Developing Observational Rating Scales

John Whyte, MD, PhD & Tessa Hart, PhD

Introduction

Cognitive rehabilitation research requires measures of the neural and cognitive processes and/or functional domains targeted by the treatment. Measurement of neural and cognitive processes may be particularly relevant for clarifying the mechanism of treatment and the types of patients who can benefit from treatment, and is frequently done using neuroimaging, neurophysiologic, and/or cognitive psychological measures. However, particularly in the more advanced stages of treatment research, one needs to address whether the treatment leads to readily observable and functionally important behavioral changes. Although measuring these real-world treatment effects can sometimes be done through standardized testing or participant self-report, it may be more advantageous to use rating scales that are completed by an external observer of the participant's behavior. Although behavioral rating scales do have disadvantages of their own, they avoid the environmental restriction common to many standardized tests, and some of the biases inherent in self-report measurement.

The process of developing observational rating scales of behavior involves multiple steps. Because validation of a new scale is an iterative process involving different types of research, this process is best carried out in conjunction with a program of research where one anticipates an ongoing need to assess treatment with respect to its impact on observable behavior in the natural environment. Below we describe the steps involved in constructing an observational rating scale, drawing on our own efforts to develop such a scale focusing on observable problems with attention that are frequently seen after traumatic brain injury – the Moss Attention Rating

Scale (MARS). In this discussion we review the steps in sequence, although there may be instances where iteration among steps is necessary based on the results in the subsequent step.

Gathering and Endorsement of Content

An initial step is determining what domain is to be assessed by the scale, its overall scope, and the relevant subdomains to be captured. In the case of the MARS, we knew that the simple word, "attention", is associated with a wide range of interrelated concepts that appear in lay conversation, clinical discussions of individuals with TBI, and basic science discussions of the definition and cognitive architecture of attention and arousal systems. Thus, we generated a list of words and phrases subsumed by "attention" and drawn from lay, clinical, and basic science sources. We then distributed this list of concepts to a set of clinicians and attention researchers and asked them to add any concepts they felt were missing, as well as to comment on existing concepts. Individual content possibilities were then grouped to allow categorization. In this way, although the investigator must exercise judgment about what to ultimately include, and how to collapse similar concepts into larger categories, a systematic attempt is made to gather the full range of relevant concepts for incorporation into the scale.

Item Writing and Revising

The range of concepts relevant to the scale needs to be transformed into specific scale items, and a decision must be made about the type of rating scheme to be used. This is an interactive process, since the type of items chosen will constrain the type of rating that is logical. For example, one might create items that are adjective phrases describing the person being rated, and then rate each item according to how "characteristic" it is of the person in question. Alternatively, one could create items that describe specific behaviors, and then rate the frequency

or intensity of their occurrence. It is desirable to use a single rating scheme for the entire scale whenever possible, but not at the expense of using an insensitive rating method for one subgroup of item content.

In the case of the MARS, we believed that different behaviors relevant to attention would occur with vastly different frequencies (e.g., visually orienting to a stimulus of interest, vs. self-correcting a mistake). Consequently, we chose a format in which "trait-like" statements were made, and the rater had to endorse how characteristic or typical of the individual being rated these were (with responses ranging from "definitely true" to "definitely false" on a 5-point Likert-like scale). Aside from the overarching decision about the measurement scale, the wording of each specific item needs to be decided, based on an attempt to define as clearly as possible in terms of observable behavior, the "evidence" for the particular concept that the item is assessing. For example, if we want to rate "sustained attention" we must decide on the behavioral manifestations of that construct. Is a deficit in sustained attention evidenced by randomly erratic performance on tasks; performance that deteriorates systematically over a session; stopping performance of an ongoing task altogether? One or all of these must be phrased as ratable scale items. Typically at this phase, one will write a surplus of items that tap the various concepts of interest, since subsequent phases of development are likely to result in the elimination of some items.

Qualitative Item Feedback

The process of item refinement can proceed from qualitative to quantitative analysis. Initially, one should assess the clarity of the items and the feasibility of rating them. For the MARS, we began by writing draft items and asked a set of therapists to rate a pilot sample of 10

rehabilitation inpatients. We then convened a focus group of the clinicians who participated in the pilot, showed them the agreement level of the items, and solicited their help in understanding the sources of disagreement and in re-wording, discarding, and/ or replacing poor items. In a few instances, the meaning of items was simply unclear. In more items, there were concerns about how they could be rated in the presence of other confounding conditions. These discussions resulted in a revised scale suitable for more intensive study. A slightly larger pilot study involving 20 patients and multiple therapists revealed agreement substantial enough to advance that draft of the scale into the phase of more rigorous psychometric assessment. Small pilot studies followed by qualitative and quantitative analysis are relatively inexpensive and can go far in producing a final or semi-final version of a scale for larger-sample reliability testing.

Psychometric Evaluation

When more formal psychometric assessment is appropriate, one is typically interested in both test-retest and inter-rater reliability. Both should be assessed on the range of individuals expected to be rated by the measure, and with a sufficient sample size to ensure relatively narrow confidence intervals around the reliability scores obtained. Test-retest reliability should be done without access to the prior rating and with a sufficiently long interval to minimize the rater's memory, but short enough so that stability of the rating would be expected.

Inter-rater reliability should be done with all raters completing the scale as close to one another in time as is feasible, and with the raters blinded to each other's ratings. If there are different categories of individuals who may be using the rating scale (e.g., physical and occupational therapists in the initial work with the MARS) then it is important to explore any systematic differences in agreement by those categories. When the ratings are based purely on

observation, it is reasonable to have the 2 raters observe the same session of behavior as long as their ratings are not discussed. But when the observations must be "elicited" in some way by the rater, then ratings should be obtained in separate sessions since analyses based on 2 observations of one "elicitation" will overestimate the agreement that would be obtained in independent clinical use.

Validation of the scale is a more complex process which depends on the intended use of the scale, and which is likely to proceed in gradual steps. More complete discussions of validation can be found.[e.g., 1] It is particularly helpful to be able to predict opposite or differential relationships between the new scale and multiple existing measures—some that should be positively related to the new scale and others that should be negatively related or unrelated. However, validation at a minimum typically includes comparison to some other measure of conceptual relevance, such as different measure of the same construct (concurrent) or a measure of a different construct that would be expected to covary with the construct of interest. The comparison data set should be obtained independently from the scale ratings. One should keep in mind that conditions with a severity dimension (e.g., traumatic brain injury, size of left hemisphere stroke) may induce severity-driven correlations among unrelated domains. Thus, a correlation between 2 measures alone, may not serve as adequate concurrent validation.

Factor Analysis

Many psychosocial domains have discernible subdomains. For example, the Agitated Behavior Scale includes items that assess aggression, disinhibition, and lability.[2] In the case of the MARS, as mentioned, we believed that attention is a multifaceted construct but could not predict how those facets, which are often defined with respect to specific cognitive processes,

would relate to categories of observable behavior. A principal components analysis can help to reveal underlying interrelationships among items which can then be interpreted in conceptual terms. Such exploratory factor analyses should be independently confirmed in a subsequent sample, both of which require large samples. Orthogonal analyses, which presume that the subcategories are completely uncorrelated may not be realistic in some behavioral domains either because they are intrinsically or mechanistically related, or because they are all subject to a disease severity dimension. In the case of the MARS, for example, three factors (initiation, restlessness/ distractibility, and sustained/ consistent attention), composed of 3, 5, and 3 items (out of the total pool 45 items) were identified with factor intercorrelations ranging from .46 to .75.[3]

Item Spacing and Scoring

The items comprising a rating scale may map onto a single behavioral dimension, or on several distinct dimensions. However, in addition, several items tapping the same dimension may differ in their “difficulty level” (actual difficulty, likeliness to be seen or other factors that affect the overall likelihood of low vs. high ratings for different items) on that dimension. Item response analyses can clarify the dimensionality of the rating scale, as well as the “fit” and location of individual items on the dimensions that are identified. For example, a scale may appear primarily unidimensional, but with several items that don’t fit the dimension and therefore should be discarded. Or two dimensions may be identified with an additional few items that fit neither dimension well. In addition, it may be determined that the bulk of the scale items like at one end of the dimension with sparse sampling of the other end, in which case new items may need to be written, and some of the previous steps repeated. Alternatively, the dimension

may be well covered by the available items with a number of closely spaced items, some of which can be deleted due to redundancy.

In the case of the MARS, we determined that 42 of the original 45 attention items fit a single dimension, whereas 3 items misfit.[4] Inspection of these items revealed clear reasons for the poor fit of 2 of these items ("keeps eyes open even when not directly stimulated" and "reacts to dramatic stimuli such as loud noises, alarms, shouts, etc."), since, in retrospect, they can be true of patients in the vegetative state. The third misfitting item ("tends to speak less than he/she is capable of") appeared to have relevant content but probably was more difficult to rate reliably because it required an assumption about the patient's "capability" independent from his/her actual performance. Nevertheless, this item was retained at this point in the project because of an interest in the subdomain of initiation (this item was later dropped because it was redundant with other items – see below). Note that in the case of the MARS, item response analysis supported unidimensionality of the scale while principal components analysis suggested the presence of 3 correlated factors. These are not mutually exclusive conclusions but presumably indicate that the factors (which are, after all, correlated) are more tightly interrelated clusters of items within a broader dimension.

Final Scale Construction

In the above steps, one may be able to shorten the scale to the minimum number of items that adequately covers the dimension(s) of the scale and any factors of interest. In the case of the MARS, we began with 45 attentional items that had adequate fit to the attention dimension. We then deleted the 3 items that fit poorly based on Rasch analysis, leaving 42 items. We then mapped the 11 items that were necessary to assess the 3 dimensions onto the overall attention

dimension to examine their spacing. We noted, in particular, that item spacing was sparse at the higher “difficulty” end of the dimension, perhaps because complex attention tends to involve deployment of multiple subprocesses. Thus, we selected additional items in the pool from the more difficult end of the dimension to supplement the factor-based items, while deleting items that were of similar difficulty to the factor items, resulting in a 22-item scale, essentially a 50% shortening.

Summary

The development of a reliable and valid observational rating scale is a complex and multi-step process. However, the development process itself may be very revealing in terms of how different observers view the construct in question and the factors that underlie systematic disagreements. In the case of the MARS, we conducted several small qualitative and quantitative pilot studies before constructing the first version of the scale that was subjected to large-scale research. In a multicenter study involving 228 patients, we then assessed interrater agreement between just two rehabilitation disciplines who assessed patients at a single point in time.[4] This documented sufficient reliability to allow us to proceed to the next step, and also allowed us to shorten the scale and characterize its factor structure.[3] A subsequent study on 149 patients allowed us to examine interrater agreement among a broader set of rehabilitation disciplines, to assess changes in MARS scores during the recovery period, to validate the scale against other measures of attention, and to assess its response to pharmacologic treatment.[5] Although the scale is now sufficiently developed for dissemination and use, there clearly remain further opportunities to explore its validity and clinical utility.

References

1. Sireci, S.G., *On validity theory and test validation.* Educational Researcher, 2007. **36**(8): p. 477-481.
2. Bogner, J.A., J.D. Corrigan, R.K. Bode, et al., *Rating scale analysis of the agitated behavior scale.* Journal of Head Trauma Rehabilitation, 2000. **15**(1): p. 656-669.
3. Hart, T., J. Whyte, S. Millis, et al., *Dimensions of disordered attention in traumatic brain injury: further validation of the Moss Attention Rating Scale.* Arch Phys Med Rehabil, 2006. **87**(5): p. 647-55.
4. Whyte, J., T. Hart, R. Bode, et al., *The Moss attention rating scale (MARS)ã for traumatic brain injury: Initial psychometric assessment.* Archives of Physical Medicine and Rehabilitation, 2003. **84**(2): p. 268-276.
5. Whyte, J., T. Hart, C. Ellis, et al., *The Moss Attention Rating Scale for traumatic brain injury: Further explorations of reliability and sensitivity to change.* Archives of Physical Medicine and Rehabilitation, 2008. **89**: p. 966-973.