phase 1) spoke very positively about training and tool development, stating that they found the processes beneficial and impactful. Some noted the benefits of learning from each other through consensus-building discussions and two mentioned feeling less isolated now, having gone through the process. Many reported feeling more comfortable with the tools because of their role in developing them, with one noting that the tools became “ingrained”. Some found that they gradually changed their approach to scoring throughout the development process. Some also stated that their expectations of physicians had changed as a result of the process, either becoming more stringent, more lenient, or being better able to adjust their expectations to the physician’s context; those not involved in PDT did not report a change to their scoring expectations.

Almost all participants not involved in tool development (those in PDT phase 2) stated that the new program seems very similar to the previous program and indicated that their conduct of
assessments will be mostly the same. One noted that the training session had a minimal impact on building consensus among assessors and another found the training unmemorable. Whereas most participants involved in PDT phase 1 felt comfortable with using the tools in their first assessment, some involved in PDT phase 2 mentioned that they feel like they are still transitioning to the new format and are not yet familiar with it.

Overall, it seems that the perceived effects of the new program were more pronounced for participants involved in tool development. These assessors felt more positively about the new program and perceived a greater effect of training and the tools on their conduct of assessments. The data to support these findings are displayed in Table 4.

Table 4. Supporting data for ‘Differences in Perceived Effects’

<table>
<thead>
<tr>
<th>Perceived Effects for Assessors Involved in PDT (Tool Development)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Perceived benefit of new program</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Learning from other assessors throughout process</strong></td>
</tr>
</tbody>
</table>
| Increased comfort with tool | “I think it was ingrained in us over the period of time that we developed the process as to how the process was going to unfold… It may not be the same for those who would have had some input into consensus building but weren’t involved in the development so much. So, I think from my point of view, I feel relatively comfortable with it.” – P3  

*Interviewer: How have you been able to integrate those two things? Using your own judgement, but also relying on these external consensus-derived external materials.*  

“A lot easier for me because I helped develop them.” – P2 |
| Change in conduct of assessments over time | “I’ve only done this once with the new pilot, but in the process, while we were doing this, I think my feedback was changing, and I was getting a better grasp. Because we were developing the document… we got more confident about where we had gone with this. Some of my colleagues have changed my mind about certain things that I’ve always thought about, and certain ways that I had about thinking… So I’d be doing peer assessment even before we did the pilot, and I was already expressing different views.” – P10  

“I think the whole process had a bit of an impact. Even before, for example with the charting [patient record selection], I started to ask people to identify some charts rather than doing it randomly. So in that sense, some of the things influenced my assessments even before the new tool came out.” – P13 |
<table>
<thead>
<tr>
<th>Perceived Effects for Assessors not Involved in PDT (Tool Development)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Minimal perceived difference in program</strong></td>
</tr>
<tr>
<td>“I didn’t find that it was… it wasn’t much different than the previous one. I just felt that it was perhaps a little better organized and easier to follow than the previous tool.” – P6</td>
</tr>
<tr>
<td>“I have to say, after I did it the first time, my biggest … it seems very, very similar to what it was before. Except for some maybe, just minor kind of details changed.” – P8</td>
</tr>
<tr>
<td>“It wasn’t that different than how we were doing them before.” – P9</td>
</tr>
<tr>
<td>“I don’t think it will change to be honest with you… the only thing that’s change is how I need to draft the report, but I don’t think it will change anything for me.” – P12</td>
</tr>
<tr>
<td><strong>Minimal impact of training</strong></td>
</tr>
<tr>
<td>“So I think, actually, if you really wanted consistency, you’d have to put a lot more effort into training the assessors to setting standards across the board. Because what they did was just kind of dipping their toe in the water, in terms of trying to get us to have a consensus approach to things.” – P8</td>
</tr>
<tr>
<td><strong>Transitioning to new program</strong></td>
</tr>
<tr>
<td>“The first time I did the assessment in the new format, it was really just getting used to the way they want things done now. Because, if you’re aware of it, that there is a much more stringent instructions about how to select charts and what they want to see, specifically.” – P8</td>
</tr>
<tr>
<td>“In this new way, it’s more global, and so, especially with my first one, I didn’t feel I had a good grasp of how to explain in the new way. But I’ve done a couple now in this new format, and it was good.” – P11</td>
</tr>
<tr>
<td>“I think if one had done several of them, then one would get more and more at ease with doing that and picking the right scores perhaps more quickly or more intuitively.” – P4</td>
</tr>
</tbody>
</table>

### 12.4 Perceptions of Improved Standardization

Many participants stated that a key benefit of the redeveloped peer assessment program is increased standardization, which they report will make assessments more consistent, credible,
objective and fair. Participants referred mostly to standardization of processes, such as sampling (e.g., patient record selection), interviewing, and report writing, but also of assessor judgement, since consistent standards will now be applied. Despite this perceived increase in standardization, participants also tended to note that their own scoring had remained mostly unchanged. The data to support these findings are displayed in Table 5.

Table 5. Supporting data for ‘Perceptions of Improvement Standardization’

| Standardization in process | “With the College’s direction we began to develop new ways to standardize the process because now it was felt to be one of the concerns that the assessors needed to have a more standardized approach which would make it more streamlined across the whole domain of the peer assessments.” – P3

“Going forward, there’s going to be less variability in these reports, and I think that leads to a much more credible process.” – P1

“I think it’s going to be more structured and I think there’s going to be more consistency between all of us in terms of how we do it… in terms of how we select charts, in terms of some of the questions that we ask before, in terms of how we do our rating, it should all be much more consistent, I’m hoping.” – P13

| Standardization in assessor judgement | “We all come from different backgrounds. We all have different ideas of what should be happening and how it should happen so I think it doesn’t leave a lot of room for biases. I think it’s clear. Actually I think that that’s the most beneficial thing about this whole reassessment, about the process and the outcome… consistency among assessors.” – P7

“I was very happy to see that actually the peer redesign was happening. I liked it, in that I think that there is just more guidelines and it’s much clearer what the College is looking for and considers acceptable. It’s much more clear what is a recommendation as opposed to what is a concern whereas in the past, I felt that sometimes has been a little subjective, which was challenging when you’re doing the assessment. I mean, there is obvious clear concerns but then it hasn’t been obvious sometimes, to me anyway.” – P5

“I think the fleshing out of all the elements of the SOAP format or the assessment, the history, physical and so forth was important for us and it helped me to more
| Own scoring unchanged | “I’d like to think that I’m reasonably objective with my assessments and I think this certainly helped to standardize it but I’m not sure that I was scoring people differently.” – P3  
“I think there will be more consistency, not individual assessments… but across the board among the assessors. Because we do have, we each have our own ideas about what’s appropriate and what isn’t. So this is giving us guidelines that say well whatever your ideas are you can use them in your own practice but these are the requirements of the College, so I think that that’s important.” – P7  
*Interviewer: Did that mean that you had to change your approach to some things?*  
“No, it didn’t really but I’m an educator. I was a teacher. I was in education before I came into medicine so I’m used to evaluating and having a set of rules to go by too, in my own mind, when I’m evaluating it.” – P7 |

### 12.5 Perceptions of Assessor Role

Participants noted that their role within peer assessments is to provide educational feedback to physicians and to record information about performance for the Quality Assurance Committee (QAC) at the CPSO. This dual role was salient, with many participants indicating the dual formative and summative aims of the program. Most stated that they give verbal feedback to the physician on the day of assessment for educational purposes, and then provide a written report to the QAC for decision-making; the QAC is viewed as the primary audience for the reports. Most participants spoke positively about their role in providing educational feedback to physicians; however, one noted discomfort with this aspect of the assessor role, stating that he does not feel he is in a position to educate other physicians since most are highly qualified and competent.

Participants noted that their role within report writing is to collect and record evidence about physician performance, whereas the role of the QAC is to make decisions about that performance. Participants noted that they communicate this to assessed physicians, seemingly to distinguish their role as one of recording facts as opposed to determining outcomes. Two participants mentioned that they now feel they are providing more judgement than they used to, with the increased focus on narrative comments in the new program. One participant was
comfortable with this perceived role change whereas the other was displeased that assessors seem to be expected to make decisions now. The data to support these findings are displayed in Table 6.

Table 6. Supporting data for ‘Perceptions of Assessor Role’

| Dual role of assessors given the dual formative and summative aims of the program | “And the goals here were two-fold, overarching quality assurance, we as physicians have to do a reasonable job clinically on a medical record keeping, but also there was the education component. So, this peer redesign very importantly bolstered an area that was underappreciated and not emphasised so much on by the College nor assessors in the past, that was education.” – P1  
“I’m not a harsh taskmaster, but I also have really high standards for what I expect to standard of care… But I also am not a very imposing person. And so, I want to be gentle and affirming and positive in my feedback to a physician. In the same breath, I want to make sure that they really get the message about what’s been great and what hasn’t been great. And so, I want to make sure that the language I use is clear enough for the Quality Assurance Committee to understand that, yes, this has been a concern. But I want that language to not sound punitive when it’s read by the physician who gets a copy of my report. So, I don’t want it to come off sounding punitive, but I need the QAC to completely understand that, yes, this is an area of concern.” – P2 |
| Formative role – providing feedback for physicians | “I try to conduct them in a collegial manner, I make myself available as a resource before, during, and after. A significant minority, I would say about 40% of the people whom I’ve assessed, keep in touch with me, and run things by me.” – P10  
“I point out the good points first and the things you do to give good criticism. And then to reinforce the fact that this not meant to penalize, this is meant to educate, which of course you’ve talked to them about at the beginning anyway” – P7  
“You don’t want the physician to feel like they’ve been attacked. You want them to actually move forward from it, knowing that this can be a positive thing, that it can be a good outcome with positive change and stuff like that, even though there are some significant areas that might need improvement.” – P2 |
“Essentially, you have to choose the feedback that you’re going to give to the physician in a way that’s actually going to be collegial, non-threatening and educational. And then that’s exactly how I try to write it on the form.” – P8

“I don’t find the educational component to be all that important when it comes to the Peer Redesign, unless of course, flagrant deficiencies are being identified. I mean, [physicians] who are being assessed are working in a specialized area and should have full competencies in their clinical areas. So, I find the educational value fairly minimal myself… I suppose there’s room for education if somebody is deficient in certain clinical practice areas but I haven’t really come across that. I mean, it’s something I’ve always struggled with, what is the real educational value? Well, sometimes I’m assessing individuals who have far more expertise in the area of practice that I do.” – P6

<table>
<thead>
<tr>
<th>Summative role – writing reports for QAC without making decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>“I always tell the physician during my feedback session that whatever I tell them will be in the report. I'm not putting anything in the report that we don’t discuss today. So, I feel that I've already gone through all the areas with the physician, so when I write the report, it’s like I'm writing to the College about this doc.” – P9</td>
</tr>
<tr>
<td>“I do indicate that my primary … I write this for the quality assurance committee… But you’re going to get a copy of it, so there should be no surprises.” – P1</td>
</tr>
<tr>
<td>“One of the things that I’m pretty clear on in my own mind, and I tell the physicians, is that I’m writing a fact-based report and I can certainly give my impressions and opinions, but I don’t make any decisions. The decisions about what to do are made by the committee.” – P8</td>
</tr>
<tr>
<td>“We don’t say, this person needs a reassessment, this person needs to take a medical record keeping course. We make suggestions and recommendations… QAC wants the evidence. We lead the horse to water. It’s the QAC that ultimately makes the decisions. So.. in my report, I try to be as clear as I can without overstepping my boundaries as an assessor and to present the information in a way that I hope will lead to the best-informed decision by the QAC and what to do.” – P1</td>
</tr>
</tbody>
</table>
| “I make it clear that I’m not the one who decides what happens as a result of this
report… That it will be reviewed by the committee and that if they detect any issues that I’ve missed that they will be in contact with them” – P7

| Perception of assessors providing more judgement | “We’re still trying to give the committee the information they need. Perhaps now with the new form, more is wanted, more judgement. It always used to be said that the assessor is the eyes of the committee, and now there’s a bit more not just eyes but also judgement. And, yes, we always made judgements, but sort of the feeling I have is that the new form wants you to be more of a judge, which is fine. It’s not a problem. But that’s the feeling I have.” – P4

“I think, sometimes, that they would like me to make their decisions for them. And that came across a little bit, in the feedback I got after doing the new assessment. And, as an assessor, I wasn’t impressed by that.” – P8 |

13 Analytical Summary and Synthesis

Cumulatively, these findings suggest that FORT revealed the variability of assessor judgements which led assessors to realize that their judgements do not represent the “truth” of performance but rather one’s own subjective perspective. This, in turn, seemed to facilitate an understanding that there are not “true” scores of performance or universal standards that all physicians can be held to. Whereas participants’ comments suggest they may have previously assumed that their perspectives about performance reflected an objective reality, they now seem to understand that there is not one correct assessment of a physician’s performance, only subjective interpretations of that performance within a particular context.

Most participants responded positively to this realization of variability and inherent subjectivity, noting it encouraged more reflective scoring and motivated them to use the structured assessment tools to bring consistency to assessments. However, for a small minority, it detracted from their confidence in the program and prompted unintended changes to their behaviours. For example, one participant stated that she now focuses more on the educational aspect of the program, seemingly to lower the stakes of the program. While enhancing the educational impact for physicians is an intended outcome of the redeveloped program, this was not the intended mechanism.
The findings also suggest that the use of the new assessment tools had perceived effects on sampling and feedback but minimal perceived impact on numeric scoring. Participants stated that they now review patient records more comprehensively, comment more globally on physician performance, and provide more detailed and specific written feedback to physicians, but that they do not perceive a change in their rating of performance. Although they stated that they refer to the scoring rubrics to make ratings, they do not find their ratings to be different now than what they would have been previously. They did note feeling more confident in their scoring and feedback because of the tools.

Our analyses indicated a difference in perceived effects between participants who were involved in tool development (PDT phase 1) and those who were not (PDT phase 2). Participants involved in the iterative, consensus-building process of developing the tools mentioned internalizing the tools over time and noticing incremental changes in their scoring behaviours throughout the process, with some stating that their expectations of physicians are different now. Conversely, participants who were not involved in this process found the new program to be very similar to the previous program, suggesting that the tools and training had minimal effects on their approach to assessments or their frames of reference.

Despite these potential differences in effect between training groups, almost all participants spoke about how the new program is more standardized and therefore more consistent, objective, and fair. However, most also stated that their own scoring has remained mostly unchanged. This perception that the overall program will be different but that one’s own conduct of assessments will be the same is paradoxical. Our analysis provided several explanations for this, outlined below.

Firstly, it could be that assessors assume that while their own scoring is unchanged, other assessors are going to alter their behaviours. For example, one participant stated that she has found that other assessors do not write as detailed reports as she does, so sees the value in standardizing this aspect of assessments: “So, I think for me it hasn’t changed, but I have seen other reports so I’m going to say that I could see where this definitely fills a role” (P12). Assessors may therefore perceive their own way of conducting assessments to be “correct” and that standardization would entail other assessors adjusting their scoring behaviours accordingly.
Secondly, assessors may be unable to perceive differences in their own cognitive processes. Even if their scoring has changed over time, they may not be able to detect such changes, particularly if they considered themselves to have always conducted high quality assessments.

Thirdly, assessors may be unable to perceive how changes in their assessment behaviours may affect their ratings. For example, although participants noted differences in their sampling and recording of information, they may be unable to think about how their scoring would differ had they not sampled and recorded performance in this way. This might be equivalent to asking someone how they would have performed on a test if the questions were different.

Fourthly, it could be that while assessors are using the scoring rubrics, they have not internalized them so do not perceive a change in their cognitions. In other words, they may be referring to an external frame of reference during assessments but their own frame of reference is unchanged. Conversely, it could be that assessors, particularly those involved in developing the tools, altered their way of scoring slowly over time. In this sense, their frames of reference shifted gradually so they did not perceive a change in scoring:

“I’m not convinced that I’ve changed my own method of doing things drastically based on this but perhaps that was also because as I was part of the small group that started to develop this, perhaps I was already assimilating some of this into my assessments.” – P3

Lastly, it could be that participants are referring to a subtler form of standardization wherein there is consistency in process but flexibility in assessor judgement. Some participants spoke about how they integrated the scoring rubrics with their own professional judgement, noting that experience and clinical expertise are needed to apply a scoring rubric while accounting for context. This perspective aligns with the notion that assessor subjectivity and judgement is inherent to performance assessments. However, participants seemed to imply that assessor subjectivity is problematic and were pleased with increased standardization and perceived objectivity of the new program. Some noted that whereas the former program had more potential for inconsistency, they can feel better about the program now that it will be more objective and fair.

With this value placed on standardization and objectivity, participants spoke positively about their role as objective recorders of information, rather than interpreters or decision makers. In
fact, one assessor voiced concern that the assessor role seems to be verging on decision making now that there is an increased focus on providing narrative comments. It may be that some assessors feel more confident about their role if they perceive it be one of objectively recording information. Alternatively, it may be that by not being decision-makers, assessors feel better able to foster a formative environment for assessed physicians. Participants noted that they tend to tell physicians that they are peers providing feedback rather than CPSO employees making decisions. They may do this to set a collegial tone, highlighting the educational lens many reportedly adopt when conducting assessments. Participants also noted, however, that they are focused on providing the QAC with high quality information for their decision making, suggesting that they also emphasize the summative aspect of the program. This balance that assessors must find between the formative and summative aspects of the program and between supporting decision-making without making decisions was salient.

14 Theoretical Contribution

Our findings suggest that assessor training may make visible for participants the subjectivity of their judgements, demonstrating that there is not one objective “truth” of a physician’s performance but rather multiple subjective interpretations of that performance. While participants reported that this learning prompted behaviours that support scoring (such as more representative sampling of performance, adherence to the scoring rubrics, and provision of more detailed and robust feedback to physicians), they did not perceive a substantial change in their numeric ratings. At the same time, however, being involved in tool development and training seemed to lead participants to sense an increased standardization of the assessment program. This finding highlights a “standardization paradox” wherein participants perceive a general increase in the assessment program’s standardization but no change in their own numeric ratings. This paradox could imply that participants perceived the onus of standardization to be on other assessors, that our participants could not detect or articulate changes in their implicit scoring behaviours, or that participants acknowledged the inevitable balance of standardization and subjectivity in WBAs. The results also highlight the potential importance of assessor involvement in tool development and suggest that this may be the most influential aspect of PDT and FORT.
Chapter 4
Discussion

We aimed to explore the effects of Performance Dimension Training (PDT) and Frame-of-Reference Training (FORT) on assessors’ scoring behaviours within WBAs delivered by the College of Physicians and Surgeons of Ontario (CPSO). Using a constructivist grounded theory approach, we explored assessors’ perceptions of what they learned through training, how they utilized consensus-derived tools, and how the training and tools influenced their conduct of assessments.

We found that both PDT and FORT were reportedly valuable learning experiences for participants that provided opportunities to learn from each other and highlighted the variability of assessor judgement. This learning appeared to motivate participants to use the scoring rubrics and prompted them to be more reflective about scoring and feedback. The use of the consensus-derived tools reportedly affected participants’ sampling and recording of performance data and increased their confidence in scoring and feedback. While participants utilized the scoring rubrics to choose or confirm their scores, and demonstrated an integration of the tools with their own professional judgement, they did not perceive substantial changes to their numeric ratings.

Participants who were involved in PDT phase 1 (tool development) appeared to respond more positively to the training and tools, citing more substantial benefits of the new program, whereas those who were not tended to find the new program similar to the previous one. However, almost all participants noted that the new program will result in greater standardization and consistency among assessors. Participants also commented on their role within the assessment process, highlighting their balance of the formative and summative aims of the program and the boundary of their role of supporting decision making but not being decision makers. A discussion of these findings as they relate to the literature is presented below.

15 Discussion of Findings

15.1 Perceived Effects of Training

Through PDT (phase 1 – tool development), participants reported learning from each other about different assessment approaches and becoming comfortable with the new standards reflected in
the tools. Some stated that they began to incorporate elements of the new tools and assessment process during the development process and changed their approach to scoring over time. Through FORT, participants reportedly learned about the variability of assessor judgement and where their perspectives fell within the range of assessor perspectives. This learning had largely positive effects on participants, motivating them to use the scoring tools, to be more reflective during scoring, and to consider how expectations may differ in different practice environments.

Almost all participants reported learning about assessor variability through FORT but only those who were involved in tool development (PDT phase 1) stated that they adjusted their expectations of physicians as a result of the process. It is possible that prolonged exposure to the tools throughout the tool development process may have increased their comfort and facilitated assimilation of their perspectives to the assessment criteria. This potential exposure effect would suggest that assessors not involved in tool development may gain comfort over time with repeated use of the new tools. Alternatively, it could be that the tool development process itself affected assessors; directly contributing to the tools through discussion and debate may have educational value and may enhance assessors’ investment in the program. Kogan et al. (29) found that those involved in tool development responded more positively about training and found it more educationally beneficial. Comparatively, those who were given pre-defined tools did not find training as impactful; they questioned the tools’ value and found the assessment framework “forced”. By not participating in the development process, it is possible that assessors do not perceive the tools to reflect their way of thinking or that they do not feel as invested in them.

These findings may shed light on why other studies had conflicting findings about the effects of training on assessor scoring. In Holmboe et al.’s (25) study, wherein reliability improved after training, PDT consisted of participants defining the elements of competence and developing the criteria for high quality performance. Conversely, in Cook et al.’s (28) study, wherein there were not improvements in rating accuracy or reliability, PDT involved only a discussion of the performance domains that comprised a standardized scoring tool, not engagement in defining the standards or developing the tools. Additionally, relatively less time was spent on PDT in this study; the majority of training was dedicated to FORT. It may be that being directly involved in developing the tools that define a new standard or frame of reference is integral to assessors applying that standard in subsequent assessments. Kogan et al. (29) noted that most participants
valued the process of creating the tools and found this to be more beneficial than the resultant tools themselves. The act of developing the shared mental model was reportedly a significant learning experience for assessors and not only the means to an end.

Actively contributing to the development of scoring tools may be a critical component of assessor training. However, despite the potential benefits of PDT phase 1 (tool development) in this study, FORT was still reportedly a valuable learning experience for all participants. FORT alone may not have changed participants’ frames of reference, but it highlighted that other valid frames of reference exist. This, in turn, appeared to motivate participants to utilize scoring tools and to think about how other assessors might rate the same performance when scoring. It is possible that this may ultimately have positive effects on scoring, even if internal frames of reference were not altered; by realizing that one’s perspective does not represent the “truth” of performance and being prompted to use the scoring rubrics accordingly, assessors may begin to score more consistently due to their adherence to the tools.

Overall, results suggest that both PDT and FORT contributed to assessor learning but in different ways. The iterative, consensus-building process of PDT phase 1 seemed to foster assessors’ comfort with the tools and helped to calibrate their judgements and expectations. FORT appeared to open assessors’ eyes to the variability of their perspectives and the importance of putting effort towards consistently using the scoring tools. While we do not know the effects of this learning on assessor reliability, we have an understanding of what assessors learned through training and how this influenced their approach to scoring.

These findings contribute to the scoring inference in our validity argument. Assessors’ perceptions of their learning, their motivation to use scoring tools, and their change in way of thinking about scoring (i.e., thinking about how others would rate performance, thinking about how physicians might receive feedback) suggests scoring will be more aligned with the agreed upon standards outlined in the tools. It also suggests that feedback may be constructed with the recipient in mind, providing some support for the implications inference. Further evidence will need to be collected to support these claims, but assessors’ perceptions of their scoring behaviours provide a valuable first step in our validity argument.
15.2 Perceived Effects of the Tools

The use of the new tools appeared to influence scoring behaviours. Specifically, participants stated that they are now sampling performance more systematically and comprehensively and are organizing their observations in a way that is more representative of global performance or patterns of care, rather than single patient encounters. This type of comprehensive and representative sampling is a key contributor to the reliability and generalization of an assessment program (14,36,37). Generalization in Kane’s validity framework refers to the extent to which the sample of performance reviewed during an assessment is reflective of the entirety of possible assessment performance; in other words, how well the patient records (and the specific aspects of the records) reviewed represent all of a physician’s records. Given that assessors report they are reviewing more of each patient record, assessing patterns of care over multiple patient interactions, and choosing a more structured cross-section of patient presentations, we can argue that the performance reviewed more adequately reflects the entirety of a physician’s performance (that is demonstrable through patient records). The fact that assessors are all using the same approach to record selection, in that they are choosing the same types of patient presentations and looking across multiple visits, also supports generalization by contributing to the consistency of sampling across assessors.

Participants stated that they are using the scoring rubrics to choose or confirm their scores, they are providing more detailed and specific narrative comments, and they are more confident in their scoring and feedback because of the tools. Some stated that they are providing more robust verbal feedback because of the quality improvement resources and that their feedback seems more credible now that it reflects a consensus-based process. These findings support the scoring and implications inferences in Kane’s validity framework. Assessors’ use of the scoring rubrics supports accuracy and consistency in scoring (45). The increased specificity of assessors’ written and verbal feedback supports the intended formative purpose of the program, as feedback is most useful when it is specific (47,61). Increased specificity also supports the summative purpose, as rich narrative is necessary for trustworthy decision making (9). Assessors’ reference to the scoring tools during the feedback process further supports the intended formative purpose of the program; transparency in evaluation criteria can help learners know how to improve their performance (62) and perceived credibility of feedback is an important factor in its uptake and utility (63).
While participants noted changes in their sampling, scoring, and feedback processes, many did not perceive any direct influence on their numeric ratings. This finding could imply that changes in these processes do not ultimately affect assessors’ scores. Alternatively, it could reflect that assessors have difficulty identifying and articulating their own cognitive processes as they relate to scoring (64). Assessors tend to make judgements implicitly, through unconscious processes (65). Therefore, while differences in behaviours, such as sampling, use of tools, and delivery of feedback may be apparent to assessors, differences in their thought process, such as synthesis of information and arrival at a numeric score, may be much more difficult to observe and articulate. Noticing changes in one’s scoring may have been particularly challenging for assessors involved in tool development who indicated that they may have assimilated to a new way of rating over time.

15.3 Perceptions of Standardization

Despite not noticing differences in their own scoring, many participants stated that the redeveloped program would lead to more consistency in assessments. This standardization was perceived by many as a key benefit of the new program. The discordance between perceiving an increase in the standardization of the assessment program but little or no difference in one’s own numeric scoring highlights a “standardization paradox” which could be explained in a few ways. Since participants did not perceive themselves as having changed their scoring, this might imply that they think their own scoring is “accurate” and that standardization would entail other assessors adjusting to their ratings. This would align with the self-assessment literature that demonstrates physicians are not good judges of their own abilities (66). In general, people tend to rate themselves as “above average” in most areas, including their ability to make self-assessments (65,67). Participants in this study therefore may have overestimated the accuracy of their scoring and thus perceive the onus of standardization to be on other assessors.

Alternatively, participants may be referring to a more nuanced form of standardization wherein there is standardization in process (e.g., sampling, reporting) and in scoring framework (e.g., assessment domains and associated rubrics), but allowance for flexibility in assessor judgement. Some participants spoke implicitly of balancing standardization with flexibility, using the scoring rubrics only in “gray areas” or making judgements then confirming them with the
rubrics. They also noted that the rubrics provide a framework but judgement is still needed to apply the criteria in context.

This balance of standardization and subjectivity reflects a broader balance in medical education of needing to ensure program validity without compromising the authenticity of performance assessments (68). Program administrators must ensure that assessments are defensible, in that they are accurately measuring what they are purporting to, but assessors must also utilize professional judgement to evaluate performance in context. Tension exists between standardizing assessments while honouring assessor subjectivity, “calibrating” assessor judgement without trivializing the assessment or detracting from authenticity (5,69). Administrators must navigate this tension when developing assessor training programs and consider the amount of variability that is acceptable within their WBA.

There are different perspectives on the nature of assessor variability as it relates to an assessment program’s validity. Some view it as a source of unreliability, detracting from a program’s validity whereas others view it as acceptable and often meaningful (42). Also, studies on the impact of assessor variability on program decisions and outcomes have conflicting findings. Some have shown that assessor inconsistency accounts for more variance in assessments than actual differences in trainees (70,71), which can lead to differing decisions about which students should pass or fail (72). Conversely, others have demonstrated that although assessors may employ different cognitive processes and frames of reference when evaluating performance, they lead to similar global ratings, thus proposing that variation in scores is mostly related to variation in performance (38).

A program’s desired degree of assessor consistency is often related to the purpose and stakes of an assessment program. The view that variability is error is common in summative, high stakes assessments where the reliability of decisions is of highest priority (12). The perspective that variability represents meaningful idiosyncrasy is more common for lower stakes, formative assessments, or assessments that incorporate scores from multiple assessors where variability may offer unique and important insights into different facets of performance (42). There is not always a clear delineation between formative and summative assessments, however, and performance assessments often include aspects of both (5,73). Program administrators need to determine where their program falls along the formative-summative continuum (74) and
therefore how much assessor variability will be acceptable given the nature and purpose of their program. They also need to communicate this to assessors for a common understanding of what should be standardized and where variation is appropriate.

15.4 Perceptions of Assessor Role

Participants noted the balance of the formative and summative aims of the Peer Assessment program when describing their role; that they need to both provide feedback to the physician and give clear information to QAC for decision making. The regulatory context of the Peer Assessment program necessitates a summative component, but the goal of enhancing physician performance through feedback requires a formative one. Some balance of these is necessary, but more thought may need to be given to how to better support assessors in attending to both and not compromising one over the other.

There is an inherent tension between formative and summative purposes, however, many argue that performance assessments can and should be both (5,73,75). A summative assessment without a formative function is a missed opportunity for learning (73) and may be perceived as trivial by the learner (5). One could also argue that a formative assessment without any summative elements may be perceived as of low importance by the learner.

However, while formative and summative purposes cannot be artificially dichotomized, they also cannot be haphazardly combined. Consideration must be given to how to attend to both, as a perceived summative context can hinder an assessor’s ability to provide formative feedback (76) and may interfere with a physician’s receptivity to that feedback (75,77). Of greatest importance is not necessarily the program’s intended purpose, but how formative or summative the assessed physician perceives the assessment to be (73). Therefore, WBAs with an intended formative focus need to be delivered in a way that supports physicians’ perceptions of its educational nature. To this end, assessors should be coached in how to foster a formative environment as this will help to facilitate learning and receptivity to feedback. They should also be trained in how to effectively attend to the summative aims of a program without compromising these potential formative effects. This is a fine balance and further discussion is needed about how to support assessors in this endeavour.
Assessors should also be supported in navigating the role of supporting decision making without making final decisions. By teaching assessors that they are not decision makers, administrators may reinforce a misconception that assessors should be objective fact collectors. This, in turn, may perpetuate the fallacy that subjectivity is problematic and must be avoided. Assessor training programs should aim to foster a common understanding of program purpose and stakeholder roles in order to guide assessors in their conduct of assessments.

16 Implications and Future Research

This study contributes to the literature on assessor training and cognition by highlighting the potential value of PDT and FORT. While other studies have examined assessors’ reactions to PDT and FORT (29) and the effects of these training methods on scoring accuracy and reliability (25,28), ours is the first to explore assessors’ perceptions of what was learned through training and how that learning and the use of the newly developed tools affected their conduct of assessments. Our results suggest that these training methods can demonstrate to assessors that other valid frames of reference exist outside of their own, motivating them to use scoring tools and to be more reflective about scoring. Although we cannot determine if and how these factors may affect assessors’ rating accuracy or reliability, the learning and potential behaviour change demonstrated in this study suggest that PDT and FORT are worthy of further study.

Given that both PDT and FORT were reportedly valuable educational opportunities for assessors, future training programs would likely benefit from including elements of both types of training: engagement in iterative, facilitated discussions about how performance should be evaluated (and reflected in the tools) as well as practice and feedback in using the tools. Our results suggest that a main benefit of this training may be demonstrating to assessors the potential variability of their judgements and motivating them to use structured scoring tools. Future studies should replicate these findings in other settings and follow assessors longitudinally to determine whether their adherence to the tools is sustained over time.

Our results also suggest that being involved in tool development may be most effective for influencing assessors’ conduct of assessments; however, it is difficult to determine whether this is due to the nature of tool development or simply the increased amount of time that assessors were involved in PDT. Future research should explore differing lengths of time of FORT to examine if such effects can be accomplished with increased iterations of practice and feedback,
rather than involving assessors in tool development. Additionally, future research could investigate if there is something unique about being involved in tool development that influences rating behaviour, independent of the amount of time or number of iterations assessors participate in. While it is often infeasible to include all assessors in developing tools, especially if new assessors are recruited after tools have already been developed, there may be critical elements of the tool development experience that could be incorporated into FORT in a feasible way. Lastly, future research could also focus on following assessors longitudinally to investigate changes in perceived effects over time. It is possible that assessors not involved in tool development may become increasingly comfortable with the new tools as they conduct more assessments, and may therefore find greater perceived effects after repeated use of the tools. The current study provides insight into the initial effects of training on assessors’ conduct of assessments, thus the longer-term effects of training are worth further exploration.

Future training initiatives should also consider the potential unintended consequences of training and tools. In this study, despite the mostly positive effects of learning about assessor variability on scoring behaviours, training detracted from some participants’ confidence in the program and caused concern about the potential unreliability of assessments. It may therefore be important for future training sessions to include discussions about the role of assessor judgement in WBAs and the contemporary understanding that subjectivity does not necessarily equal unfairness and objectivity does not necessarily equal reliability (5,10). It is worth noting that the unintended consequences of training were only found for participants who were not involved in tool development, thus it may be that including all assessors in the tool development process, or increasing the number of training sessions, may mitigate these adverse effects. With more frequent iterations of consensus-building over time, assessors may be able to see larger increases in consistency and see the value in the level of variability that remains.

Additionally, while many found the new tools to facilitate their processes, some reported difficulty choosing which assessment domain to organize performance information under. This difficulty may simply reflect unfamiliarity with the tools that will correct itself over time, or it may highlight an aspect of the tool that needs revision. Tavares et al. (78) found that requiring assessors to attend to multiple dimensions of performance at a time can increase cognitive load, causing assessors to feel overwhelmed or to unconsciously attend to fewer dimensions to reduce mental demands. Although it was assessors who decided on the eight-domain framework in this
study, it is possible that it was appropriate conceptually but too demanding practically. If assessors are unable to parse out the differences between domains, they may spend unnecessary time on the organization of information or may engage in idiosyncratic strategies to reduce cognitive load, which may result in inconsistency. Some participants also noted that they spent more time conducting assessments or were able to review fewer charts in the time allotted, in some cases because of the new domain framework. Given that feasibility is an important component of any assessment program (12,51), efforts must be made to ensure that the time needed to complete assessments remains reasonable.

Results also highlighted that further consideration should be given to how assessor training programs might be tailored to a program’s intended position along the formative-summative continuum. Given that WBAs often have both formative and summative goals, suggestions for how best to support assessors in balancing these and not compromising one in pursuit of the other would be helpful. Consideration should also be given to how assessor training should align with a program’s perspective on assessor variability. Accuracy and reliability have been at the forefront of many assessor training initiatives, but given the increasing awareness of the value of assessors’ subjective judgements (9,10,14), discussion is needed on how subjectivity can be acknowledged and preserved within the context of training.

Some scholars have offered suggestions for how assessor training can coincide with the notion of assessor subjectivity. Schuwirth and van der Vleuten (14) state that training should not impede judgement nor strive for objectivity, but rather be used to develop assessors’ expertise and enhance their (idiosyncratic) judgement processes. Kogan et al. (29) also note that while some assessor variability is inevitable and potentially meaningful, training can be useful for expanding assessors’ feedback vocabulary and providing opportunities for assessors to share common challenges. Training may therefore provide a community of practice for assessors and may be useful for honing their skills rather than eliminating their subjectivity. However, discussion is needed about how program administrators can effectively develop and deliver such training programs. Our results suggest that participants found a balance between the structured tools and their own professional judgement; however, not all assessors may be comfortable with this integration and some may integrate ineffectively. Assessors may also have different ideas of how much standardization is required within a given program, which may cause confusion and potentially compromise the efficiency and quality of assessments (79). Scholarly dialogue is
needed to help administrators navigate the tension between standardization and subjectivity in order to effectively educate assessors and deliver WBAs that are both authentic and valid.

The results of this study also provide validity evidence for the CPSO’s redeveloped Peer Assessment program. The perspectives of assessors about their conduct of assessments provide initial insight into how the new program operates and what its effects may be. As part of our ongoing validation efforts, other perspectives will also be explored to collect comprehensive validity evidence and to understand the impact of the program on all stakeholders. Specifically, data will be collected from Quality Assurance Committee members about the influence of the new assessment tools on their decision-making as well as from assessed physicians about the impact of the new program on their learning and professional development. These perspectives will be compared to assessors’ perspectives, as gleaned from this study, to gain a more comprehensive understanding of program effects. As is the nature of program evaluation and validation, evidence will be continuously collected and appraised to build an argument that the Peer Assessment program is achieving its intended purpose of ensuring and enhancing physician performance.

17 Strengths and Limitations

There are some limitations of this study. Firstly, this study relied on the introspection of assessors. The degree to which assessors can identify and articulate the cognitive processes they employ when conducting assessments is uncertain and the literature suggests it may be an area of difficulty for most (64,65). Therefore, it is possible that the participants in this study had limited insight into their cognitive processes and/or had difficulty articulating if and how they drew upon training or the tools when conducting assessments. However, it could be argued that one’s perceptions influence one’s actions, whether or not those perceptions reflect true cognitive processes. Therefore, attempting to gain information about what assessors perceive to have learned in training and transferred to practice is an important first step in understanding the influence of training on rating behaviours.

Secondly, the gap in time between training, the subsequent assessment, and participation in this study may have adversely affected participants’ ability to recall what was learned in training and how they transferred this to the assessment context. However, since training is intended to have lasting effects on rating behaviour, our results may reflect the most salient effects of training. In
the future, researcher may want to follow assessors over time, exploring immediate as well as longer-term perceived effects of training, to investigate what may be learned but not transferred over time.

Thirdly, it is unclear whether the differences in perceptions found between those involved in tool development and those involved only in training are related to the nature of the training methods or the amount of time assessors were involved in training. Future studies are needed to systematically examine how effects may differ based on type and amount of training.

Lastly, although assessor consistency was a goal of training, there is no way to directly test consistency of assessment scores in a single-assessor program such as the Peer Assessment Program. Therefore, we must rely on assessors’ perceptions of how they utilized tools and adhered to processes to make inferences about consistency. Other programs with multiple assessors may be able to examine assessor consistency directly and explore its relationship with assessors’ perceived conduct of assessments.

This study also has a number of strengths. It provided a qualitative exploration of assessors’ perceptions about how training and scoring tools affected their rating behaviours. Assessor training research is relatively new within the field of medical education and there is still much to be learned; our study adds to the current literature base by offering a more nuanced understanding of what assessors perceived to learn through training and how this influenced their perceived conduct of assessments. It also sheds lights on how assessment tools can influence scoring and feedback.

The rigour of our qualitative methods is also a strength. The interviewer and primary analyst had in-depth knowledge of the training process and context, which facilitated the interview process as well as analysis and interpretation. The addition of two additional researchers without this contextual knowledge helped to ensure that a balanced perspective was employed during analysis and that the primary analyst’s prior knowledge did not hinder the ability to notice unexpected trends in the data.

Finally, this study adds to the current literature by examining assessors who conduct assessments of practicing physicians within a regulatory context using WBA methods other than direct
observation. Understanding the perceptions of assessors within this environment fills a considerable gap in the literature.

18 Conclusion

This study contributes to the growing body of research on assessor training and cognition by providing insight into what assessors perceive to learn through PDT and FORT and how this, and the use of consensus-derived scoring tools, influences their conduct of assessments. Our results provide support for the continued study of PDT and FORT, with specific implications and suggestions for the development of future assessor training programs. This study also contributes to the validation of the CPSO’s redeveloped Peer Assessment program by providing evidence for scoring and implications. While validation will be ongoing, assessors’ perceptions of their scoring and feedback is an important first step in understanding the delivery of assessments and the potential effects of this WBA program.
References


75. Harrison CJ, Könings KD, Schuwirth L, Wass V, van der Vleuten C. Barriers to the


Appendix A: Sample Assessment Tools for Family Medicine

1. Patient record selection protocol

Patient Record Selection

Number of records: In total, the assessor will review approximately 15-18 patient records.

Timeframe: All patient records should be for index visits dating at least 6 months prior to the notification of assessment date\(^1\). An index visit is the date of a patient’s visit as recorded on the clinic’s day sheet. It represents the entry point in the patient record where the assessor begins their review (which may extend forwards or backwards in time to other appointments, as required, to collect an informed impression of patient care and documentation).

Selection Process:

- Prior to the assessment, the physician to be assessed will retrieve appointment schedules (day sheets or equivalent) corresponding to approximately 110 different patient visits. Appointment schedules should minimally specify patient name, date of index visit, and diagnosis/presenting complaint for that index visit. Of the list of 110 visits, the assessed physician will select 14 associated patient records that are representative of his/her practice (see Types of Records list below).
- On the day of the assessment, the assessor will select 7 of the 14 patient records chosen by the assessed physician and will also select an additional 8-11 patient records from the initial list of 110 appointments.

Types of Records: The 15-18 records reviewed by the assessor should include the following index visits, when possible:

- Preventive Care\(^2\) (e.g., cancer screening, immunization, etc.) (minimum 1 male, 1 female)
- Prenatal care (minimum 1)
- Well baby check (minimum 1, including an 18 month check if possible)
- Chronic health conditions (minimum 1 hypertension, minimum 2 diabetes, minimum 1 complex care case (i.e., patient with multiple co-morbidities))

---

\(^1\) The notification of assessment date is the date on the “notification of peer assessment” letter sent from the CPSO to the physician to be assessed and the date in the engagement letter sent from the CPSO to the assessor.

\(^2\) The incorporation of preventive care into regular appointments with patients is replacing the former practice of dedicated periodic health assessments.
- Psychosocial/mental health (minimum 1 depression or anxiety, and minimum 1 other which may include domestic violence or non-prescription substance abuse)
- Chronic non-cancer pain management with or without opioids (minimum 1)
- Acute care management (minimum 2, including at least 1 respiratory tract infection)

**Patient Record Review**

The assessor will review patient records in sufficient detail, forwards and backwards in time from the date of selected index encounter, to evaluate:

- The care and documentation provided for the presenting condition on the index date
- Management of chronic, ongoing medical conditions over time
- Cumulative Patient Profile
- Comprehensiveness of family/general practice and continuity of care (e.g., immunizations, screenings, blood pressure monitoring, blood glucose monitoring, etc.)
- Completeness of the “narrative of the patient” (from CPP, patient encounter notes, etc.)
- Physician-initiated “Opportunistic Care” (e.g., preventive healthcare interventions, etc.)

When relevant, the assessor will review the following Electronic Medical Record components/screens (or hard copy equivalents) to ensure the entire patient record has been reviewed:

- Preventive Care (immunizations, cancer screening, etc.)
- Medications
- Allergies
- Medical History / Risk Factors
- Social History
- Family History
- Physicals
- Lab work
- Referrals
- Reminders
- Chronic Disease Flowsheets
2. Scoring rubric for Examination domain:

**EXAMINATION:**

Guided by the presenting problem, a systematic evaluation of the patient’s physical and/or mental state.

Key CPSO Policies: Medical Records

**ELEMENTS OF QUALITY**

1) **Physical examinations** were completed based on presenting complaint, with relevant documentation of:
   a. Pertinent positive and negative findings
   b. Physical measurements and vital signs, where appropriate
   c. Relevant descriptive information (e.g., dimensions indicating spread of cellulitis at presentation, quality of respiratory sounds, description of rash)
   d. Illustrations of conditions, where appropriate (e.g., location of rash, laceration, abdominal tenderness)

2) **Mental health examinations** were completed when indicated, with relevant documentation of:
   a. Mental Status Examinations (MSEs) (e.g., mood and affect (including risk of harm to self/others), appearance, attitude, behavior, speech, thought process, thought content, perception, cognition, insight and judgment)
   b. Interplay of psychological and physiological factors

3) **Standardized Measures** were completed when indicated, with relevant documentation of:
   a. Scoring flow sheets (e.g., PHQ-9, min-mental state exam, pain scale)

**EVALUATION CRITERIA:**

<table>
<thead>
<tr>
<th>Score</th>
<th>Opportunities for Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Little to no improvement is needed when the trend shows that most elements of quality were evident and deficiencies, if any, were minor. Examples include:</td>
</tr>
<tr>
<td></td>
<td>• Examinations sometimes included components not relevant to the presenting complaints</td>
</tr>
<tr>
<td></td>
<td>• Mental status examinations were present but could be expanded upon</td>
</tr>
<tr>
<td>2</td>
<td>Moderate improvement is needed when the trend shows some elements of quality were lacking, but the likelihood of adverse patient outcomes was low. Examples include:</td>
</tr>
<tr>
<td></td>
<td>• Descriptions of general appearance, level of alertness, and comfort level were minimal</td>
</tr>
<tr>
<td></td>
<td>• Relevant physical measurements were not consistently present (e.g., height, weight, and BMI for preventive care and other assessments)</td>
</tr>
<tr>
<td></td>
<td>• Physical examinations tended to lack focus on presenting complaints and relevant histories</td>
</tr>
<tr>
<td></td>
<td>• Physical examinations were often not thorough enough to fully assess current presentations (e.g., repeated diabetic assessments with no evidence of a foot examination)</td>
</tr>
<tr>
<td></td>
<td>• Important, relevant descriptive information (e.g., dimensions indicating spread of cellulitis at presentation) was often not included</td>
</tr>
<tr>
<td></td>
<td>• Illustrated/described conditions (e.g., location of rash, laceration, abdominal tenderness) were often not included when appropriate</td>
</tr>
<tr>
<td></td>
<td>• Observations tended to be poorly described</td>
</tr>
<tr>
<td></td>
<td>• Key elements of examinations (e.g., pertinent positive and negative findings) were often not documented</td>
</tr>
</tbody>
</table>
**Significant improvement** is needed when the trend shows many elements of quality were lacking, or when patient outcomes could be adversely affected. Examples include:

- Pertinent vital signs (e.g., temperature and weight in child with infectious complaint) were consistently not documented
- Mental status examinations were often not included when relevant
3. Assessment Report Template (excerpt)

**PEER ASSESSMENT REPORT**

**Relevant Background Information:**

**Ratings and Comments**

**History:** A record of information gathered through questioning the patient or others (e.g., family members, substitute decision-maker) and reviewing pertinent documents to determine the next steps in care.

<table>
<thead>
<tr>
<th>Rating</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
</table>

**Areas of Quality Care and Suggestions for Quality Improvement:**

**Specific Concerns Requiring Attention and Recommendations for Remediation:**

**Examination:** Guided by the presenting problem, a systematic evaluation of the patient's physical and/or mental state.

<table>
<thead>
<tr>
<th>Rating</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
</table>

**Areas of Quality Care and Suggestions for Quality Improvement:**

**Specific Concerns Requiring Attention and Recommendations for Remediation:**
4. Quality Improvement Resource for Chronic Non-Cancer Pain Management

**QI Resource 3: Chronic Non-Cancer Pain Management**

*Peer Assessment: Family Medicine/General Practice*

Chronic non-cancer pain includes any painful condition that persists for three months or longer and is not associated with malignant disease (DeGreeff, 2017)

Last Reviewed: September 2017 / Next Planned Review: September 2020

Quality Improvement Resources (QIRs) provide a reference for topics that are of particular relevance to a medical discipline and which may arise during peer assessment. They are intended to promote a common framework for assessor feedback and to provide educational material for physicians. QIRs are developed by College peer assessors and regularly reviewed and updated in accordance with an established cycle. They are not clinical standards, clinical guidelines, or intended to replace the knowledge, professional skill and judgment of physicians.

QIRs are intended to be inclusive of many potential elements of quality of care related to a specific topic; however, they are not exhaustive. In addition, not all elements listed in a QIR will be applicable for every patient encounter, assessment, or plan of care.

**SUBJECTIVE:**

- **Initial assessment:** (may be done over several patient appointments)
  - The patient’s pain is well described including the type, cause and nature of the pain including past investigations and previous interventions
  - The patient’s functional impairment is described and documented
  - Medical history including past medical and surgical history, allergies (to opioids), family history of substance use/addiction and psychiatric history is documented
  - Previous history of addiction (illicit drugs or alcohol), active substance abuse disorder and current psychiatric illness are considered and documented
  - Non-pharmacological modalities and non-opioid pharmacotherapy modalities have been tried

- **Follow up patient visits:**
  - Pain intensity and quality is reassessed using a numerical scale
  - Function is assessed and functional goals are considered
  - Side effects to medications are elicited including constipation, nausea, vomiting and sedation
  - Consider possible associated psychiatric concerns and mental health concerns (e.g., depression)
  - Reassessment of use of opioids if there has been a trial and pain has not improved

**OBJECTIVE:**

- **Initial assessment:**
  - Physical examination is problem based, clearly documented and supports diagnosis
  - Investigations and consultations confirm diagnosis and prescribed treatment
  - The pain intensity is documented preferably using a numerical scale

- **Follow up visits:**
  - Physical examination is carried out with changes in patient history or condition
  - Documentation of the SAS: Activity, Analgesia, Adverse effects, Affect
  - The pain intensity and functional goals are documented preferably using a numerical scale at every visit
  - Investigations and/or consultations are arranged as indicated with changes in patient pain profile, medication complications, etc.

**ASSESSMENT:**

- **Initial assessment:**
  - Diagnosis is clearly stated and is supported by investigations (X-rays, CT/MRI), assessment and/or consultant notes
  - Addiction risk is considered – completion of “Opioid Risk Tool” and/or “CAGE questionnaire” ideally
  - Trial of tapering and referral to a pain clinic/or addiction medicine consultant when indicated

- **Follow up patient visits:**
  - Urinary drug screening (UDS) is carried out at intervals depending on risk
  - Compliance to medications is assessed
  - If opioid misuse is determined according to DSM-V criteria for substance use disorder, consideration be given to referrals for addiction, Suboxone, Methadone Maintenance Therapy, counseling
  - If patient is using ≥ 90 mg morphine equivalent per day, consideration should be given to tapering opioids and/or referral to specialist (as appropriate)
**QI Resource 3: Chronic Non-Cancer Pain Management continued**

**PLAN:**

- **Initial Assessment:**
  - Review goal of therapy: pain reduction and improved function (with or without pain elimination)
  - Physical therapy/modalities, education and counseling are prescribed where appropriate
  - Prescribed medications are indicated for diagnosis, e.g., opioids are not effective for fibromyalgia, headaches, low back pain, neck pain
  - Alternative non-opioid medications are prescribed where appropriate for the pain syndrome
  - Narcotic prescribing agreement is signed, in patient chart, and easily referenced

- **Follow-up patient visits for chronic non-cancer pain:**
  - At each patient visit pain level, compliance, medication side-effects and patient function are assessed
  - Documentation of the SAP: Activity, Analgesia, Adverse effects, Aberrant behaviours, Affect
  - Patient is reassessed more frequently at the current recommended "watchful dose" of morphine equivalent/day
  - Pain medications are prescribed in a stepwise fashion, slowly titrated and their effect is assessed at appropriate intervals
  - Morphine equivalent dose is documented and used to determine doses with changes in prescribed opioids
  - If patient is using > 90 mg morphine equivalent per day, consideration should be given to tapering opioids and/or referral to specialist (as appropriate)

- **General**
  - Patient information sheets/websites (as appropriate)
  - Integration of chronic non-cancer pain history, physical, treatment, and monitoring into CPP (as appropriate)

---

**Recommended Guidelines/Resources:**

  [http://nationalpaincentre.mcmaster.ca/guidelines.html](http://nationalpaincentre.mcmaster.ca/guidelines.html)

  [http://cfpc.ca/UploadedFiles/CPD/Opioid%20poster_CFP_ENG.pdf](http://cfpc.ca/UploadedFiles/CPD/Opioid%20poster_CFP_ENG.pdf)

  [http://thecapwellhealth.ca/ccep](http://thecapwellhealth.ca/ccep)

- Opioid Manager App (and additional resources) (UHN Rehabilitation Institute). Last accessed: July 12, 2016.
  [www.opioidmanager.com](http://www.opioidmanager.com)

- RxFiles Drug Comparison Charts, 10th edition. Subscription required. Products listed and available for order at this link:

  [www.opd.utoronto.ca/opioidprescribing/](http://www.opd.utoronto.ca/opioidprescribing/)

- Project Echo Ontario (University Health Network — Queen’s University Pain Project). Last accessed: July 12, 2016.
  [www.echoontario.ca](http://www.echoontario.ca)

  [http://ocfp.on.ca/cpd/collaborative-networks/mmap](http://ocfp.on.ca/cpd/collaborative-networks/mmap)
Recommended Guidelines/Resources continued:


Additional resources are available from the CPD/Practice Improvement section of the CPSO website: www.cps.on.ca/CPSO-Members/Continuing-Professional-Development/CPD-Practice-Improvement-Resources/Medical-Expert-Role-Resources