What's in a Number? Problematizing Interpretations of Observation Rubrics

Mark White, University of Oslo

**Abstract** (300 words). Recently, the MET, the TALIS GtI, and the LISA studies (among others) have conducted large-scale observations of naturally occurring classroom instruction. The same conclusion was reached in each study: the quality of teaching varies systematically across instructional domains, being highest for classroom management, moderately high for classroom culture, and lowest for instructionally focused dimensions. This paper problematizes this conclusion, arguing that scores from observation systems cannot support it. These studies draw this conclusion by comparing scores from observation systems across measured dimensions or domains. However, there is a no guarantee that a score of a three represents the same level of quality across dimensions of observation systems. This is only true if dimensions are equated and observation systems have generally not equated scores across dimensions. I further problematize the implications of this finding. The general implication seems to be that there is a need to improve the quality of the aspects of instruction that receive the lowest score. This, however, does not follow from the findings. Conceptually, the best target of intervention is the dimension of teaching that can most easily be changed and whose local change leads to the largest shifts in outcomes. Neither of these criteria are necessarily connected to the average score on a dimension. Overall, then, this paper problematizes interpretations of scores from observation systems. In doing so, I highlight the importance of carefully considering what is being measured and how theory links the construct of teaching quality to the observed measures of teaching quality.

**Extended summary** (1000 words, excluding reference list) Include introduction, theoretical background, methods, aims, preliminary findings/findings, results, theoretical and education significance, relevance to the QUINT ambition and the reference list.

**Introduction**: Many studies of teaching quality have recently sought to characterize the nature of naturally occurring teaching across the world (e.g., Kane et al., 2012; Klette et al., 2017; OECD, 2020; Tengberg, et al., 2020). These studies routinely use observation rubrics, which code the quality of teaching into qualitatively distinct ordinal categories across a number of dimensions. These scores are then combined and summarized to provide broad characterizations of teaching quality. This process of transforming ordinal rubric scores into numerical summaries of teaching contains a large number of implicit assumptions. Ignoring these assumptions has led researchers to make broad characterizations of what teaching looks like that cannot be supported by the empirical evidence collected by researchers.

**Theoretical Background**: The demands of constructing a rubric leads rubric designs to focus on identifying the range of observable behaviors that are relevant for each dimension of teaching quality and then to divide those behaviors into a small set of qualitatively distinct levels of teaching quality. This process is

**Methods**: The methods are a theoretically and logical analyses of the nature of converting rubric scores into summaries that are meant to describe teaching.

**Aims**: The goal of this paper is to problematize common interpretations of observation scores and the associated characterizations of what teaching looks like in practice.

**Results**: I show that there is not rational basis to compare observation scores across dimensions in order to identify where teaching quality is relatively high and where it is relatively low. Further, the within-

dimension comparison of observed teaching quality is often equally problematic, due to the fairly weak conceptualizations of aspects of teaching quality connected to instructional support (i.e., the way teachers support students interacting with content and each other). Further, even if one could make these comparisons and identify where teaching quality is lower and where it is higher, the implicit assumption that we should intervene first where teaching quality is lowest does not follow from this point.

**Theoretical and Educational Significance**: This paper problematizes the way that researchers use empirical evidence to characterize the nature of teaching. It helps to push researchers towards more valid and defensible characterizations of teaching. It also provides guidance for research designs and improvements to observation systems that can move the field to more theoretically justifiable and meaningful interpretations of teaching quality.

**Relevance to QUINT Ambition**: This gets at the heart of the QUINT ambition by discussing issues around measuring and characterizing teaching quality.