Questioning the rater idiosyncrasy explanation for error variance by searching for multiple signals within the noise

Summary

CHAPTER 1

The program of research described in this dissertation was motivated by a desire to improve the utility of rater-based assessments. It germinated out of a curiosity of how multiple people could observe the same performance and make very different interpretations, judgments and assessments of that performance. We started to look for explanations for why it is so difficult to support raters to make consistent assessment judgments. This questioning led to the need to better understand how raters make judgments and prompted a journey into the psychology literatures to learn more about human judgment in social situations.

We searched for comparable circumstances where variability in judgments for the same performance had been studied and found an extensive research collection in the field of impression formation. In this literature it had been discovered that people form different impressions of the same target person but they were not idiosyncratic impressions. Instead, multiple people formed one of limited number of impressions for a person. This was intriguing because there was consensus of opinion but not one global point of consensus. Given the way our rater-based assessments are designed and analyzed, if something similar was occurring for clinical performance assessments, then these multiple points of consensus would be interpreted as measurement error. Hence, we were motivated to study the contribution of physicians' judgments to inter-rater variability for workplace-based assessments.

CHAPTER 2

This paper was written after Chapter 3 and it gives a broader review of the literature pertaining to inter-rater variation. It developed out of conversations with an international group of researchers sharing a common interest in rater cognition. Given the similar themes among our recent publications, Dr. Eric Holmboe had encouraged the group to meet, share ideas and discuss ways in which we could collaborate. During these discussions it became apparent that we were all thinking about the causes of inter-rater variation, and subsequently the solutions for it, in very different ways. The literature review is a result of us defining, comparing and contrasting our different perspectives in understanding variability in ratings and judgments. The three resulting perspectives are not mutually exclusive but they do represent some incompatible ideas. In the first two perspectives, variability is seen as the result of raters being under-trained or due to inherent limitations of human cognition. Either way, variability is not ideal, and as a result, is something to be minimized or compensated for. The third perspective better represents the philosophy of this dissertation in that variability is seen as potentially resulting from informed differences of opinion by experts judging a

complex social interaction. It is this paper that provides the context in which our program of research exists; how it aligns with, and is differentiated from, other rater cognition investigations.

CHAPTER 3

This is the paper that marked the beginning of our program of research. It summarizes our understanding of the research investigating how social categorization and social judgment processes contribute to variability in the impressions we form of others. It is drawn from an expansive collection of literatures broadly considered social psychology but more specifically referred to as the domain of social cognition. Within the numerous research studies that were reviewed, there were findings of consensus and agreement amidst the very subjective and potentially idiosyncratic social judgments. We describe three overarching theories of social categorization that each propose different mechanisms for how impressions are formed. We, as humans, must have evolved to benefit from our skills in observation, perception, judgment and decision-making during social interactions. However, these processes may not align well with the tasks asked of raters in rater-based assessments and contribute to variability in ratings. These literatures provided us with the theoretical and methodological support to begin investigating rater cognition in clinical performance assessments.

CHAPTER 4

This article represents a proof of concept as we conceptually replicated a social cognition study by translating it into the medical education context. It also introduces the methodology of latent partition analysis used to search for multiple clusters of consensus within the opinions offered by physicians. Consistent with the social cognition literature, we were able to identify more than one impression, each described by multiple physicians, for every trainee. Despite the possibility for each physician to have provided a unique social impression of the trainee, we identified as few as two and no more than five distinct impressions for each trainee. The content across the set of impressions for a given trainee was not merely different but often contained conflicting judgments. We found that physicians describing similar social judgments also assigned more similar ratings. The findings suggested there may be multiple signals within the noise of inter-rater variability and set the way for further investigations.

CHAPTER 5

Having established a methodology for identifying clusters of consensus within physicians' comments regarding a clinical performance, we were able to extend the investi-

gations to include all three theories of social categorization described in Chapter 3. In comparing their relative ability to explain variance in clinical performance ratings, we discovered each could account for significant rating variance. That is, each theory of categorization could be used to group physicians' comments pertaining to a single performance into 2-4 distinct categories. When we examined the content of the predictive categories, we observed that, consistent with other medical education research findings, the differences in raters' impressions reflected disagreement regarding the same aspect of performance as well as emphasis on different aspects of a single clinical performance. We had identified clusters of consensus that appeared to be meaningfully different interpretations of a single performance but the findings were exploratory and needed to be further tested.

CHAPTER 6

In this article we describe our efforts to triangulate the clusters of consensus finding using a new set of participants and a different methodology. Once again we find two or three different points of view for the same clinical performance, that when accounted for, explain a substantial proportion of variance. The content of the points of view was consistent with the content of the previously identified clusters of consensus and it confirmed that physicians can have opposing judgments of the same aspect of the performance. The consistency of these findings prompts us to seriously consider that we may need to conceptualize physicians functioning as constructivist raters rather than post-positivist assessment instruments.

CHAPTER 7

In the final chapter we further consider the implications of finding clusters of consensus within variable assessment judgments. Multiple points of consensus may reveal competence could be better conceptualized as having multiple signals or "truths". If so, our measurement models will need to be interpreted in terms of multiple "true scores". This finding presents challenges for designing an assessment system that can analyze variable and subjective assessment judgments to contribute to robust decision-making regarding trainee competence. We discuss possible non-psychometric approaches for systematic analysis of physicians' judgments regarding trainee competence in terms of a fictitious assessment system. Further investigations are needed to more fully understand what clusters of consensus represent and what the implications may be for rater-based assessments.