Maastricht University

Inter-rater variability as mutual disagreement

Citation for published version (APA):

Gingerich, A., Ramlo, S. E., van der Vleuten, C. P. M., Eva, K. W., & Regehr, G. (2017). Inter-rater variability as mutual disagreement: identifying raters' divergent points of view. *Advances in Health Sciences Education*, *22*(4), 819-838. https://doi.org/10.1007/s10459-016-9711-8

Document status and date:
Published: 01/10/2017

DOI:
10.1007/s10459-016-9711-8

Document Version:
Publisher's PDF, also known as Version of record

Document license:
Taverne

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 09 Aug. 2022

Adv in Health Sci Educ (2017) 22:819–838
DOI 10.1007/s10459-016-9711-8

Inter-rater variability as mutual disagreement: identifying raters' divergent points of view

Andrea Gingerich[1] · Susan E. Ramlo[2] · Cees P. M. van der Vleuten[3] · Kevin W. Eva[4] · Glenn Regehr[4]

Received: 13 April 2016 / Accepted: 9 September 2016 / Published online: 20 September 2016
© Springer Science+Business Media Dordrecht 2016

Abstract Whenever multiple observers provide ratings, even of the same performance, inter-rater variation is prevalent. The resulting 'idiosyncratic rater variance' is considered to be unusable error of measurement in psychometric models and is a threat to the defensibility of our assessments. Prior studies of inter-rater variation in clinical assessments have used open response formats to gather raters' comments and justifications. This design choice allows participants to use idiosyncratic response styles that could result in a distorted representation of the underlying rater cognition and skew subsequent analyses. In this study we explored rater variability using the structured response format of Q methodology. Physician raters viewed video-recorded clinical performances and provided Mini Clinical Evaluation Exercise (Mini-CEX) assessment ratings through a web-based system. They then shared their assessment impressions by sorting statements that described the most salient aspects of the clinical performance onto a forced quasi-normal distribution ranging from "most consistent with my impression" to "most contrary to my impression". Analysis of the resulting Q-sorts revealed distinct points of view for each performance shared by multiple physicians. The points of view corresponded with the ratings physicians assigned to the performance. Each point of view emphasized different aspects of the performance with either rapport-building and/or medical expertise skills being most salient. It was rare for the points of view to diverge based on disagreements regarding the interpretation of a specific aspect of the performance. As a result, physicians' divergent points of view on a given clinical performance cannot be easily reconciled into a single coherent assessment judgment that is impacted by measurement error. If inter-rater variability does not wholly reflect error of measurement, it is problematic for our current

✉ Andrea Gingerich
andrea.gingerich@unbc.ca

[1] Northern Medical Program, University of Northern British Columbia, 3333 University Way, Prince George, BC V2N 4Z9, Canada

[2] Department of Engineering and Science Technology, University of Akron, Akron, OH, USA

[3] School of Health Professions Education, Maastricht University, Maastricht, Netherlands

[4] Centre for Health Education Scholarship, University of British Columbia, Vancouver, BC, Canada

measurement models and poses challenges for how we are to adequately analyze performance assessment ratings.

Keywords Inter-rater variability · Mini-CEX · Q methodology · Rater-based assessment · Rater cognition · Workplace-based assessment

Introduction

Human judgment has been considered indispensable to programs of assessment in medical education (Schuwirth and Van der Vleuten 2011). However, whenever multiple observers provide ratings, even of the same performance, inter-rater variation is prevalent (Crossley and Jolly 2012). This inter-rater variation is often interpreted as the result of raters committing mistakes, making omissions or being biased (Albanese 2000; Downing 2005; Williams et al. 2003). The resulting 'idiosyncratic rater variance' is considered to be unusable error of measurement in psychometric models (O'Neill et al. 2015) and can be of sufficient magnitude to threaten the defensibility of our assessment decisions (Crossley et al. 2002; Downing 2004). Thus, the majority of research exploring rater cognition has searched for the controllable judgment processes as well as the unconscious cognitive biases that may underlie rating variability (Gauthier et al. 2016; Gingerich et al. 2014a; Kogan et al. 2011; Tavares and Eva 2013; Williams et al. 2003; Wood 2014).

Recently, medical education researchers investigating inter-rater variability have found that raters sometimes emphasize different aspects of the performance (i.e. seeing different aspects as most important), sometimes outright disagree on the same aspects of the performance (i.e. seeing the same aspect differently) and sometimes make unchecked social inferences (e.g. inferences regarding personality traits and motives) (Gauthier et al. 2016; Govaerts et al. 2013; Herbers et al. 1989; Kogan et al. 2011; Mazor et al. 2007; Yeates et al. 2013). These findings and interpretations are certainly consistent with the conceptualization of rater variability as idiosyncratic 'rater error'. In a study exploring inter-rater variation, however, Gingerich et al. (2014b) discovered patterns in raters' responses that might suggest their judgments are not as idiosyncratic as they have been characterized. More specifically, multiple clusters of consensus were identified for each clinical performance with each cluster of consensus containing several physician raters who reported similar impressions or interpretations of the given clinical performance. The content and valence of the impressions described often varied widely between clusters, suggesting that the multiple clusters of consensus represented different, but not entirely idiosyncratic, perspectives on the performance. These findings were consistent with research from the social psychology literature which suggests that in social interactions people will tend to differently (but not entirely idiosyncratically) categorize those they are observing based on the social inferences they make about the performer (Fiske et al. 2007; Macrae and Bodenhausen 2000; Mohr and Kenny 2006; Park et al. 1994). Importantly, these clusters of consensus for a given clinical performance (i.e. accounting for the cluster to which each rater belonged) often explained a significant proportion of variance in raters' scores of the performance.

While the research to date has offered important insights into rater cognition, all of the previous investigations have used open response formats, such as interviews or text boxes, as tools for collecting raters' comments and justifications for their ratings (Chahine et al. 2016; Gauthier et al. 2016; Gingerich et al. 2014b Govaerts et al. 2013; Herbers et al. 1989; Kogan et al. 2011; Mazor et al. 2007; St-Onge et al. 2016; Tavares et al. 2016; Tweed and Ingham 2010; Yeates et al. 2013, 2015). This has been an important design element in

these early studies–allowing raters the freedom to articulate their own ideas without being limited by the researchers' preconceived notions. However, it is also a potentially limiting design element because participants have been found to provide *less* detailed responses when using an open text format than when they were prompted to provide more structured responses (Herbers et al. 1989). Therefore, the flexibility of the open response format may allow raters to disproportionately emphasize some features, while omitting other features that were nonetheless salient and potentially influential in their assessment judgments. If so, the resulting open text responses would provide a distorted representation of the underlying rater cognition and their subsequent analysis would identify variability in rater cognitions that is better explained as an artifact of the study design than it is attributable to meaningful differences in raters' cognitions.

This study, therefore, aims to triangulate recent rater cognition findings by using a methodology that removes any variability that might be introduced into raters' responses by use of an open response format. In Q methodology every participant is presented with the same set of statements and is instructed to indicate which statements are most salient (according to their personal viewpoint on the topic of investigation) by sorting them in relation to all of the other statements (Stephenson 1953; Watts and Stenner 2012). Q methodology was chosen because it requires that participants explicitly reflect on each of the potentially salient features of the clinical encounter (after, not during their ratings) and requires them to create a relative ranking that more clearly indicates their opinions about things that might simply be left unsaid in a free response format (Brown 1980). A specified set of analytic procedures can then be used to identify the different points of view held by the participants (Stephenson 1953; Watts and Stenner 2012). Q methodology, therefore, offers the opportunity to extend the rater cognition literature both by replicating previous findings using a different method and by providing novel insights that arise from the analyses that the methodology affords. If groupings of raters with similar viewpoints (based on Q-factors) exist and explain substantial variability in ratings, then examination of the basis for the groupings would help determine if the commonly held view that inter-rater variability is simply representative of measurement error can be maintained. That is, Q methodology allows us to explore the extent to which the differences in the groups' perspectives arise from disagreements about what happened in the encounter, or disagreements about the importance of what happened.

Our ultimate goal for the overarching line of research is to discover new ways to improve rater-based assessments by better understanding the variability in ratings introduced by raters. The specific aim of this study was to use the structured response procedures of Q methodology to investigate inter-rater variability by identifying (1) consensus within assessment impressions (2) divergence between multiple assessment impressions for a given performance and (3) the relationship between shared assessment impressions and the clinical assessment ratings that are assigned.

Methods

Q-sort procedure and analysis

Q methodology consists of philosophy, theory and a set of procedures that focus on describing a population of viewpoints rather than describing a population of people

(Newman and Ramlo 2010; Watts and Stenner 2012). Q methodology can be understood as following five steps that align with the three aims of our study.

Step 1: Identifying salient aspects within the clinical performances

The first step in the design of a Q-study is identifying a set of statements that will be sorted by participants. It begins with gathering a large array of statements that are broadly representative of the topic of interest, known as the concourse (McKeown and Thomas 1988; Watts and Stenner 2012). Our concourse needed to include as many salient aspects of the clinical encounters that influence assessment judgments as possible. Therefore, it was essential to utilize data collected from previously studied clinical performances to generate a set of statements that contained all salient aspects of the performances. To make the eventual task described in Step 2 feasible for participants (i.e. to limit the amount of time required to approximately 1 h), we selected a set of four video-recorded clinical performances (videos 1, 2, 6, 7) from seven used in an earlier study (Gingerich et al. 2014b). The four videos were selected because they represented: (1) a set of clinical performances demonstrating a mix of competency levels in interpersonal skills and clinical skills; and (2) a range (based on the previous study) in the proportion of variance in the mini-clinical evaluation exercise (Mini-CEX) ratings explained by cluster membership (from 9 to 57 %). In the previous study, physicians provided open-text responses to three questions requesting comments on the resident's clinical competence along with their own social judgments and impressions of the resident. These were compiled from the four videos and the resulting concourse contained 317 complete open text responses ranging from a couple of words to a couple of paragraphs in length.

Once the full concourse is identified, a subset of statements is selected to form the Q-sample. The Q-sample should be a reasonable size for participants to work with but remain representative of the concourse and be balanced across the possible points of view that might be found during analysis (McKeown and Thomas 1988; Watts and Stenner 2012). We used the most formal approach to developing a structured Q-sample, known as Fisher's Design of Experiments approach (Brown 1980; Stephenson 1953; Watts and Stenner 2012). We began with one author (AG) reducing the concourse to the most distinctive responses for each video. This reduced the number to 195 complete open text responses. These were then parsed into shorter statements of suitable length to be used in a Q-sort. Duplicates were removed and semantically similar statements were combined (e.g. statements such as "a likeable person", "very likeable, very personable", "well respected by peers and patients", "well-liked by fellow residents" were combined to form a single statement: *Is liked by peers and patients*). Two authors (AG and SER) then conferred to select the most iconic statements in a way that balanced the number of statements referring to each of the subscales on the Mini-CEX and different types of social judgments (e.g. naming personality traits, inferring state of mind etc.) for each video. This resulted in a set of 60 statements. Because the Q-sort process enables one statement to be placed in different grid positions to convey opposing opinions (see Fig. 1 and more details below) there was no need for opposing descriptors to be included (e.g. we could include 'efficient' or 'inefficient' but did not need both). Removal of one item from each pair of "opposites" in the sample reduced the number of statements to 44. Through two rounds of pilot testing, the phrasing of the statements was refined in response to participants' expressions of uncertainty, but the original phrasing from participants in the previous study was maintained as much as possible. The final list of 44 statements (displayed in Table 1) contained 11 statements representing the Mini-CEX subscales of counseling and humanistic qualities/

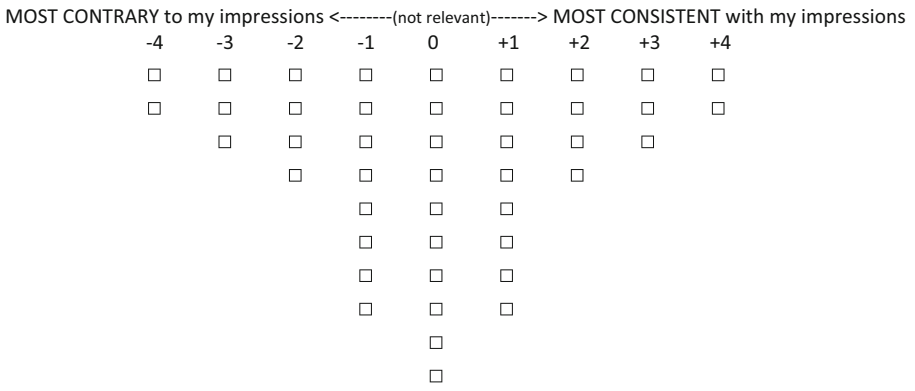

Fig. 1 Grid positions on Q-Sorting Score Sheet. Participants rank-ordered 44 statements by placing them on a quasi-normal distribution scoring sheet. They had previously sorted the statements into three piles: consistent with my impression, contrary to my impression and not relevant to my impression. When presented with the scoring sheet they were instructed to select from the 'consistent' pile the 2 statements that were most consistent with their impression and place them under '+4'. The next three statements that were most consistent with their impression were placed under the '+3' and so on. The same process was used to rank the statements from the 'contrary' pile. The statements in the 'not relevant' pile were then placed in the remaining spots on the score sheet

professionalism (grouped under the heading of rapport-building skills); 18 statements representing the Mini-CEX subscales of medical interviewing skills, clinical judgment and organization/efficiency skills (grouped under the heading of medical expertise skills); and 15 statements containing social inferences not related to rapport-building or medical expertise.

Step 2: Data collection using the structured response format of a Q-sort

The second step in Q methodology is the completion of a Q-sort by participants. During a Q-sort each participant's point of view on the specified topic (in this case, the performance of the resident) is collected by having them sort the statements in the Q-sample using a forced quasi-normal distribution ranging from "most consistent with my perspective" to "most contrary to my perspective" (Brown 1980; Watts and Stenner 2012). This technique, therefore, allowed the researchers to capture the subjectivity of physicians' assessment impressions in the form of their Q-sorts using the Q-sample and distribution provided. These Q sorts, therefore, offer a structured response format to indicate the most salient aspects of the clinical performance selected from a large set of potentially relevant statements by a diverse set of raters.

To operationalize this process, an invitation email was sent to participants that included a link to the website QSortOnline. After giving consent to participate, participants viewed a randomly presented video and completed Mini-CEX ratings. The Mini-CEX form (Norcini et al. 2003) consisted of 6 dimensions: Medical Interviewing Skills; Humanistic Qualities/ Professionalism; Clinical Judgment; Counseling Skills; Organization/Efficiency; and Overall Clinical Competence), each associated with 9-point scales (1-2-3 unsatisfactory; 4-5-6 satisfactory; 7-8-9 superior). The other commonly used Mini-CEX dimension (Physical Exam Skills) was not included because it was not applicable to the videos. Upon completing the Mini-CEX rating, participants were asked to complete a Q-sort using the following instructions:

Table 1 Comparison of the factor arrays produced by Q methodology to determine differences in the Q-sort configurations of the same 44 statements, conducted separately for four different video-recorded clinical performances

Statement number and statement		Video 1		Video 2		Video 6		Video 7	
		Q-Factor 1 'disinterested'	Q-Factor 2 'diligent competence'	Q-Factor 1 'compassionate competence'	Q-Factor 2 'incomplete attempt'	Q-Factor 2 'no rapport'	Q-Factor 1 'incompetent'	Q-Factor 2 'competent rapport'	Q-Factor 1 'inexperienced incompetence'
Rapport-building statements[a]									
#7	Appeared compassionate or showed genuine concern	**–4**	**–4**	**4****	–1	**–4**	–1**	**3**	1**
#8	Connected well with patient	**–4**	**–4**	**3****	–2	**–4**	**–3****	**4**	1**
#15	Was friendly and engaged	**–3**	**–3**	**3****	0	–2	1**	**3**	**3**
#42	Was willing to listen and unhurried	**–3**	**–3**	**4****	–1	**–3**	–2**	**3**	1**
#43	Welcome and closing carried out with courtesy	**–3***	**-3**	**3**	2	–2	–1**	**4**	1**
#5	Body language demonstrates disinterest	**4****	2	**–4***	**–3**	**3**	–1**	**–3**	**–3***
#4	Used a rapid-fire questioning style	**4****	**4**	**–3****	0	0	0	–1	–1
#12	Seemed dismissive and disrespectful	2**	0	**–4****	–2	**3**	1**	**–4**	**–3**
#9	Was detached and distant	**3***	2	**–3****	–1	2	–2**	**–3**	–2
#14	Focused on the task instead of the patient	**3***	**4**	–1*	2	1	1	–2	1**
#31	Used a paternalistic approach	1	2	–1**	1	2	1**	–2	–2
Medical expertise statements[b]									
#41	Thorough, organized and efficient history	–1**	**3**	2**	–2	**–3**	**–4****	2	**–4****
#40	Systematically considering all differentials	–2**	2	1**	**–4**	**–3**	**–4****	0	**–4****
#38	Is intelligent and has solid clinical knowledge	–1*	3	2**	0	–1	–1**	1	–2**
#11	Did not fully investigate chief complaint	1**	–1	–1**	**3**	**4**	**4**	1	**4****
#6	Used a checklist style of questioning	**3**	**3**	0**	2	0	–2**	1	2**
#27	Too many closed and leading questions	2**	1	0**	1	0	0	–1	–1
#3	Asked questions but didn't follow through on responses	2**	–1	0**	2	0	0	0	1**
#26	Many questions were not relevant	0*	–2	–1**	**–3**	–1	–1	–1	2**
#13	Failed to inquire about red flag symptoms	1**	–1	–2**	0	1	**4****	0	**3****
#10	Details about patient's life are missing	1	1	1**	**3**	1	2	0	–1
#18	Inadequate baseline established	0	0	0*	1	2	1	–1	1**
#29	No idea what the diagnosis was	0*	–2	–2	–1	–1	0	0	**4****
#1	Anchored to a diagnosis too early	1**	–1	0**	**3**	**4**	**3**	0	–2**
#35	Omitted relevant tasks	1	0	0**	1	1	**3****	0	2**
#22	Jumps too quickly to treatment	0	0	0**	**4**	**3**	**3**	0	–1**
#20	Insufficient non–pharmacological options	0	0	2**	**4**	1	2	0	0
#37	Was ill–prepared for this visit	0*	–1	–2	–1	0	1	–1	1**
#19	Inexperienced with this condition	–1	–1	0**	1	0	0	0	**3****
Social inferences and judgments[c]									
#33	Is reliable and trustworthy	–1*	1	1*	0	**–3**	–1	0*	2
#21	Is mature; has some life experience	–1*	0	0*	0	0*	–1	–1*	1
#25	Is liked by peers and patients	–1	–1	1*	1	–1	–1	0**	2

Table 1 continued

#39	Someone I'd look forward to working with	−2**	−2	2**	0	−3	−1	0**	1
#28	Is really trying hard	−2*	1	1*	0	−2	−2	0*	1
#16	Has good intentions	−1*	1	1	1	0*	0	2	2
#34	Seems receptive to feedback	−1*	0	1	0	−1**	−2	0*	1
#44	Will do OK in the long run	−2*	1	1	1	−1	−1	0*	1
#2	Is arrogant and overconfident	1	1	**−3**	**−3**	0**	1	**−3****	**−4**
#24	Is lazy or not diligent	0**	−2	−1**	**−4**	1	1	−1**	−3
#23	Going through the motions	2**	0	−1*	0	0*	1	0**	−1
#24	Is lazy or not diligent	0**	−2	−1**	**−4**	1	1	−1**	−3
#23	Going through the motions	2**	0	−1*	0	0*	1	0**	−1
#30	Looks tired	1*	0	0**	−1	0	0	−1	−1
#32	Patient will not follow advice	0	1	−1	−1	2	2	0**	−1
#36	Should not have made it this far	0**	−1	−2	−2	1*	0	0**	−2
#17	This resident is going to be sued one day	0	0	−1	−1	2**	0	−1**	−2

Shading is used to indicate the point of view's more salient aspects within the assessment impressions

Bold font is used to highlight the most distal grid positions on the Q score sheet

* Statements distinguish Factor 1 from Factor 2 ($p < .05$)

** Statements distinguish Factor 1 from Factor 2 ($p < .01$)

[a] Rapport statements included those referring to Mini-CEX subscales counseling skills and humanistic qualities and professionalism

[b] Medical expertise statements included those referring to Mini-CEX subscales medical interviewing skill, clinical judgment and organization/efficiency

[c] Social judgment statements included those referring to social inferences not directly related to skills in rapport-building or medical expertise

> We would like you to share your honest and unfiltered impressions of this resident and ask that you rank the statements from 'MOST *consistent* with my impressions of this resident and their performance' to 'MOST *contrary* to my impressions of this resident and their performance'.

FlashQ software was used to facilitate the Q-sort, which was completed in multiple steps (see Fig. 1 for more details) as is recommended for Q-sorting tasks (Newman and Ramlo 2010; Watts and Stenner 2012). After completing the Q-Sort, participants were prompted to explain (a) why they selected the two statements in the "−4" grid positions as being most contrary to their impression, (b) why they selected the two statements in the "+4" grid positions as being most consistent with their impression, and (c) any problems they encountered with performing the sorting task. They could then choose to exit the program or to move on to the next randomly presented video and repeat the process for any number of the remaining three videos. Finally, they were asked to complete a short demographics questionnaire.

Step 3: Identifying consensus in assessment impressions through analysis of the Q-sorts

The third step in Q methodology is the analysis of the Q sorts which enables us to identify how many points of view there are for a given clinical performance along with how many physicians share each of those points of view. This is possible because participants with similar points of view are expected to sort the statements in a similar way. Similar Q-sorts are highly correlated and, therefore, participants and their Q-sorts can be grouped together

into Q-factors via factor analysis (Stephenson 1953; Watts and Stenner 2012). This is called a 'by-person' factor analysis because it groups together participants with highly correlated Q-sorts into a factor, just as a conventional factor analysis uses a 'by-item' matrix to group together highly correlated items into a factor (Stephenson 1953; Watts and Stenner 2012).

We analyzed the Q-sort data using free custom software PQMethod 2.35 (Schmolck 2014). We used the classic centroid technique for factor extraction followed by varimax rotation of the factors. In considering how many factors to extract, we paid special attention to those with eigenvalues >1 (Watts and Stenner 2012); those exceeding Humphrey's rule (i.e. those for which the cross-product of the two highest loadings for a factor in the unrotated matrix exceeded twice the standard error) (Watts and Stenner 2012); and those on which at least two raters loaded significantly ($p < .01$) (Brown 1980; Watts and Stenner 2012). It is important to note, however, that theoretical significance is more important than statistical significance in Q methodology (McKeown and Thomas 1988). As a result, each factor solution was examined for fit and interpretability with the best solution selected. Each video was analyzed separately and all Q methodology analysis was completed with researchers blind to the corresponding Mini-CEX ratings.

Step 4: Characterizing each point of view through Q-factor interpretation

The fourth step of Q methodology is interpretation of the Q-factors to reveal the points of view reflected by each. All Q-sorts that are grouped into a given factor have similar sorts or 'configurations' of the statements. However, those configurations are not identical. Thus, one of the analyses performed by the PQMethod software is the identification of a representative Q-sort for each Q-factor (each column of Table 1), known as the 'factor array' (McKeown and Thomas 1988; Newman and Ramlo 2010). This factor array is used to interpret the point of view associated with each factor using procedures described by Watts and Stenner (2012).

Step 5: Identifying points of divergence between Q-factors

The 'factor matrix' displays all the factor arrays side by side, showing every statement and its grid position for each Q-factor. This enables comparisons of each statement across factors. For example, the matrix could indicate that a particular statement was placed in the "+3" grid position in Factor 1 and the "−4" position in Factor 2. This comparison provides two pieces of information that can be used as indications of how the same performance features may have been differently interpreted. First, if participants with Q-sorts that highly correlate with one Factor place a particular statement in a distal grid position (e.g. "−4", "+4", "−3" or "+3") and participants with Q-sorts that highly correlate with another Factor place the same statement in a central grid position (e.g. "−1", "0", "+1") it can be inferred that the performance feature is more prominent or salient for the first set of participants compared to the other. Second, if a statement is placed on the 'contrary to my impression' side of the grid (e.g. "−4", "−3", "−2") in one Factor and on the 'consistent with my impression' side of the grid (e.g. "+4", "+3", "+2") in another Factor, this could be an indication of disagreement in the interpretation of the performance feature. Interpreting these patterns across items allows us to determine how the two factors represent different points of view (i.e. systematic differences in rater cognition).

Identifying the relationship between Q-factors and Mini-CEX ratings

To determine if differing points of view were related to the ratings physicians assigned, participants were assigned to the Q-factor with which their Q-sort was most highly correlated and then Q-factor assignment was used as the independent variable in a one-way ANOVA to determine the proportion of variance that could be explained (partial eta squared) in the 'overall clinical competence' Mini-CEX ratings. Analyses were performed separately for each of the four videos.

Participants

Q methodology requires purposeful recruitment of participants to cover all possible viewpoints on the topic because each participant is considered a *variable* in the by-person factor analysis (McKeown and Thomas 1988; Newman and Ramlo 2010; Watts and Stenner 2012). The goal was, therefore, to include a diverse range of clinical assessors who were responsible for judging the competence of medical residents in real-life. Because we strove to capture their authentic assessment impressions and ratings no rater training was provided. Thus, to recruit participants for this study, we asked colleagues to approach, on our behalf, physicians who they considered to be good, well-respected and experienced assessors of residents. To be eligible, assessors must have been licensed to practice medicine in North America at some point in their career. In addition, they must have been trained in emergency, family or internal medicine because those were the specialties deemed relevant to the content of the cases presented on video. Experts in assessment who are widely known for their contributions to assessment research, rater training, or assessment policy (and who met the inclusion criteria) were individually invited to participate. Since participants are the study variables in Q methodology, it is important to undertake a recruitment strategy that avoids study of an overly homogeneous and unrepresentative set of individuals. As a result, as is recommended, Q-analyses were performed periodically throughout data collection and whenever a factor associated with unique demographic characteristics or a particular assessment style (e.g. severe/lenient) started to emerge, we asked our colleagues to invite additional participants who had similar characteristics in an effort to ensure adequate sampling of that variable. Recruitment was stopped once the factors became established and the number of gathered Q-sorts was between the total number of statements in the Q-sample and half that number (i.e. 22–44 participants per video) (Watts and Stenner 2012).

This study was approved by the Behavioural Research Ethics Board at the University of British Columbia and the Research Ethics Board at the University of Northern British Columbia.

Results

Participants

Between November 2014 and February 2015, 46 unique participants submitted a total of 128 Q-sorts by sorting the same 44 statements in response to 1–4 videos. The participants were from 19 different cities in 5 provinces in Canada and 5 states in the USA. There were 24 (52 %) from internal medicine, 13 (28 %) from emergency medicine, 9 (20 %) from

family medicine. Twenty-three participants were female (50 %). The physicians had been in practice for an average of 13.3 years (range 1–35) with an average of 10.8 years of experience with assessing residents (range 1–35).

Identifying clusters of consensus through Q-Factor Analysis

If all participants had shared a single point of view on the clinical performance, we would expect them to sort the statements in a similar configuration and a single Q-factor to be identified. This did not occur. Instead, a 2-factor solution was determined to be the best fit for each of the four videos (see Table 2 for details regarding factor extraction and rotation). In other words, based on subsets of similar Q-sorts the analysis revealed two major clusters of consensus among participants' impressions for each of the clinical performances.

It is important to note that it is possible for each Q-factor to have two points of view associated with it. This is possible because participants' Q-sorts can be either positively or negatively correlated with a single Q-factor. Therefore, two Q-factors could represent four points of view, in theory. In other words, if there are physicians with Q-sorts that are positively correlated with a Factor and physicians with Q-sorts that are negatively correlated with the same Factor, they might be considered to have directly opposing points of view with regard to the resident's performance. Physicians with Q-sorts that correlate with different Factors might best be described as having different, but not necessarily opposing points of view. For three videos (1,2,7), participants' Q-sorts were only positively correlated with either one factor or the other. As such, there were two distinct (but not opposite) points of view, each shared by 6 or more physicians, for these videos. For video 6, however, 17 participants' Q-sorts were positively correlated with one factor, 14 were positively correlated with the other and one was *negatively* correlated with the latter factor. Thus, there were three different points of view, one of which was represented by only a single participant. In summary, when physicians' assessment impressions were provided in the structured response format of a Q-sort, the similarity of their responses generally converged on two clusters of consensus.

Characterizing each perspective through Q-factor interpretation

After identifying the number of points of view associated with each of the four clinical performances, we examined the configuration of the statements in the factor arrays. An interesting pattern emerged as we identified the clinical features that had been differently interpreted within each point of view. As highlighted using grey shading in Table 1, one point of view for every video (the factor listed first) used the distal position on the grid (± 3 and ± 4) almost exclusively to represent rapport-building statements; with the medical expertise and social judgment statements being placed in less extreme positions (0 to ± 2). Conversely, the other point of view used the distal positions almost exclusively to represent statements referring to medical expertise leaving the less extreme positions to represent the rapport building and social judgment statements. Based on these sorting configurations, it appears that for most videos, one group of physician raters emphasized rapport-building skills most prominently in their assessment impressions whereas the other group emphasized medical expertise skills as most salient.

We will describe this analysis in more detail using Video 2 as an example. For the first Q-factor (third column of numbers in Table 1), the rapport building statements were consistently assigned large positive or negative values (± 3 or ± 4) with the high positive valences being associated with positive statements (compassionate, unhurried, friendly,

Table 2 Summary of Q-Factor extraction and interpretation

Video 1	
Summary of clinical performance	Female resident interviews female patient regarding respiratory symptoms: scripted to have superior medical expertise and satisfactory interpersonal skills
Factor extraction and rotation	We noted 3 factors had eigenvalues greater than 1 (18.36, 2.55 and 1.04) but only a single factor exceeded Humphrey's rule. A 2 factor solution was chosen because it created two interpretable points of view and had 5 sorts that significantly loaded onto the second factor in the unrotated correlation matrix. Varimax rotation grouped 18 participants into factor 1 and 14 into factor 2
Consensus between points of view	Participants grouped into both factors agreed that she used a rapid-fire (#4; +4, +4) checklist-type (#6; +3, +3) questioning style that focused more on the task instead of the patient (#14; +3, +4). She did not connect well with the patient (#8; −4, −4) and the welcome and closing were not carried out with courtesy (#43; −3, −3). She came across as detached and distant (#9; +3, +2), hurried and not willing to listen, (#42; −3; −3) not friendly or engaged (#15; −3, −3) or compassionate or genuinely concerned (#7; −4, −4)
Factor 1 point of view (based on positively correlated Q-sorts)	*Disinterested* Her body language was demonstrating disinterest (#5; +4) and she seemed dismissive and disrespectful (#12; +2) like she was just going through the motions (#23; +2) and not trying very hard (#28; −2). Concerned for how she will do in the long run (#44; −2) as she asked too many closed and leading questions, (#27; +2) didn't follow through on the responses (#3; +2) and really didn't systematically consider all the differentials (#40; −2)
Factor 2 point of view (based on positively correlated Q-sorts)	*Diligent competence* She was intelligent with solid clinical knowledge (#38; +3) and performed a thorough, organized and efficient history (#41; +3) while systematically considering all differentials (#40; +2) to arrive at the diagnosis (#29; −2). She seemed diligent and not lazy (#24; −2)
Demographics	The Q sorts were submitted by 32 participants (14 male, 18 female; 16 internal medicine, 9 emergency medicine, 7 family medicine). Factor 1: 18 participants (6 male and 12 female) from 11 different cities; 10 internal medicine, 5 emergency medicine and 3 family medicine. Factor 2: 14 participants (8 male and 6 female) from 8 different cities; 6 internal medicine, 4 emergency medicine and 4 family medicine
Video 2	
Summary of clinical performance	Female resident interviews male patient, diagnoses depression and provides clinical management: scripted to have satisfactory medical expertise and satisfactory interpersonal skills
Factor extraction and rotation	We noted 3 factors had eigenvalues greater than 1 (16.03, 3.91 and 1.60) but only 2 factors exceeded Humphrey's rule and the third factor only had a single sort that loaded significantly in the unrotated matrix. We tried extracting and rotating the three factors and it resulted in nine sorts loading onto factor 3 but these sorts also loaded moderately well onto factor 1. Examination of the content of Factor 3 found it to be very similar to the content of factor 1 and we determined it was not a distinct point of view. A 2 factor solution was chosen because two interpretable impressions were created. When the factors were rotated, Factor 1 grouped 28 participants and Factor 2 grouped 6 participants

Table 2 continued

Consensus between points of view	The only aspects the participants agreed on was that the resident carried out the welcome and closing with courtesy (#44; +3, +2) using body language that demonstrated interest (#5; −4, −3) and she was not arrogant or overconfident (#2; −3, −3)
Factor 1 point of view (based on positively correlated Q-sorts)	*Compassionate competence* She showed genuine concern and appeared to be compassionate (#7; +4) by the way she was willing to take time and listen (#42; +4). She seemed friendly and engaged (#15; +3) and connected well with the patient (#8; +3) by not using a rapid-fire style of questioning (#4; −3) or coming across as dismissive or disrespectful (#12; −4). Her history was fairly thorough, organized and efficient (#41; +2) and she seemed fairly intelligent with good clinical knowledge (#38; +2) although she could have offered some non-pharmacological options (#20; +2)
Factor 2 point of view (based on positively correlated Q-sorts)	*Incomplete attempt* She jumped to treatment too quickly (#22; +4) and did not offer non-pharmacological options (#20; +4). She did not fully investigate the chief complaint (#11; +3) and although she asked relevant questions (#26; −3) she did not ask enough of them so important details about the patient's life were missing (#10; +3). Her history was not thorough or organized or efficient (#41; -2) and even though she used a somewhat checklist style of questioning (#6; +2) questions were asked but not followed up on (#3; +2). She anchored to a diagnosis too early (#1; +3) and did not work through all the differentials systematically (#40; −4). She did not connect well with the patient (#8; −2) since she was more focused on the task than on him (#14; +2)
Demographics	The Q sorts were submitted by 34 participants (16 male, 18 female; 18 internal medicine, 9 emergency medicine, 6 family medicine). Factor 1: 28 participants (13 male, 15 female) from 15 different cities; 16 internal medicine, 6 emergency medicine and 6 family medicine. Factor 2: 6 participants (3 male, 3 female) from 5 different cities; 3 internal medicine, 3 emergency medicine and none from family medicine
Video 6	
Summary of clinical performance	Male resident interviews male patient regarding back pain and provides clinical management: role-played to demonstrate unsatisfactory medical expertise and unsatisfactory interpersonal skills
Factor extraction and rotation	We noted 3 factors had eigenvalues greater than 1 (18.84, 2.33, 1.11) and all three had two or more significantly loading sorts but there were only two factors that barely met Humphrey's rule. A 3 factor solution was investigated and we discovered three sorts were confounded across the factors (i.e. did not load onto any factor). In addition, Factor 3 was a bipolar factor meaning that there were sorts that were both positively and negatively correlated with it. In fact, there was only a single sort that loaded onto each pole of the third factor. The content of the positively correlated sort was not sufficiently different from factor 1 to justify extracting Factor 3. A 2 factor solution was chosen because all the sorts loaded onto a factor and three interpretable points of view were created. The rotated 2 factor solution grouped 17 participants into Factor 1, 14 participants had sorts that positively correlated with Factor 2, and 1 participant's sort negatively correlated with Factor 2

Table 2 continued

Consensus between points of view	Factor 1 and the participants whose sorts were positively correlated with Factor 2 agreed that this resident anchored to a diagnosis too early (#1; +3, +4) and jumped too quickly to treatment (#22; +3, +3) without fully investigating the chief complaint (#11; +4, +4)
Factor 1 point of view (based on positively correlated Q-sorts)	***I**ncompetent* He did not perform a thorough, organized or efficient history (#41; −4) as he failed to inquire about the red flag symptoms (13; +4) and did not systematically work through all the differential diagnoses (#40; −4). He also omitted relevant tasks (#35; +3) such as the physical exam and these deficiencies were significant enough that there was some concern he could be sued in the future (#17; +2)
Factor 2 point of view (based on positively correlated Q-sorts)	*No rapport* He did not connect with the patient (#8; −4) due to a lack of compassion and no display of genuine concern (#7; −4). He came across as dismissive and disinterested (#12; +3) and not willing to take the time to listen (#42; −3) as well as somewhat detached and distant (#9; +2)
Factor 2 point of view (based on negatively correlated Q-sort)	*Friendly competent* The participant's point of view whose Q-sort negatively correlated with the Factor 2 could be represented as the opposite point of view of 'no rapport' described above. The point of view according to the participant's actual Q-sort and comments is that this resident should be fine in the future (#44; +1) because he started with broad differentials (#1; −4) and took his time with the questions (#42; +3) so that he knew the diagnosis (#29; +1) by the end of the interview. He showed genuine concern (#7; +4), was friendly and engaged (#15; +2) and not dismissive or disrespectful (#12; −3)
Demographics	The Q sorts were submitted by 32 participants (17 male 15 female; 17 internal medicine, 7 emergency medicine, 8 family medicine). Factor 1: 17 participants (9 male, 8 female) from 11 different cities; 10 internal medicine, 5 from emergency medicine and 2 from family medicine. Factor 2 (positively correlated): 14 participants (8 male, 6 female) from 12 different cities; 7 internal medicine, 1 emergency medicine and 6 family medicine. Factor 2 (negatively correlated): 1 female participant from emergency medicine with <5 years of extensive involvement with assessing residents
Video 7	
Summary of clinical performance	Male resident interviews male patient regarding chest pain: depicts a second year medical student practising history-taking skills with a standardized patient; demonstrates unsatisfactory medical expertise and superior interpersonal skills
Factor extraction and rotation	We noted 2 factors had eigenvalues greater than 1 (13.06 and 5.46), there were 2 factors that exceeded Humphrey's rule and 16 sorts significantly loaded onto Factor 2. Therefore, a 2 factor solution was chosen which when rotated grouped 17 participants into factor 1 and 13 into factor 2
Consensus between points of view	Participants in both factors agreed he was friendly and engaged (#15; +3, +3) with body language that demonstrated interest (#5; −3, −3) and not at all dismissive or disrespectful (#12; −3, −4) or arrogant or overconfident (#2; −3, −4)

Table 2 continued

Factor 1 point of view (based on positively correlated Q-sorts)	*Inexperienced incompetence* He did not fully investigate the chief complaint (#11; +4) because his history was unorganized, inefficient and not thorough (#41; −4), he did not systematically consider all the differentials (#40; −4) and had no idea what the diagnosis was (#29; +4). He asked many questions that were not relevant (#26; +2) while omitting relevant tasks (#35; +2) such as inquiring about red flag symptoms (#13; +3). He seemed inexperienced with this condition (#19; +3) and without solid clinical knowledge (#38; −2)
Factor 2 point of view (based on positively correlated Q-sorts)	*Competent rapport* He performed a thorough, organized and efficient history (#41; +2) while connecting well with the patient (#8; +4) by being courteous with the welcome and closing (#43; +4), appearing compassionate and showing genuine concern (#7; +3) and being willing to listen without hurrying (#42; +3). He came across as diligent (#24; −3), reliable and trustworthy (#33; +2)
Demographics	The Q sorts were submitted by 30 participants (16 male, 14 female; 15 internal medicine, 10 emergency medicine, 5 family medicine). Factor 1: 17 participants (9 male, 8 female) from 10 different cities; 8 internal medicine, 6 emergency medicine and 3 family medicine. Factor 2: 13 participants (7 male, 6 female) from 9 different cities; 7 internal medicine, 4 emergency medicine and 2 family medicine

connected well) and negative valences being associated with negative statements (disinterest, dismissive, detached, rapid fire questioning). Further, although the medical expertise statements were clearly less salient (mostly 0 to ±2), again the positive valences were generally associated with largely positive statements about the candidate (good history, solid clinical knowledge, systematically considering differentials) and negative valences with problematic statements (no idea of diagnosis, ill-prepared for visit). This produces an overall impression of "compassionate competence" (as labeled and described in Table 2, Video 2, "Factor 1 point of view").

By contrast, in the second Q-factor for Video 2 (fourth column of numbers in Table 1), the rapport building statements show relatively small values (mostly 0 to ±2) suggesting relatively less salience of these items for physicians associated with this factor, and the higher values (±3 or ±4) are found amongst the medical expertise statements. Further, there is less consistent association of positive valences with positive statements in the rapport building items (connected well with patients = −2) and more negative associations in the medical expertise statements (systematic consideration of differentials = −4, incomplete investigation = +3, etc.). Thus, the overall impression for this group of physician raters might be interpreted as "incomplete attempt" (as labeled and described in Table 2, Video 2, "Factor 2 point of view").

Similar interpretations can be found for each of the various perspectives on each video in Table 2. The distinction between the two perspectives was more striking for some videos (e.g. video 2 and 7) and less so for others (e.g. video 1 and positively correlated points of view for video 6). For example, there is much more agreement across the points of view for Video 1 in emphasizing deficient rapport building. However, for all four clinical performances the content of the assessment judgments was sufficiently different as to be identifiable as distinct points of view despite some instances of agreement between

them. Examination of physician membership within these points of view revealed that membership was *not* stable across the four performances (i.e. it was not the same group of physicians emphasizing rapport-building skills over medical expertise every time) and membership could not be attributed to demographic factors (as shown in Table 2).

Identifying divergences between clusters of consensus by comparison across Q-factors

Further examination of the factor arrays for different points of view also revealed possible indications of disagreement in the interpretation of certain aspects of the performance. For example, again referring to Table 1, for videos 2 and 7, the two Q-factor configurations assigned statement #41 to opposite sides of the distribution. This pattern can be interpreted as reflecting clear disagreement among the two groups of raters about whether "thorough, organized and efficient history" was consistent with or contrary to their impression of the same performance. A more prominent example of disagreement is given in Table 2 for video 6. The point of view negatively correlated with Factor 2 and summarized as 'friendly competent' represents the opposite assessment judgment of the point of view positively correlated with Factor 2 and summarized as 'no rapport'. However, it is worth noting that the number of clear disagreements across factors is small and that the 'opposite' perspective seen in video 6 included only a single rater.

Overall, it appears that different assessment impressions of a given clinical encounter can include similar interpretations of many of the performance features and yet the collated assessment judgments can diverge due to physicians differently weighting and sometimes disagreeing on the interpretation of a few performance features. The resulting set of points of view can then be understood to represent conflicting rater judgments of rapport-building and/or medical expertise skills for a single performance.

Identifying the relationship between Q-factors and Mini-CEX ratings

As shown in Table 3, physicians belonging to different points of view were associated with significantly different Mini-CEX ratings for all four videos. The mean ratings for the points of view for each video differed in a direction consistent with the content of the point of view. For example, the points of view emphasizing a greater number of deficiencies in the clinical performance (such as factor 1 for videos 1 and 7 and factor 2 for video 2) were associated with lower mean ratings. Accounting for the two points of view for videos 1, 2 and 7 explained 35–53 % of variance in the Mini-CEX ratings. For video 6, the Q-sort that was negatively correlated with Factor 2 had a rating that was significantly higher than the average ratings of the other two Q-sorts (the group associated with Factor 1 and the group *positively* associated with Factor 2), with the three perspectives accounting for 21 % of Mini-CEX rating variance.

Discussion

Medical education researchers studying inter-rater variability by investigating rater cognition have used open response formats to collect raters' assessment judgments (Chahine et al. 2016; Gauthier et al. 2016; Gingerich et al. 2014b; Govaerts et al. 2013; Herbers et al. 1989; Kogan et al. 2011; Mazor et al. 2007; St-Onge et al. 2016; Tavares et al. 2016;

Table 3 Variance explained in overall clinical competence Mini-CEX ratings by accounting for Q-factors

Video	Q-factors[a]	Number of Q-sorts[b]	Mean (sd)[c]	Standard error of EMM[e]	Partial eta squared (%)	F value (*p* value)
1	Factor 1 'disinterested'	18	3.2 (.94)	.23	39.5	19.6 (.000)
	Factor 2 'diligent competence'	14	4.8 (1.1)	.27		
2	Factor 1 'compassionate competence'	28	6.8 (1.0)	.18	35.4	17.6 (.000)
	Factor 2 'incomplete attempt'	6	5.0 (.63)	.39		
6	Factor 1 'incompetent'	17	2.4 (1.1)	.23	21.3	3.9 (.03)
	Factor 2 (positive) 'no rapport'	14	2.4 (.75)	.25		
	Factor 2 (negative) 'friendly competent'	1	5.0[d] (na)	.93		
7	Factor 1 'inexperienced incompetence'	17	3.7 (1.2)	.29	53.0	31.5 (.000)
	Factor 2 'competent rapport'	13	6.2 (1.1)	.33		

[a] Q-factors were entered as independent variables in one-way ANOVAs using overall clinical competence Mini-CEX ratings as the dependent variable, conducted separately for each video

[b] Participants were assigned to the Q-factor on which their Q-sort was most highly loaded

[c] Mean and standard deviation

[d] The single overall clinical competence rating corresponding with this Q-sort configuration

[e] Standard error of estimated marginal mean

Tweed and Ingham 2010; Yeates et al. 2013, 2015). Since the flexibility of the open response format might have amplified the variability between participants, we sought to triangulate the findings by collecting raters' judgments using a structured response format. We used Q methodology, choosing to have participants sort the same 44 statements describing salient features of performances (previously collected in (Gingerich et al. 2014b) using three different open question formats) for each of the four performances. This design could have easily resulted in each performance being interpreted from a single point of view by all raters, or in a set of completely idiosyncratic points of view unique to each participant. However, neither of these two possibilities emerged. Instead, two or three distinct points of view were identified for each clinical performance. Physicians' membership in the two or three different points of view for each performance could not be attributed to their medical specialty, gender, geography or experience with assessing residents despite the inclusion of participants with varied demographic backgrounds.

Consensus and divergence of multiple points of view

Identifying more than one point of view for a given clinical performance replicates the Gingerich et al. (2014b) finding of multiple physicians sharing one of a limited set of distinct impressions for a clinical performance. Examination of the different points of view indicates that physicians differently emphasized a few aspects of the performance within their assessment impression and rarely outright disagreed on a given aspect. This adds support to previous medical education research findings of differential salience and rater

disagreement regarding a single clinical encounter (Govaerts et al. 2013; Herbers et al. 1989; Kogan et al. 2011; Mazor et al. 2007; Yeates et al. 2013). The variations in the sorting configurations are unlikely to be spurious since 21–53 % of variance in the Mini-CEX ratings could be explained when physician's membership in a Q-factor was accounted for. The consistency of finding multiple clusters of consensus within raters' responses for a given clinical performance across two samples of participants and using two different methodologies challenges the assumption that inter-rater variability is simply measurement error.

It is also worth noting that contrary to our original theorizing (Gingerich et al. 2011, 2014b) these data suggest that physician raters do not see social judgments (such as inferences about intelligence, laziness, or arrogance) as particularly salient aspects of their impressions of the performance. When statements containing such social judgments were put head-to-head with statements containing inferences and judgments regarding clinical skills, physicians did not appear to find them compelling and generally relegated them to positions of 0 or ± 1 on the scoring sheet. It could be that participants were able to avoid forming or being influenced by social judgments while making assessment judgments. However, it is noteworthy that these statements were generated by a previous cohort of physician raters watching these videos. Moreover, due to our use of a self-report design feature we cannot determine the extent to which this discounting of social inferences reflects socially desirable responses or if any unconscious biases influenced the responses. Thus, further research using additional triangulating methods will be needed to rule-out social judgments as a significant source of inter-rater variability.

Re-conceptualizing inter-rater variability and rater cognition

The identified points of view for each performance reflect differing assessment judgments of skill in rapport-building and medical expertise. The identification of two factors underlying performance assessment ratings is consistent with prior medical education research (Chahine et al. 2016; Nasca et al. 2002; Ramsey and Wenrich 1993; Silber et al. 2004; Verhulst et al. 1986). It also aligns well with the two-dimensional theories of social categorization which posit social judgments are made based on judgments of sociability/morality versus competence/ability (Beauvois and Dubois 2009; Fiske et al. 2007; Wojciszke 2005). Although physicians were asked to provide their assessment impressions using Q-sorts and not rating scales, the resulting points of view seem to represent differential judgments on these two underlying dimensions. If so, inter-rater variability could be conceptualized as differential emphasis on rapport-building and/or medical expertise rather than idiosyncratic rater variations. Likewise, rater cognition could be conceptualized as the formation and combination of two judgments: was what needed to be done sufficiently done and was it done while building an alliance with the patient. The fact that individual raters did not consistently emphasize one or the other suggests that there is some cue in the rater-candidate interaction that makes one or the other aspect of performance more salient for a given rater at a given time.

Limitations and areas requiring further investigation

While this work adds to our understanding of inter-rater variation at a time when rater cognition researchers are divided on the value of variability in raters' judgments (Gauthier et al. 2016; Gingerich et al. 2014a; St-Onge et al. 2016), important questions remain. Since we utilized previously studied clinical performances in this investigation, the prevalence of

differing points of view for clinical performances is unknown and could potentially be confined to this set of videos. Each video depicted a different trainee and we have yet to examine the stability or predictability of raters' perspectives for a given trainee across different clinical encounters. In addition, the relative legitimacy of assessment information provided in one point of view compared to another for the same performance has not yet been determined. Finding stability in raters' points of view for trainees' performances on different occasions would have promising implications for the refinement of measurement models.

We also note that Q methodology is only one of several methodological options that we might have used to examine whether or not structured response techniques help to clarify the research questions asked. Other options include the regression-based methods used in decision-making and judgment task research, such as policy capturing or conjoint analysis (Aiman-Smith et al. 2002). Given our research context, there were several potential limitations in using these alternative approaches. First, these methods require a large number of objects to be rated by each rater as well as intentional manipulation of the variables of potential interest across the objects being rated (Karren and Barringer 2002). It was not feasible do so in the context of this study both for logistical reasons and because we were seeking to mimic the design of the Gingerich et al. (2014b) study. Second, because each dimension of interest is intentionally manipulated and/or explicitly presented, it is assumed that all characteristics are equally salient and it is merely weighting that affects the decision process (Gibson and Hobson 1983). The alternative approaches would seem to assume that a particular rater's weighting policy is constant across all objects of measurement (in order to obtain a predictive model for each person about which dimensions are important to that person in general) (Brehmer and Brehmer 1988). We did not wish to adhere to the assumption of roughly equal salience of dimensions across raters nor the assumption of a constant policy across objects of measurement. Thus, a more naturalistic model that allowed us to understand physician raters' variable interpretations of a single performance was necessary. It is for these reasons that we selected Q methodology; although this choice does limit our ability to compare our findings to what might have been found using policy capturing techniques, the benefits of Q methodology outweighed using that technique.

Implications of this research

Regardless of the relative accuracy of the points of views or the actual cognitive processes involved with forming differing points of view for the same clinical encounter, the finding of multiple interpretations that cannot be easily reconciled into a single judgment is problematic for the analysis of ratings. Most critically, this study provides preliminary evidence that raters are not interchangeable. This would violate the homogeneity assumption and result in excess variance being attributed to the raters in psychometric measurement models (Kane 2002). If raters could be expected to report different assessment judgments or assign different ratings because they belonged to one of multiple unknown points of view, our current measurement models would be inefficient in extracting and summarizing the relevant assessment information. If subsequent research confirms multiple shared interpretations of the same encounter are to be expected and that one is unable to predict which population raters will belong to for any given encounter, then we may need to search for alternate assessment designs that treat inter-rater variation as something more meaningful and informative than idiosyncratic rater variance.

Acknowledgments The authors wish to thank everyone who assisted with recruiting participants and especially those who took the time to participate in this study. We also wish to thank Rick Hoodenpyle for designing the online data collection system and hosting it on QSortOnline.

Funding This study was funded by a National Board of Medical Examiners® (NBME®) Edward J. Stemmler, MD Medical Education Research Fund Grant. The project does not necessarily reflect NBME policy, and NBME support provides no official endorsement.

References

Aiman-Smith, L., Scullen, S. E., & Barr, S. H. (2002). Conducting studies of decision making in organizational contexts: A tutorial for policy-capturing and other regression-based techniques. *Organizational Research Methods, 5*(4), 388–414.

Albanese, M. (2000). Challenges in using rater judgements in medical education. *Journal of Evaluation in Clinical Practice, 6*(3), 305–319.

Beauvois, J.-L. O., & Dubois, N. (2009). Lay psychology and the social value of persons. *Social and Personality Psychology Compass, 3*(6), 1082–1095.

Brehmer, A., & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing? *Advances in psychology, 54*, 75–114.

Brown, S. R. (1980). *Political subjectivity: Applications of Q methodology in political science*. New Haven: Yale University Press.

Chahine, S., Holmes, B., & Kowalewski, Z. (2016). In the minds of OSCE examiners: Uncovering hidden assumptions. *Advances in Health Sciences Education, 21*(3), 609–625.

Crossley, J., Davies, H., Humphris, G., & Jolly, B. (2002). Generalisability: A key to unlock professional assessment. *Medical Education, 36*(10), 972–978.

Crossley, J., & Jolly, B. (2012). Making sense of work-based assessment: Ask the right questions, in the right way, about the right things, of the right people. *Medical Education, 46*(1), 28–37.

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38*(9), 1006–1012.

Downing, S. M. (2005). Threats to the validity of clinical teaching assessments: What about rater error? *Medical Education, 39*(4), 353–355.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83.

Gauthier, G., St-Onge, C., & Tavares, W. (2016). Rater cognition: Review and integration of research findings. *Medical Education, 50*(5), 511–522.

Gibson, C. J., & Hobson, F. W. (1983). Policy capturing as an approach to understanding and improving performance appraisal: A review of the literature. *Academy of Management Review, 8*(4), 640–649.

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014a). Seeing the 'black box' differently: Assessor cognition from three research perspectives. *Medical Education, 48*(11), 1055–1068.

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-based assessments as social judgments: Rethinking the etiology of rater errors. *Academic Medicine, 86*(10), S1–S7.

Gingerich, A., van der Vleuten, C. P. M., Eva, K. W., & Regehr, G. (2014b). More consensus than idiosyncrasy: Categorizing social judgments to examine variability in Mini-CEX ratings. *Academic Medicine, 89*(11), 1510–1519.

Govaerts, M. J. B., Wiel, M. W. J., Schuwirth, L. W. T., van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2013). Workplace-based assessment: Raters' performance theories and constructs. *Advances in Health Sciences Education, 18*(3), 375–396.

Herbers, J. E., Jr., Noel, G. L., Cooper, G. S., Harvey, J., Pangaro, L. N., & Weaver, M. J. (1989). How accurate are faculty evaluations of clinical competence? *Journal of General Internal Medicine, 4*(3), 202–208.

Kane, M. (2002). Inferences about variance components and reliability-generalizability coefficients in the absence of random sampling. *Journal of Educational Measurement, 39*, 165–181.

Karren, R. J., & Barringer, M. W. (2002). A review and analysis of the policy-capturing methodology in organizational research: Guidelines for research and practice. *Organizational Research Methods, 5*(4), 337–361.

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: A conceptual model. *Medical Education, 45*(10), 1048–1060.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology, 51*(1), 93–120.

Mazor, K. M., Zanetti, M. L., Alper, E. J., Hatem, D., Barrett, S. V., Meterko, V., et al. (2007). Assessing professionalism in the context of an objective structured clinical examination: An in-depth study of the rating process. *Medical Education, 41*(4), 331–340.

McKeown, B. F., & Thomas, D. B. (1988). *Q methodology (quantitative applications in the social sciences series* (Vol. 66). Thousand Oaks, CA: Sage.

Mohr, C. D., & Kenny, D. A. (2006). The how and why of disagreement among perceivers: An exploration of person models. *Journal of Experimental Social Psychology, 42*(3), 337–349.

Nasca, T. J., Gonnella, J. S., Hojat, M., Veloski, J., Erdmann, J. B., Robeson, M., et al. (2002). Conceptualization and measurement of clinical competence of residents: A brief rating form and its psychometric properties. *Medical Teacher, 24*(3), 299–303.

Newman, I., & Ramlo, S. (2010). Using Q methodology and Q factor analysis in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social and behavioral research* (2nd ed., pp. 505–530). London: Sage Publications.

Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The Mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine, 138*(6), 476.

O'Neill, T. A., McLarnon, M. J., & Carswell, J. J. (2015). Variance components of job performance ratings. *Human Performance, 28*(1), 66–91.

Park, B., DeKay, M. L., & Kraus, S. (1994). Aggregating social behavior into person models: Perceiver-induced consistency. *Journal of Personality and Social Psychology, 66*(3), 437–459.

Ramsey, P. G., & Wenrich, M. D. (1993). Use of peer ratings to evaluate physician performance. *JAMA: Journal of the American Medical Association, 269*(13), 1655–1660.

Schmolck, P. (2014). *PQMethod* 2.35, software adapted from Mainframe-Program QMethod by John Atkinson at Kent State University. http://schmolck.userweb.mwn.de/qmethod/.

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher, 33*(6), 478–485.

Silber, C. G., Nasca, T. J., Paskin, D. L., Eiger, G., Robeson, M., & Veloski, J. J. (2004). Do global rating forms enable program directors to assess the ACGME competencies? *Academic Medicine, 79*(6), 549–556.

Stephenson, W. (1953). *The study of behavior: Q-technique and its methodology*. Chicago, IL: University of Chicago Press.

St-Onge, C., Chamberland, M., Lévesque, A., & Varpio, L. (2016). Expectations, observations, and the cognitive processes that bind them: Expert assessment of examinee performance. *Advances in Health Sciences Education, 21*(3), 627–642.

Tavares, W., & Eva, K. W. (2013). Exploring the impact of mental workload on rater-based assessments. *Advances in Health Sciences Education, 18*(2), 291–303.

Tavares, W., Ginsburg, S., & Eva, K. W. (2016). Selecting and simplifying: Rater performance and behavior when considering multiple competencies. *Teaching and Learning in Medicine, 28*(1), 41–51.

Tweed, M., & Ingham, C. (2010). Observed consultation: Confidence and accuracy of assessors. *Advances in Health Sciences Education, 15*(1), 31–43.

Verhulst, S. J., Colliver, J. A., Paiva, R., & Williams, R. G. (1986). A factor analysis study of performance of first-year residents. *Academic Medicine, 61*(2), 132–134.

Watts, S., & Stenner, P. (2012). *Doing Q methodological research: Theory, method and interpretation*. Thousand Oaks, CA: Sage.

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine, 15*(4), 270–292.

Wojciszke, B. (2005). Affective concomitants of information on morality and competence. *European Psychologist, 10*(1), 60–70.

Wood, T. (2014). Exploring the role of first impressions in rater-based assessments. *Advances in Health Sciences Education, 19*(3), 409–427.

Yeates, P., Cardell, J., Byrne, G., & Eva, K. W. (2015). Relatively speaking: Contrast effects influence assessors' scores and narrative feedback. *Medical Education, 49*(9), 909–919.

Yeates, P., O'Neill, P., Mann, K., & Eva, K. W. (2013). Seeing the same thing differently: Mechanisms that contribute to assessor differences in directly-observed performance assessments. *Advances in Health Sciences Education, 18*(3), 325–341.