Clinician Judgments of Functional Outcomes: How Bias and Perceived Accuracy Affect Rating

Aaron M. Wolfson, PhD, Jason N. Doctor, PhD, Stephen P. Burns, MD


Objective: To evaluate the accuracy of clinician judgments of patient function, the susceptibility of judges to bias, and the relation between a judge’s degree of belief in his/her accuracy of classification to observed accuracy when using the FIM™ instrument.

Participants: Fifty rehabilitation professionals.

Setting: 3 urban medical centers.

Design: Four randomized experiments among subjects to examine the effect of potentially biasing information on FIM ratings of patient vignettes. Participants answered 60 true/false questions regarding patient function and FIM score and indicated confidence in the accuracy of their answers.

Interventions: Manipulation of patient information.

Main Outcome Measures: The standard FIM 7-point scale, observed proportion of correct responses to the 60 true/false questions, and a 6-category confidence scale for each of the 60 questions were used as dependent measures.

Results: FIM ratings assigned to others biased participants’ FIM ratings of patient vignettes. Functional ability was overestimated when ratings in other domains were high and underestimated when they were low. Participants were overconfident in their ability to answer FIM questions accurately across all professional disciplines.

Conclusion: Bias and poor judgment of level accuracy play a significant role in clinician ratings of patient functioning. Blind ratings and training in debiasing are potential solutions to the problem.

Key Words: Activities of daily living; Judgment; Observer bias; Rehabilitation.

© 2000 by the American Congress of Rehabilitation Medicine and the American Academy of Physical Medicine and Rehabilitation

FUNCTIONAL OUTCOMES are central to rehabilitation research and practice.1 Treatment planning, program coordination, and allocation of financial resources hinge on functional assessment of physical and cognitive ability within rehabilitation clinical care.2 In addition, these outcomes are important to medicine in general.3-6 Rehabilitation researchers have taken a leading role in developing assessment devices that capture patient functional status.7-10 Although rehabilitation outcomes research has focused on traditional psychometric criteria of measurement reliability and validity,11-12 less attention has been given to the validity of the judgment and decision-making processes of clinicians who assess patient function. Consideration of human judgment factors is important when individuals are intermediaries in scientific observations.

Assessment in medical settings usually requires direct observation of a particular behavior and a subsequent recorded observation or judgment of that behavior. Magnetic resonance imaging and blood tests are examples of assessments in which recorded observations are based only on physical or biological theory; there is no human filter that could introduce bias and/or error. However, for most forms of functional assessment, the clinician’s judgment and recorded observations of target behaviors may significantly impact final assessment. The clinical judge as an intermediary may become susceptible to external biases (fig 1).

Judgments of patient functional status by using standardized assessment instruments require a complex integration of information. They involve recalling interactions with the patient and the mapping of these recollections to prespecified functional categories. The inherent difficulty of the task may occasion short cuts in reasoning; normally termed heuristics.13 One could argue that raters use time-saving heuristic-based decision rules to maneuver more efficiently through functional assessment. Therefore, if raters are vulnerable to heuristic-based biases, it is likely that they are biased by additional external information.14 This study examines whether biases and heuristics influence assessment of functional outcome by clinicians by using the FIM™ instrument.

The FIM and Clinical Judgment

The FIM is an important measure of functional outcome; it relies heavily on clinician-based judgment and decision making. The primary tool for collecting outcomes data for the national rehabilitation data set,15 the FIM instrument it assesses basic activities of daily living using 18 items from 6 subscales: self-care, sphincter control, mobility, locomotion, communication, and social cognition.9 Data from the FIM are managed nationally by the Uniform Data System for Medical Rehabilitation (UDSmr), which is also responsible for collecting, organizing, and reporting FIM data.9 The UDSmr was established at the State University of New York at Buffalo in 1987; today, approximately 1300 facilities in the United States, Canada, and Australia subscribe to its database, which is comprised of 1,000,000 patient records.15 Subscribers must maintain a predetermined performance level by showing a minimum FIM rating proficiency to receive quarterly reports from the data set. UDSmr requires that all subscribing facilities be examined yearly to ensure that raters maintain a minimum skill level so that the database is not contaminated with invalid information.

Currently, the Medicare system of payment to hospitals for rehabilitation services is being redesigned under a new prospective payment system. Beginning in 2000, Medicare, which funds 70% of all inpatient rehabilitation hospital stays, will link reimbursement to functional assessment outcome measures.
When this article was written, it was undecided whether reimbursement would be based on FIM ratings, Penn Ability Systems\textsuperscript{8} FIM-FRG, or the minimum data set (MDS). However, the relationship of FIM outcomes to the financial solvency of rehabilitation facilities underlines the importance of accurate FIM ratings by clinicians.

Researchers have performed many studies of the FIM and have found it to be a reliable and valid measure of disability for an acute rehabilitation population using standard psychometric criteria.\textsuperscript{16-20} These studies have typically used multiple ratings of the same patients by different clinicians and have examined interrater agreement. To our knowledge, however, there has been no analysis of whether the FIM is susceptible to different biases that have been shown in other domains involving human judgment.

**Potential Heuristics and Biases in Functional Assessment**

For the past 25 years, researchers in human judgment and decision making have accumulated evidence that judges use shorthand cognitive strategies, or heuristics, to conserve mental energy and to reach expedient decisions.\textsuperscript{21} Decisional heuristics deviate from the normative model of decision making, which posits that judgments and decisions result from careful comparisons and implementations of logical or rational rules.\textsuperscript{22} Heuristics represent rules of thumb that judges use as guidelines to make tasks less complex.\textsuperscript{23,24} Because heuristics typically have high validity to judges, they are particularly attractive as intuitive and efficient methods for decision making.\textsuperscript{22} However, individuals rarely understand the actual constraints on their cognitive abilities; consequently judges may be unaware when the probability of making erroneous decisions is particularly high.

The heuristics of representativeness and availability are useful in understanding how rehabilitation professionals may be susceptible to biasing information when completing functional measures such as the FIM. These 2 heuristics may also explain why clinicians might be vulnerable to inaccuracies in estimating their own abilities as judges of patient function. The representativeness heuristic relies on the degree to which element A resembles element B, and the likelihood that A and B belong to the same category, based on this resemblance.\textsuperscript{16} For instance, a 65-year-old man with morbid obesity is a more representative member of a type 2 diabetes population than a 30-year-old man of average weight. Within rehabilitation, clinical judges of patient function may be influenced by the degree of sameness between a patient's diagnosis and demographic profile and the typical level of functional independence observed for patients similarly treated.

In the medical context, the availability heuristic occurs when previous clinical observations or experience may bias a current observation. For example, a group of individuals is asked to estimate the probability of full recovery from a spinal cord injury (SCI). If a member of the group has observed a recent patient make a full recovery, that fact may cause the observer to increase his/her estimate for the entire population with SCI. The observer has just relied on the availability heuristic. Just as with the representativeness heuristic, various factors influence the propensity to use the availability heuristic. For example, the more familiar an individual is with a particular group the easier it will be to mentally retrieve instances of the current situation. Similarly, the circumstance in which the information was initially encoded would also be influential (ie, if someone witnesses a fire, as opposed to reading about it). A more detailed account of these and other heuristics and biases is found elsewhere.\textsuperscript{23}

Several demographic and process variables common in medical rehabilitation may temper a clinician's FIM rating performance, and may be affected by decisional heuristics. For example, ageism, or the tendency to make judgments about people based solely on their age, has been well documented.\textsuperscript{25-29} Researchers have examined health care trainees and their potentially biased attitudes toward the elderly in both community and nursing home populations.\textsuperscript{27,28} Researchers have also shown that older patients seeking admission to multidisciplinary pain rehabilitation programs are discriminated against.\textsuperscript{30} This indicates that professionals in different settings may have discernable biases toward, and stereotypes of, older people that may affect the types of inferences made about them. Similarly, a patient's psychiatric diagnosis or history may bias clinician FIM ratings of that patient. Public opinions of persons with severe mental illness are not favorable, and this population is often viewed as dangerous and unpredictable.\textsuperscript{31} Psychiatric labels elicit stigmatizing responses separate from those directly attributable to symptomatic behavior.\textsuperscript{32} Researchers have argued that bias regarding psychiatric diagnosis exists among trained health care workers, including psychiatrists.\textsuperscript{33}

Traditionally, different rehabilitation team members are responsible for rating various FIM items. For example, speech therapists may rate communication, whereas psychologists may rate memory and problem solving. Although differences between team members' ratings have been documented,\textsuperscript{17} to date there has been no analysis of how FIM scores assigned by 1 team member may influence the scores assigned by other team members. Therefore, to determine how bias may be introduced into the ratings, we must draw inferences from the literature reporting psychologic research. For example, Asch\textsuperscript{34} showed how individuals can be manipulated to conform to subjective judgments in perceptually ambiguous situations. It is possible that a clinician who is uncertain about a patient's functional level in 1 domain of the FIM may base his/her rating on the ratings assigned by other clinicians in other domains.

**Discrepancy Between Perceived and Observed Accuracy**

In the medical decision-making literature, an individual making a judgment is said to be calibrated if “over the long run, for all propositions assigned a given probability, the proportion that is true equals the probability assigned.”\textsuperscript{35} Calibration can...
be examined empirically by asking judges to answer true/false questions followed up by a question about their degree of belief (over the range of 50% to 100%) that their answers are correct, where “50% chance” = completely uncertain and “100% chance” = completely certain that an answer is correct. Mean proportion correct is then evaluated against mean degree of belief for all questions.

If clinician perceptions about patient functioning are susceptible to bias, one could argue that rehabilitation staff members may also be miscalibrated to estimates of their own levels of accuracy. In describing rehabilitation staff ratings of patient functioning, it is important to address both the accuracy of the decision and the subjective degree of confidence the raters have that their decisions are correct. Clinical judges whose perceptions of their performance as FIM raters correspond to their observed performance are more likely to consult with others, or with assessment manuals when appropriate. Conversely, persons who are overconfident with their abilities to measure function accurately may be less likely to seek consultation, leading to more errors in measurement.

The purpose of this study was to examine whether rehabilitation staff is susceptible to externally biasing information when assigning FIM ratings to hypothetical patients and whether staff are overconfident that their decisions are correct. We hypothesized that raters will rate older patients functionally lower than younger patients, and will rate a patient with schizophrenia functionally lower than a patient without schizophrenia. Raters may also be biased by FIM ratings by other team members in other domains of functioning (ie, when functional ratings in other domains are high, participants should be biased upward and when functional ratings are low, the bias should be downward). Finally, it was hypothesized that staff members will overestimate their ability to correctly assess patient functional ability because of their reliance on heuristics in decision making.

METHODS

Overview

A single experimental questionnaire consisting of 2 distinct sections was answered by rehabilitation specialists who volunteered. The first section dealt with bias and the FIM and required participants to evaluate 4 hypothetical patient scenarios and to provide a FIM score for each item. The 4 scenarios represented 1 of 2 between-sample experimental conditions. All scenarios contained objective information invariant across conditions about patient function. This information was sufficient for identification of an appropriate FIM category, based on UDSMr training materials. One of 2 sets of nonessential, but potentially biasing, information was included in each question for 50% of the participants. Participants were randomly assigned to a condition (eg, positive vs negative bias) for each question, and each question was presented in a random order. Thus, each patient participated in 4 different experiments on bias. Participants were randomized to conditions separately within each experiment.

The second section was concerned with confidence analysis and required participants to complete a 60-item questionnaire containing FIM descriptions and confidence estimation questions. Between the 2 sections, participants responded to a total of 64 questions about patient functional status. Total participation time was approximately 35 to 45 minutes.

Participants

Participants included 50 medical rehabilitation staff members recruited from inpatient rehabilitation units at the University of Washington Medical Center, Harborview Medical Center, and the VA Puget Sound Health Care System in Seattle. All participants had completed FIM training, which was assessed by written examination. In addition, members of Division 22 (Rehabilitation Psychology) of the American Psychological Association (APA) were contacted via their Internet listserver and asked to participate in the study if they were using the FIM on a regular basis. Interested parties contacted by us (AMW) and experimental packets were mailed to them. Experimental procedures met with ethical standards, and the research was approved by each facility’s institutional review board. Subspecialties represented by potential participants included nursing, physical therapy, occupational therapy, social work, rehabilitation counseling, speech and hearing, rehabilitation psychology, and recreational therapy. Any rehabilitation staff member trained to use the FIM was eligible to participate.

Design and Measures

Bias experiments. The first half of the experimental questionnaire contained the 4, 2-group randomized experimental designs for evaluating a separate aspect of bias. Participants were randomly assigned to receive 1 of 2 versions of each of the 4 scenarios. Similarly, scenario presentation was randomized for each participant to avoid confounding by order. Paragraphs included functional behavioral descriptions adapted from UDSMr training materials. Level 4 functional descriptions from FIM training materials were included in the scenarios. At this level, patients perform 75% of the activity and require minimal assistance. Descriptions of patient function were identical for each pair of scenarios, with the potentially biasing information varying between groups. Specifically, participants received a question in which patients’ ages varied (ie, 25 years old vs 80 years old; asked to assess eating); 1 question in which psychiatric diagnoses varied (ie, no diagnosis, schizophrenia diagnosis; participants asked to assess memory); motor FIM questions (ie, bathing) in which FIM scores in other domains varied (ie, other FIMs 1 or 2 vs other FIMs 6 or 7); and cognitive FIM questions (ie, problem solving) in which FIM scores in other domains also varied. Each description included information about patient function that was constant across both conditions, and was sufficient for participants to identify accurately an appropriate FIM level. Participants read each scenario and assessed the functional level within a single functional domain by using the standard FIM 7-category scale. An example scenario for the experiment involving high versus low FIM ratings in other domains is presented in the appendix.

Perceived and observed accuracy evaluation. For the second portion of the study, participants read 60 functional descriptions followed up by a FIM numerical code for the associated domain of function. Functional descriptions included only the 5 “interior” FIM levels of function. Descriptions from levels 1 and 7 (ie, “exterior” FIM levels), which deal with clear absolutes in functioning, were assumed too easy to discriminate and were excluded. All functional descriptions were adapted from FIM training materials; there were 6 cognitive and 6 motor items for each of the 5 interior FIM levels. Questions were constructed from all 18 functional categories and distributed as equally as possible among interior FIM levels. Because there are only 5 subtypes of cognitive items on the FIM (ie, memory, comprehension, problem solving, expression, social interaction), 1 item was repeated for each functional level and presented as a “true” question and as a “false” question. Because some of the descriptions included percentages that might have cued respondents, alternative language was systematically substituted in a manner consistent with accepted subjective probability research methods. For exam-
RESULTS

Participant Demographics
Fifty rehabilitation professionals participated in the current study. Demographic information about the 50 participants is reported in Table 1. Approximately 42% of participants were employed at Harborview Medical Center, 25% came from the University of Washington Medical Center, 18% from the VA Puget Sound Health Care System, and roughly 15% volunteered from APA’s Division 22 listserver. Rehabilitation staff often rotate through 1 or more facilities during training. Because of the general continuity of training experiences, as well as the overlapping patient populations, we expected that participant characteristics and performance would be comparable across locations. No subsequent differences were evident between groups, therefore, data were collapsed across sites.

Bias
To assess whether participants were biased by information added to functional descriptions of hypothetical patient scenarios, 4 Mann-Whitney U tests were run on FIM rating by scenario type. Medians and interquartile ranges are reported in Table 2. There were no significant differences between group medians for scenarios 1 (age 25 vs age 80) and 2 (no psychiatric diagnosis vs schizophrenia), suggesting that neither extreme changes in age nor psychiatric status influenced partici-

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Cases (no.)</th>
<th>Median (IQR)</th>
<th>Z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 25</td>
<td>25</td>
<td>3 (2–3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 80</td>
<td>25</td>
<td>2 (2–3)</td>
<td>−.525</td>
<td>.600</td>
</tr>
<tr>
<td>Scenario 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No DX</td>
<td>29</td>
<td>4 (4–5)</td>
<td>−1.373</td>
<td>.170</td>
</tr>
<tr>
<td>SZ</td>
<td>21</td>
<td>4 (4–4.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 3 (Motor)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low other FIM</td>
<td>20</td>
<td>3 (3–3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High other FIM</td>
<td>30</td>
<td>4 (4–4)</td>
<td>−4.144</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Scenario 4 (Cognitive)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low other FIM</td>
<td>27</td>
<td>4 (3–4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High other FIM</td>
<td>23</td>
<td>5 (4–5)</td>
<td>−3.721</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Abbreviations: No DX, no diagnosis of schizophrenia; SZ, schizophrenia diagnosis; Low, FIM levels 1 and 2; High, FIM levels 6 and 7.

Table 2: Medians and Interquartile Ranges (IQRs) for Participants’ Ratings of Hypothetical Patient Scenarios

Table 1: Participant Demographic Information

<table>
<thead>
<tr>
<th></th>
<th>Psychologists (n = 12)</th>
<th>Nurses (n = 8)</th>
<th>PT (n = 13)</th>
<th>OT (n = 10)</th>
<th>Speech (n = 7)</th>
<th>Total (n = 50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional experience*</td>
<td>9.8 ± 7.7</td>
<td>18.2 ± 5.6</td>
<td>9.0 ± 7.7</td>
<td>6.8 ± 4.8</td>
<td>5.9 ± 5.1</td>
<td>9.5 ± 7.7</td>
</tr>
<tr>
<td>Rehabilitation experience*</td>
<td>9.5 ± 6.7</td>
<td>14.6 ± 3.8</td>
<td>5.5 ± 5.3</td>
<td>6.2 ± 4.1</td>
<td>4.4 ± 4.1</td>
<td>8.3 ± 6.5</td>
</tr>
<tr>
<td>Years since last FIM training*</td>
<td>2.1 ± 1.2</td>
<td>2 ± 1.5</td>
<td>1.9 ± 2.5</td>
<td>2.4 ± 2.6</td>
<td>1.9 ± 1.5</td>
<td>1.9 ± 1.9</td>
</tr>
<tr>
<td>Age*</td>
<td>39.5 ± 6.9</td>
<td>44.5 ± 5.8</td>
<td>35.3 ± 6.6</td>
<td>37.2 ± 10.4</td>
<td>31.2 ± 6.8</td>
<td>37.0 ± 9.1</td>
</tr>
<tr>
<td>Men</td>
<td>6 (50%)</td>
<td>2 (25%)</td>
<td>3 (23%)</td>
<td>2 (20%)</td>
<td>1 (14%)</td>
<td>14 (27%)</td>
</tr>
<tr>
<td>Women</td>
<td>6 (50%)</td>
<td>6 (75%)</td>
<td>10 (77%)</td>
<td>8 (80%)</td>
<td>6 (86%)</td>
<td>36 (73%)</td>
</tr>
</tbody>
</table>

* Values presented as mean years ± SD.
Abbreviations: PT, physical therapy; OT, occupational therapist; Speech, speech therapist.

Statistical Analysis
Because FIM data are ordinal, a nonparametric Mann-Whitney U test for 2 independent samples was performed for each of the 4 bias experiments. A Bonferroni correction was applied a priori to control for experiment-wise error across experiments (ie, α for each test was set at .0125). Perceived and observed accuracy was compared by computing the mean and standard deviation, within and across rehabilitation discipline for observed accuracy, and by comparing the same descriptive statistics for perceived accuracy. Under a conservative set of assumptions regarding FIM variance, difference in central tendency between groups, and the power characteristics of the Mann-Whitney test, the 4 experiments on bias each had a greater than an 80% power to detect a significant effect with 50 participants.37 Statistical analyses were performed using SPSS for Windows, version 9.0.
Participants’ FIM ratings. There were, however, significant differences in group ratings for scenarios 3 and 4. In scenario 3, participants were asked to give FIM ratings for a motor item in which hypothetical team ratings in other domains were either high or low. Median FIMs for the “low” scenario differed significantly from median FIMs for the “high” scenario (median_FIM low = 3, median_FIM high = 4; Z = −4.144, p < .0001). Participants gave significantly lower FIM ratings when low team FIMs were presented in other domains, but gave FIM ratings closer to the true level (ie, level 4), when high team FIMs were presented in other domains. In scenario 4, participants were asked to give FIM ratings for a cognitive item in which hypothetical team ratings were either high or low. Participants endorsed significantly lower ratings when team FIMs in other domains were low (median_FIM low = 4) and endorsed higher ratings when team FIMs in other domains were high (median_FIM high = 5; Z = −3.721, p < .0001). These findings suggest that team member ratings in other domains of functioning tended to bias clinician ratings on a target FIM item.

Confidence

To determine whether participants were overconfident in their ability to assess patient functional status, mean accuracy was calculated by summing the percentage correct on part 2 of the questionnaire. Mean confidence was calculated by summing confidence results across participants. Confidence versus accuracy is reported in table 3. Results indicate that participants were overconfident in their ability overall, with psychologists giving the greatest degree of overconfidence and physical therapists showing the smallest, but still sizeable discrepancy. Because FIM dimensions are typically assessed by different professional specialties, confidence and accuracy values were analyzed further by cognitive and motor FIM items. Psychologists, physical therapists, and occupational therapists performed approximately the same for both cognitive and motor items. Speech therapists performed better on cognitive items than on motor items, and were overconfident for both. Nurses, who are usually responsible for rating motor items, performed better for cognitive items than for motor items, and were overconfident for both. Statistically significant differences between specialties were not tested because of sample size limitations. Although mean values for accuracy and confidence merely represent rough measures of the overconfidence effect, the participants’ inconsistent beliefs about their performances is interesting.

DISCUSSION

Results generally support the hypotheses that external information about hypothetical team member FIM ratings affects judges’ FIM ratings, and that staff members were typically overconfident in their ability to identify correctly hypothetical patients’ FIM ratings. Of the 3 potential sources of bias studied, only information provided by other team members ratings of the patient in other domains of functioning produced a significant bias effect. Participants responded to other team members’ ratings, overestimating functional ability when ratings in other FIM domains were high, and underestimating functional ability when ratings in other FIM domains were low. Additionally, raters were overconfident across all professional disciplines. Although judges’ ratings were not significantly different for either version of scenario 1 (age 25 vs 80), ratings were consistently lower than would be expected for a level 4 functional FIM description (median_FIM = 3, median_FIM = 4). In both versions, the patient was described as having sustained a high-level SCI in a fall from a height of 10 feet. The nature of the particular injury was perhaps more influential than age for both experimental conditions. Furthermore, the functional category being assessed was eating, which may have been particularly difficult to assess or predict for a patient with midcervical level tetraplegia. The correlation between motor scores and FIM scores in patients with tetraplegia is lower for eating than for other items such as grooming and bathing, so a range of FIM scores for eating may be seen with a given neurologic classification.

In scenario 2, participants failed to rate a patient with a history of unmedicated schizophrenia significantly lower than a patient with no history of psychiatric illness. Although there is no obvious explanation for this finding, it is possible to speculate about potential explanations. For example, the base rate for schizophrenia in the United States has been found to range between 2% to 2%. Because this condition is rare, participants may have not had well-formed mental templates of how a person with active psychosis would function in an inpatient rehabilitation unit. Moreover, few patients in an inpatient unit are actively psychotic. Similarly, because there was no mention of active psychiatric symptoms, it would be difficult to form an opinion about functional abilities. Therefore, it appears that participants focused more closely on pertinent information in the functional description rather than relying on cognitive conceptions of mental illness to make final decisions.

In scenarios 3 and 4, participants were presented with ratings in other FIM domains supposedly made by other team members and were asked to rate both a motor item and a cognitive item. For the motor item, they rated a woman diagnosed with multiple sclerosis significantly lower for bathing when FIM ratings in other domains were low (ie, ones and twos), but made accurate ratings when ratings in other domains were high (ie, sixes and sevens). Similarly, judges overestimated FIM ratings for a cognitive item when team ratings in other domains were high, and underestimated when team ratings in other domains were low. These general tendencies may be explained by a social psychology theory that considers knowledge about group decision making and normative social influence.

Table 3: Mean Accuracy and Mean Half-Range Confidence by Profession

<table>
<thead>
<tr>
<th>Professional Specialty</th>
<th>Mean Accuracy</th>
<th>Mean Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Motor</td>
<td>Cognitive</td>
</tr>
<tr>
<td>Overall sample</td>
<td>68%</td>
<td>67%</td>
</tr>
<tr>
<td>Psychologists</td>
<td>68%</td>
<td>67%</td>
</tr>
<tr>
<td>Nurses</td>
<td>71%</td>
<td>74%</td>
</tr>
<tr>
<td>Physical therapists</td>
<td>74%</td>
<td>70%</td>
</tr>
<tr>
<td>Occupational therapists</td>
<td>67%</td>
<td>73%</td>
</tr>
</tbody>
</table>

Arch Phys Med Rehabil Vol 81, December 2000
The subtle performance difference between motor and cognitive scenarios deserves comment. For the motor item, participants appeared to be biased more by lower ratings from team members, whereas for the cognitive item participants were biased by higher team ratings. The FIM is divided into 13 motor subcategories and 5 cognitive categories. Inpatient rehabilitation units commonly use various team members who are responsible for rating different areas of patient functioning. For example, nursing staff would be better able to assess a patient’s bowel and bladder functioning than would a psychologist. Although this division of labor is largely implicit, and depends heavily on the particular facility’s staff composition, occupational therapists, physical therapists, and nurses usually assess motor items, whereas speech pathologists and psychologists assess cognitive items. Because the majority of FIM categories are based on motor descriptions, participants may have found the cognitive scenario more difficult to interpret and, hence, deferred to the expertise of “other staff.” Similarly, FIM rating criteria for motor items are more clearly operationalized than for cognitive items, which increases the perceived difficulty of the task for a FIM cognitive item. Research suggests that individuals are more susceptible to social influence when confronted with a difficult or ambiguous task and are unsure about the validity of their responses. Note that participants were evaluated for bias based on their responses to questions that may fall outside of the domain of items they are most likely to assess. For example, nurses were asked to rate cognitive items that are usually assessed by speech pathologists or psychologists. Future research should more clearly evaluate bias by professional specialty.

Our results indicate that rehabilitation professionals were influenced by both the low and high team FIMs in other domains for a cognitive target item, but only by low team FIMs in other domains for a target motor item. Given these results, participants may have been less influenced by patient demographic variables, and more susceptible to social comparison when making FIM ratings. This is clinically significant, given that inpatient physical rehabilitation units typically use multidisciplinary treatment teams. Team members may have access to other’s FIM ratings before making decisions about patients, and therefore are exposed to potential bias through social comparison.

As predicted, rehabilitation professionals were consistently overconfident in their ability to accurately identify appropriate FIM ratings for a series of functional descriptions. This fact can best be explained by the hypothesized use of decisional heuristics to process the amount of information necessary to make judgments of functional ability based on a predetermined set of FIM rating criteria. Decisional heuristics allow decision-makers to conserve cognitive resources by providing short cuts in reasoning.

Researchers have argued that the medical training model emphasizes learning prototypes of medical syndromes, which are then compared with actual patient information to arrive at final diagnostic decisions. This process is particularly susceptible to the representativeness heuristic described by Tversky and Kahneman. For example, in the current instance, raters not only had to remember the FIM rating criteria, but also may have had to conjure an image of how a patient at each particular FIM level might appear. Therefore, participants may have had to recall the recommended amount of assistance associated with each FIM level and to compare each statement with a mental image of a patient fitting that description. A more efficient way would be to remember a patient with similar functional ability and compare him/her with the functional description. Clearly, there is potential for error at several steps of this decision process that could encourage judges’ overconfidence in their ratings.

Several limitations of our study must be addressed. Even the most realistically written descriptions do not equate to the actual experience of meeting a patient, reviewing his/her hospital record, and conducting a detailed initial evaluation. The patient scenarios we used were adapted from actual UDSMR training materials, were supplemented with basic demographic information, and reviewed by outside rehabilitation professionals before they were included in the study. However, in clinical practice, staff is almost never expected to make patient evaluations based on a written paragraph.

The process of completing experimental packets was also different from the way staff typically makes patient judgments. Although FIM ratings must typically be performed within 3 days of admittance and discharge, rehabilitation professionals have some freedom to consult with colleagues, refer to the FIM decision trees, order additional psychological testing, or perform multiple baseline evaluations before making a decision about a patient’s functional ability. In essence, the benefits of an in vivo, multidisciplinary approach to patient evaluation were suppressed in the current study. Our study traded off the richness and complexity of clinical practice to achieve a careful and rigorous experimental analysis of bias. Such trade-offs are generally considered worthwhile in medical and social science research. For example, the implementation of randomized clinical trials to gain knowledge about a drug’s effectiveness involves similar trade-offs.

The structure in which information was presented to judges may have added to their overconfidence. Kahneman et al showed that confidence increased when information was presented in a paragraph form as opposed to being listed separately. Other than in UDSMR training and testing, rehabilitation professionals rarely make FIM judgments based on a written synopsis of a patient’s behavior. Likewise, Gruppen et al showed that physicians were more confident in their diagnostic accuracy when the task was limited to information integration as opposed to collection and integration. This is pertinent in the current research because participants were simply presented with a limited amount of information and not allowed to control their own access to patient information before assigning a FIM rating. Determining the thought processes that resulted in our participants’ FIM ratings is not possible. However, decisional heuristics seem an appropriate alternative explanation of the well-documented phenomenon of overconfidence in decision making.

Typically, rehabilitation interdisciplinary staff meetings to discuss patient care and treatment planning leave staff members open to this bias. To counteract this bias, as well as the observed overconfidence in FIM ratings, several interventions could be implemented in rehabilitation team meetings. One solution would be to elicit “blind” ratings of each patient, wherein staff members rate their particular FIM domain without prior knowledge of colleagues’ ratings. This may help eliminate bias caused by other team members’ ratings, but may also seem unnatural to members who are used to making these decisions in group settings. If blind rating is not an option, encouraging strict adherence to the FIM decision trees included in UDSMR training materials may curb the inclination to rely on other ratings. This could be performed by designating 1 team member to be responsible for comparing group decisions with the FIM decision trees. Similarly, if FIM ratings are made in a group, it would be beneficial to elicit input from several team members to get a comprehensive impression of the patient’s actual functional ability within any particular domain. For example, it may be difficult for a speech therapist to assess
whether a patient remembers 50% of the time they meet with the patient for 1 or 2 hours per week. In this case, enlisting the help of nurses and other therapists may improve clinical decision making. In addition, including confidence measures in FIM mastery testing and providing feedback about performance may help staff realize its limitations when making functional assessments. Because of shortened inpatient hospital stays, the pressure to show functional improvement through increased FIM ratings may be a pervasive problem. Most medical equipment requires occasional recalibration to ensure accurate measurements. Therefore, periodically recalibrating rehabilitation professionals’ ability to measure patient functioning accurately could ensure more meaningful assessment.

Functional assessment is crucial in physical rehabilitation and our findings could easily be extrapolated to apply to other assessment measures being used in the field. Tools such as the Barthel index and the Craig Handicap Assessment and Reporting Technique also require raters to make decisions about patient functioning. However, the FIM was selected for our study because of its widespread use as well as the likelihood that FIM scores will be tied more closely to Medicare reimbursement in the near future.

Finally, bias introduces systematic variation in assessment data. Such systematic variation can act to increase reliability coefficients in standard psychometric evaluations of the FIM. Therefore, bias may contribute to high observed reliability estimates. Our findings also call into question the validity of Rasch analyses performed on FIM outcomes. Rasch analysis assumes that responses to items by clinical raters are conditionally independent. We have shown that conditional independence is violated when FIMs in 1 domain are completed by team members who have knowledge of FIM scores in other domains. Therefore, efforts to debias clinical judges through targeted education modules may strengthen Rasch analysis findings in future research. At a policy level, bias can result in inaccuracies in functional outcomes measurement when outcomes are used to make policy decisions regarding payment. For these reasons, the importance of decreasing bias in FIM ratings cannot be overemphasized. Future research should continue to address strategies for reducing bias and decreasing overconfidence among FIM raters.

CONCLUSION

These results indicate that having access to team members’ FIM ratings in other domains tends to bias subsequent judgments on target FIM items. Rehabilitation professionals were biased downward when presented with hypothetical team member ratings on a motor item and biased upward when presented with team ratings on a cognitive item. In addition, study participants were overconfident in the accuracy of their responses to FIM items. In the current sample, professionals generally achieved greater accuracy in the domains that they normally rate, with the exception being nurses who were more accurate rating cognitive items than motor items. The FIM is, and will continue to be, an important outcome measure in physical rehabilitation. Further research should examine the efficacy of training interventions designed to inform judges’ about their natural rating patterns.

Acknowledgments: The authors thank Leighton Chan, MD, and Mark Jensen, PhD, for their critical review of previous drafts of this manuscript, as well as David R. Patterson, PhD, for his assistance in recruiting psychologists to participate in the study. Aaron M. Wolfson also acknowledges Margot Holaday, PhD, who served as dissertation committee chairperson for the original project on which the current research was based.

References


APPENDIX: EXAMPLE EXPERIMENTAL SCENARIO

The patient is a 45-year-old woman with multiple sclerosis. The patient has been married for 23 years and has 2 adolescent children who live in the home. The patient washes in the tub by using a tub bench and hand-held shower. She needs the water temperature adjusted and assistance to wash the lower legs and feet. The staff describes the patient as pleasant and cooperative. The rehabilitation team has assessed her FIM levels in the following areas:

- **Eating:** (6 or 1)
- **Grooming:** (6 or 2)
- **Toileting:** (7 or 2)

What FIM level would you give for **Bathing**?

**Suppliers**

- a. SPSS Inc, 233 S Wacker Dr, 11th Fl, Chicago IL 60606.