Demand Characteristics and Quality Improvement: Who is Fooling Whom?

Abstract

Since some of the data used for quality assurance purposes (i.e. rating scales) requires the active participation of clinicians, administrators use various mandates or incentives to insure its collection. However, although improving clinician compliance these measures may bias clinician responses. It is suggested that the concept of 'demand characteristics' originally developed by Orne to describe how non-specific aspects of the experimental setting shape what the human subject does may be applicable. For example a measure that might increase clinician compliance with completing GAF ratings on an inpatient unit might also influence the scores to make them coincide with the expectation that all patients are very ill when admitted and improved when discharged. Why such a phenomenon would be difficult to detect and what it might say about the relationship between managers and those they manage is also explored.

.

The computerized clinical data that administrators use to assess the quality of mental health services is collected through two very different mechanisms. Some data sets are the byproduct of the computerized processes that are integral to the delivery of clinical care: admission and discharge dates, medication orders, laboratory results, etc. To the extent that clinical processes depend on this data, it should be expected to provide an accurate representation of that component of the clinical enterprise. It may not tell us whether the patient's length of stay was clinically appropriate or the test ordered was necessary, but it will accurately reflect the number of days in hospital or whether a CBC was ordered. Other data sets are not a byproduct of processes essential for delivering care, so their accuracy is not insured by such constraints, e.g. clinical rating scales, etc. Since collecting such data sets depends on the clinicians' participation, administrative mandates and various incentives are employed to increase compliance. If increased compliance was associated with increased accuracy this would improve the validity of the data. But increased compliance does not insure accuracy, on the contrary, it is possible that those measures that are most successful in prodding clinicians to provide data might be most likely to bias their responses.

It could be argued that this distortion would be mitigated by relying on valid and reliable clinical instruments; for example DSM-IV diagnoses or the Global Assessment of Function (GAF, Endicott et al, 1976) that have been shown to have good inter-rater and repeat reliability. However, these reliability studies were carried out in research settings which place high expectations on the accuracy and reliability of raters.   Clinical settings may impose very different expectations. Clinicians might adjust their ratings to be congruent with their clinical decisions or with administrative expectations related to performance improvement, utilization review activities, or the use of these ratings for billing purposes. Expectations may vary by clinical setting. For example, on an inpatient unit, patients would be expected to be rated as highly impaired on admission (otherwise they are not appropriate for admission) and to be far less impaired by the time of discharge (otherwise they are not appropriate for discharge). In an outpatient setting, patient ratings would be expected to fluctuate in a narrower range (severe psychopathology might beg hospitalization while little psychopathology would question why treatment should be continued). Also, different units or facilities might have their own ‘expectation’ of the appropriate rating that would ‘trigger’ admissions, discharges, etc; so admission, discharge and outpatient ratings might differ significantly amongst them(Greenberg and Rosenheck, 2005)

In trying to tease out the degree to which clinical ratings reflect the patients' actual clinical condition as opposed to being distorted by these expectations it may be useful to borrow the concept of 'demand characteristics' introduced by Orne(1962) to describe a somewhat similar phenomenon seen in psychological experiments done with human subjects. Here the problem is differentiating between the effect of the specific experimental variable, what is '***done*** to the subject' as a passive responder to stimuli; as opposed to how non-specific aspects of the 'experimental setting' shape what the human subject ***does.*** For example, a 'demand characteristic' of participating in almost any psychological experiment, is that human subjects believe that they need to be "good subjects" in order to validate the experimental hypothesis. To a researcher this might seem a good thing; it may help them obtain the results they want. But often it is a problem, while a researcher might want to find confirmation of their hypothesis, what they really need to find is the truth. I would suggest that the administrator or policy maker is in a similar situation. When monitoring quality of care indicators, they might want these indicators to improve, but what they really need to improve is care not just indicators.

Orne became aware of this phenomenon when he was trying to determine the differences between a hypnotic and a non-hypnotic state. He tried to find a task that subjects would do while hypnotized but not when not hypnotized. But he found it impossible to develop a task, regardless of how noxious, meaningless, or boring that would be done by subjects under hypnosis but not by waking subjects if told "this is an experiment." Because he recognized this phenomenon he was able to study it. He had some subjects actually hypnotized (experimental group) while others were told to simulate hypnosis (controls). Since trained hypnotist could not differentiate the two groups, Orne concluded that 'demand characteristics' as opposed to an altered conscious state could account for the behavior of the experimental group. An alternative experimental approach was to hold the specific experimental variable constant but alter the demand characteristics. For example some subjects (experimental) were told they were going to participate in a sensory deprivation study while others were told they were to be controls for such a study. After both groups spent four hours in a quiet room (but without real sensory deprivation) the experimental group but not the controls reported a significant increase on measures used in sensory deprivation research, suggesting that demand characteristics explain at least some of the findings commonly attributed to sensory deprivation.

Although administrators might acknowledge that the methods used to insure data collection might influence the data that is collected, the 'demand characteristics' of their own situation may make it difficult for them to recognize it in practice. Take for example GAF scores on inpatient units. Aware that their behavior is being monitored and in keeping with their perception of administrative expectations; 'good clinicians' will be vigilante to record GAF scores that justify their clinical decisions (always low on admission and higher on discharge). Unit administrators are subject to similar expectations; 'good administrators' want their clinicians' GAF ratings to reflect the same pattern as evidence that the unit, and by inference, they are doing a good job. It would not be hard to imagine how confirmation of these expectations, although illusory, might seem compelling. As administrators focus on insuring clinician compliance with GAF ratings requirements, clinical outcomes might appear to improve. When administrative oversight is relaxed, compliance would drop and the GAF scores might no longer be as consistently low on admission and higher on discharge; but when administrators re-focus on collection of inpatient GAF scores, collection rates would improve, and GAF scores might fall into line with expectations. With such an ABA design it would be hard for administration not to conclude that their oversight of GAF scores has improved the quality of care.

Let me summarize what I have been trying to sketch out. Administrators assess clinician performance and provide feedback in order to improve care. In this managerial model the administrator is in charge. But humans don't just have things done to them, they do things back. This is obvious when what they do back doesn't fulfill expectations, but may be more difficult to detect when they do. Since the quality indicators that administrators define and measure represent their operational definitions of quality care, they are delighted when they improve. They may forget that improvement on these measures does not have a one to one relationship to improvement in clinical outcomes (Luchins, 2009). So when clinicians give back to administrators the numbers they want, administrators (perhaps incorrectly) believe that the patients must be getting the care that they need. This willing suspension of disbelief is not unique to mental health. Think about body counts during the Vietnam War (Gibson, 1986)

In my mind this raises the question of who is controlling whom. Administrators, as managers, tend to assume they are in charge; they chose what to measure, control the incentives and collect and feedback these numbers to the clinicians. But these numbers depend on the clinician's behavior and are subject to their agenda. They may be 'playing along' while still not 'playing' the same game. As has been noted by the philosopher Alasdair MacIntyre (1984), in actual social situations, unlike those described by game theory, "No one game is being played… the problem about real life is that moving one's knight to QB3 may also be replied to with a lob across the net"

A wonderful illustration of managerial blindness to this phenomenon comes from a re-analysis (Gillepsie 1991) of the most famous study of industrial production, the Hawthorne Experiments. These studies was carried out in the early 1930s in a General Electric relay assembly plant in a special test room with a small group of immigrant women. In the accepted version, the research demonstrated that almost any intervention (increasing illumination, decreasing illumination) increased the workers’ productivity. This was interpreted to mean that "attention" from management increased production (the Hawthorne Effect). This is how the findings appeared to the researchers. But this is not how it appeared to the workers. They were not blind to the 'demand characteristics' of the study. They knew their productivity was being monitored and responded accordingly. In fact, it has been argued (Adair, 1984) that the "Hawthorne Effect" is not due to 'attention', 'novelty ' or even 'awareness' of being in an experiment, but to the subject's identifying a purpose for the experiment and expectations for their behavior. The workers were also sensitive to a series of factors that the experimenters chose to ignore. The study coincided with the beginning of the Great Depression. Being 'subjects' in the test room, these women continued to have a job while workers elsewhere in the factory (women first) were being laid off. Also, unlike the workers elsewhere in the factory, when these women increased their productivity their piece rate was not reduced and their earnings rose. They had an incentive to increase their productivity, an incentive that was denied to the other workers. The experiments' outcome was determined not just by what management was 'doing to' the workers but by the workers' assessment and response to the situation; "what they did" back. Instead of being manipulated, one could just as well argue that these immigrant women were manipulating the experimental situation to protect their jobs and make more money.

I remember from my undergraduate days a cartoon showing rats in a cage being watched by a couple of scientists in lab coats. The caption was, "We've really got these researchers conditioned. Every time we push the bar they've got to drop a pellet." So, who is fooling whom?

Adair G: The Hawthorne effect: A reconsideration of the methodological artifact. **Journal of Applied Psychology.** 69: 334-345, 1984.

Endicott J Spitzer RL Fleiss J L Cohen J: The Global Assessment Scale: A procedure for measuring overall severity of psychiatric disturbance. **Archives of General Psychiatry.** 33**:** 776-779, 1976.

Gibson JW: The Perfect War: The War We Couldn't Lose and How We Did. New York, Vintage Books, 1986.

Gillespie R: Manufacturing Knowledge: A History of the Hawthorne Experiments. Cambridge, Cambridge University Press, 1981.

Greenberg GA, Rosenheck RA: Using the GAF as a national mental health outcome measure in the Department of Veterans Affairs. **Psychiatry Services**. 56:420-426, 2005.

Luchins DJ: Improving care, improving performance or just improving numbers. **Psychiatric Services**. **59**: 1328-1330, 2008.

MacIntyre A: After Virtue: A study in Moral Theory (second edition). Notre Dame, University of Notre Dame Press, 1984.

Orne, MT: On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. **American Psychologist.** **17:** 776-783, 1962.