The Assessment of Motor and Process Skills Applied Cross-Culturally to the Japanese

Sawako Goto, Anne G. Fisher, Wanda L. Mayberry

Key Words: activities of daily living evaluation • culture (sociology)

Objectives. The purposes of this study were to examine the cross-cultural validity of the Assessment of Motor and Process Skills (AMPS) as well as the reliability of raters from different cultures.

Method. Six trained raters from diverse cultural backgrounds scored 10 Japanese subjects' performances on familiar and culturally relevant tasks of their own choosing.

Results. Results indicated high cross-cultural validity and interrater reliability for the AMPS, as indicated by goodness of fit of subjects and raters via the many-faceted Rasch measurement model. The relative rater severity was also evaluated cross-culturally by the standardized difference (z). The Japanese rater and one of the three American raters varied significantly in severity between their scoring of Japanese subjects versus European subjects.

Conclusion. The results support the hypothesis that the AMPS can be used as a cross cultural instrument of daily living assessment. Further study is needed to clarify the issue of possible cultural bias in rater severity.

Occupational therapists use many kinds of assessments to gather and analyze information about their clients, to know their needs and potentials, to plan and implement appropriate interventions, and to evaluate the effects of interventions in enhancing their clients' independence. Although each assessment is important, the evaluation of functional (occupational) performance, rather than the physical performance components, is the primary focus of occupational therapy (Mathiowetz, 1993).

Literature Review

The shift from focusing on physical performance components to focusing on a more holistic and naturalistic model has demanded that occupational therapists study human behavior by observing the person performing in the natural environment (Mathiowetz, 1993; Yerxa, 1991). Moreover, because occupational therapists are experts in occupational functioning, it is important that they develop, refine, and use measurement tools that consider observation of naturalistic behavior in the context of occupational performance.

Evaluations of activities of daily living (ADL) are most typically classified under the rubric of functional assessment (Fisher, 1992a) and can be divided into those that are basic tests of personal ADL (e.g., feeding, toilet-
ing, bathing, transferring) and tests of more complex, instrumental activities of daily living (IADL) (e.g., housework, shopping, budgeting, cooking meals) (Lawton & Brody, 1969). Many of the existing ADL assessments have problems, including (a) the use of the uniform tasks regardless of the client's values, interests, or familiarity with the tests (Fisher, 1992a); (b) development and validation only for specific diagnostic or age groups (Feinstein, Josephy, & Wells, 1986; Law &Lets, 1989); (c) emphasis on self-report rather than direct observation (Law, 1993); (d) limited knowledge of task hierarchy (Fisher, 1992b, 1995a); (e) use of ordinal scales (Merbitz, Morris, & Grip, 1989); (f) inappropriate length of tests (Law, 1993); and (g) lack of standardization based on extensive research (Law &Lets, 1989).

Of particular importance is the issue of interrater reliability. Raters for ADL assessments come from diverse backgrounds, and a rater's judgment is influenced by factors such as his or her profession, familiarity with a client (Lawton, 1987), and culture (Skawski, 1987). However, there are few studies of rater reliability for ADL assessments that take these into consideration. Cross-cultural interrater reliability studies for ADL assessments are lacking. As a result, the rater reliability of many assessments is questionable, especially when used cross-culturally.

The situation in Japan is even more complicated. Of the many ADL assessments used in Japan, some have been developed in Japan but are not standardized (Ando, 1992). Those assessments developed in the United States and Canada, such as the Barthel Index (Mahoney & Barthel, 1965), the Katz ADL Index (Katz, Ford, Moskowitz, Jackson, & Jaffee, 1963), and the Functional Independence Measure (Granger, Hamilton, Keith, Ziefzny, & Sherwin, 1986), although standardized, may not be valid in Japan because of cultural and contextual differences. Assessments developed and standardized in other countries should be restandardized for Japanese clients before being used, but for the most part, this has not yet been done. Consequently, Japanese occupational therapists usually rely on nonstandardized or "homegrown" ADL evaluations in practice (Taniguchi, 1994).

Because culture influences most of what we do, many problems arise when an assessment standardized for one country is applied in another. One of the most serious problems pertains to assessment norms. Those standardized ADL assessments mentioned previously are based on the sociocultural norms of white middle-class North Americans (Kinebanian & Stomph, 1992; Law, 1993; Skawski, 1987). The norms of a specific culture influence many facets of an assessment, such as rater scoring, a person's expected level of performance, a person's usual way of performing a task, or the set of tasks the person needs to perform. Therefore, North American norms might lead to misinterpretations of results when assessments are applied cross-culturally (Jitapunkul, Kamolratanaakul, & Ebrahim, 1994; Saeki, Clark, & Azen, 1985).

The Assessment of Motor and Process Skills (AMPS) (Fisher, 1995a) offers a unique opportunity to develop a standardized, contextual, cross-cultural ADL evaluation. Used as an observational IADL assessment, therapists evaluate simultaneously (a) a client's ability to perform functional daily living tasks and (b) the quality of motor and process skills necessary for effective task performance. Therapists observe directly how clients interact with a naturalistic environment and which component actions of task performance are a problem for them.

More specifically, in the context of the client's actual performance of two or three familiar, appropriately challenging, and culturally relevant functional tasks of his or her choice, the client is rated on two taxonomies that include 16 motor and 20 process skill items. In the AMPS, motor skills are defined as observable actions used to move the body or the task objects during the task performance, and process skills are defined as observable operations used to logically organize and adapt behavior over time in order to complete a client-specified task (Fisher, 1995a). Because the two taxonomies are hypothesized to be universal and culture free (Fisher, 1995a; Magalhães, Fisher, Bernspång, & Linacre, 1996), the AMPS can be used with clients with various conditions, from a variety of age groups, and from various cultures. Because the motor and process skills are assumed to be universal across tasks, the client is able to exercise choice in the evaluation process. The opportunity for choice is important because the conceptual model that underlies the AMPS assumes that a client's motivation to perform tasks is based on the task's meaning and cultural relevancy to the client; the client's function is maximized when performing a task of his or her choice (Doble, 1988).

Rasch analysis (Andrich, 1988; Wright & Masters, 1982; Wright & Stone, 1979) converts raw item data into equal-interval units of measurement through a logistic transformation of the proportion of persons receiving a score for a given item. Moreover, by using many-faceted Rasch analysis (Linacre, 1993) to develop the AMPS, the difficulty of the skill items, the challenge of tasks, and the severity of raters can each be expressed on the same linear scale (Fisher, 1993, 1994). Therefore, it is possible to compare the abilities of clients, even when they perform different tasks or are rated by different raters (Fisher, 1993, 1994, 1995a).
The AMPS has been successfully applied cross-culturally to persons who are African-American (Bennett, 1995), Mexican-born Hispanic (Clawson, 1995), Cuban-born Hispanic elderly (Dickerson & Fisher, 1995), Taiwanese (Fisher, Liu, Velozo, & Pan, 1992), and Swedish (Magalhães et al., 1996). Subjects from these diverse cultural backgrounds have been shown to have acceptable goodness of fit to the AMPS measurement model. When Caucasian-American raters scored the performances of Mexican-born Hispanic or African-American subjects, there were no significant differences between these subjects' AMPS ability measures and those of matched samples of Caucasian-American subjects (Bennett, 1995; Clawson, 1995). This suggests that Caucasian raters scored the performances of non-Caucasian-American subjects without bias.

The AMPS has not yet been applied to Japanese persons. Furthermore, although the interrater reliability of the AMPS has been high among calibrated raters from western, predominantly Caucasian countries (e.g., the United States, the United Kingdom, Sweden) (Fisher, 1995a), the interrater reliability between North American, Swedish, and British (hereafter collectively referred to as European) raters and Asian raters (e.g., Japanese) has not been compared when scoring Asian clients.

The purposes of this study were to examine the validity of the AMPS derived from European values as applied to Japanese subjects and to examine the cross-cultural interrater reliability of the AMPS. It was hypothesized that (a) the AMPS would be valid for Japanese subjects when they performed cultural-relevant tasks, and (b) raters would score consistently when evaluating the subjects. Moreover, the rater severity calibrations should not be significantly different when rating European or Japanese subjects.

Method

Subjects

The subjects selected for this study were 10 community-living Japanese persons (5 women, 5 men) without known physical, emotional, or cognitive disabilities who had lived in the United States for less than 12 months ($M = 12.4$ months; $SD = 8.8$). The subjects were known to the first author and volunteered to participate. Their ages ranged from 22 to 38 years ($M = 28.9$ years; $SD = 3.98$).

Six occupational therapists who had been trained and calibrated as reliable AMPS raters participated as raters in this study. They were three American raters, one Japanese rater, one Swedish rater, and one British rater.

Instrument

The AMPS was administered to the subjects in accordance with the standardized procedures delineated in the AMPS manual (Fisher, 1995a). Each of the 16 motor and 20 process skill items is rated on a 4-point rating scale, with scores ranging from 1 (deficit: skill item deficit is severe enough to result in task breakdown, risk, danger, or an unacceptable slowing of the task progression) to 4 (competent: no evidence of a skill item deficit affecting performance).

The following five tasks were chosen from the AMPS task choice list for use in this study: (a) preparing scrambled or fried eggs, toast, and brewed coffee or tea; (b) preparing scrambled or fried eggs, meat, and brewed coffee or tea; (c) preparing French toast and beverage; (d) preparing a tuna salad sandwich; and (e) vacuuming with moving light-weight furniture. These tasks were chosen because they are frequently done by young persons in Japan. The subjects also stated that two or more of the tasks were familiar to them. Therefore, the tasks were considered to be relevant to Japanese culture. Because we examined young subjects without disabilities, relatively difficult tasks were chosen by referring to the hierarchy of task difficulty in the AMPS manual (Fisher, 1995a).

During an AMPS task observation, each person is expected to perform the tasks in his or her usual manner with the tools and materials that he or she usually uses. The person is rated according to the criteria specified in the AMPS manual (Fisher, 1995a). A person receives a low score if the observed performance is not logical or efficient as determined or defined by (a) the skill item scoring criteria, (b) the constraints of the task performed, or (c) the person's culture. For example, the AMPS manual specifies that the task constraints for making a tuna salad sandwich include using canned meat (e.g., tuna, crab) and one chopped ingredient (e.g., celery, onion, pickle) and mixing them with mayonnaise (or equivalent). The person is expected to use two slices of presliced bread, cut the sandwich in half, serve it on a plate at a counter or table, and clean up the workspace.

A critical feature of the AMPS, intended to ensure contextual relevance, is that the person can use whatever tools he or she prefers or usually uses. For instance, Americans commonly uses either a spoon or fork to mix the tuna with the other ingredients. On the other hand, persons from Japan might prefer to use chopsticks. As long as the tool used is logical and effective for the action (e.g., mixing something), the person is not penalized. However, when a person selects a tool not well-suited to the demands of the task, his or her score is lowered on...
relevant AMPS skill items such as *uses ineffective tools* for the purpose (e.g., uses chopsticks to turn over soggy, fragile French toast in a frying pan).

Each person also can perform the task in his or her *usual* manner. For example, some persons clean up the workspace as they proceed through the task performance and others prefer to clean up afterward. Similarly, Japanese persons *often prefer* to add a lot of chopped vegetables to tuna, more than most Americans would add. Because this is an accepted procedure in Japan, the scores of a person from Japan should not be affected. However, if the performance is not efficient or logical, the score is lowered on relevant items. In the AMPS, efficient or logical is determined by such things as the person's ability to *effectively pick up, handle, or stir ingredients in a timely manner* and to perform task steps in a logical sequence. Therefore, if a person ineffective organizes his or her performance as it unfolds over time, his or her scores might be lowered on such skill items as *initiates* (evidenced by starting a step without hesitation), *continues* (evidenced by continuing an action, such as chopping an onion, to completion without interruption), or *sequences* (evidenced by logical ordering of actions or steps).

**Procedure**

To become trained and calibrated as an AMPS rater, interested occupational therapists participate in a 5-day *training workshop*. During the workshop, the raters *co-score* 8 to 10 videotaped or live task performances of European subjects. Three American raters, one British rater, one *Japanese* rater, and one Swedish rater who had been trained, calibrated, and verified as reliable raters were recruited for this study.

The AMPS was administered to the subjects by the Japanese rater in accordance with the assessment procedures described in the AMPS manual (Fisher, 1995a). The instructions were given in Japanese. All subjects were informed that the purpose of this study was to determine whether the AMPS was useful for evaluating Japanese persons.

One or two days before testing, each subject was interviewed to determine which of the task choices were familiar to him or her and which tasks he or she was willing to perform. The examiner procured needed tools and materials at this time. Each subject chose and performed two of the five task options in an environment that was familiar and relevant to the task, usually the subject's own home. Before starting the task, all needed tools and materials were made readily available and stored by the subjects in their usual places. The subject showed the examiner where all the needed tools and materials were located. Before observing each task, the examiner also verified the subject's understanding of the task criteria.

*To allow for co-scoring* by the other raters, all subjects were videotaped while performing the tasks, both of which were done in a single period.

All six raters co-scored a subset of the subjects, scoring a minimum of 10 of the 20 task performances. Before scoring the subjects, the European raters were told about Japanese cultural expectations, including the following: (a) it is appropriate for Japanese people to use chopsticks instead of other tools or utensils (e.g., spoon, fork, spatula) while cooking; (b) Japanese people wash their hands and materials more frequently than do European people; and (c) Japanese people often add more chopped vegetables to canned meat when making a sandwich than do European people. This procedure was designed to help equalize cross-cultural awareness and to control for the possibility of raters misinterpreting the subjects' behavior. The goal was that all raters would use Japanese cultural expectations when scoring the subjects, just as all raters would use European cultural expectations when scoring European subjects. During AMPS training workshops, raters become more aware of how their values and task performances may differ from those of others, including persons from other European cultural backgrounds than their own. Potential differences in Japanese cultural expectations were identified by the first and third authors through watching videotapes of Japanese persons performing AMPS tasks before the initiation of this study. These persons were not included as subjects for this study.

**Data Analysis**

Because the AMPS includes four facets (skill item difficulty, task challenge, rater severity, and person ability), the many-faceted Rasch model (Linacre, 1993) was used to analyze the data. More detailed descriptions of the use of many-faceted Rasch analysis to develop the AMPS have been reported elsewhere (Doble, Fisk, Fisher, Rivvo, & Murray, 1994; Fisher, 1993, 1994; Magalhães et al., 1996; Park, Fisher, & Velozo, 1994).

To determine whether the data met the measurement requirements of the model, they were evaluated through examination of the subject and rater goodness-of-fit statistics calculated by the many-faceted Rasch analysis computer program (FACETS) (Linacre, 1988). The mean square (MnSq) residuals, which are the differences between observed and expected scores, and the standardized score (t) residuals provided the degree to which the examined person's performance was consistent with the model expectations.
which raters and subjects fit the model (Fisher, 1993; Linacre, 1993).

The first question of this study was whether the AMPS was valid for Japanese persons. If the subjects performed better on easy skill items and tasks than on hard skill items and tasks, they would demonstrate good fit to the model, and thus the validity of the AMPS for Japanese persons would be verified. In this case, the model was defined by calibration values derived from the European subjects in the standardization sample for the AMPS (Fisher, 1995a). The fundamental question addressed by this question was whether these calibration values were valid when applied to the Japanese subjects. Therefore, we anchored (preset) the calibration values of tasks, items, and rating scale before running the analysis.

The second question was whether a rater scored consistently across cultures. If all the raters for this study assigned the subjects lower scores on hard skill items as well as higher scores on easy skill items, and if they assigned lower scores to subjects with less ability as well as higher scores to subjects with more ability, they would demonstrate acceptable goodness of fit. That is, rater scoring consistency or interrater reliability would be shown cross-culturally.

Identical criteria were used to judge the goodness of fit of the subjects and the raters. The criteria for fit to the model were based on the goodness-of-fit statistics generated by the many-faceted Rasch analysis: (a) $MSE < 1.4$ and (b) $t < 2$. Higher $MSE$ values combined with significant $t$ values were judged as failure to demonstrate acceptable goodness of fit (Fisher, 1995a; Linacre & Wright, 1994). The criterion for an overall rate of acceptable goodness of fit of subjects was 95% or more (Fisher, 1995a).

We also used a second method recommended for evaluating rater reliability—overall proportion of misfitting ratings for raters (Fisher, 1993, 1994, 1995a). Because an analysis of 672 European adults between 16 and 59 years of age ($M = 34.9$ years; $SD = 10.4$) with no disability revealed an overall rate of misfitting ratings of 2.5%, we set the criterion for an acceptable level of individual misfitting ratings to be less than 2.5% when $t = 3$ (Fisher, 1995b).

The third question was whether raters maintain the same relative severity. It was possible that each rater would demonstrate goodness of fit but become consistently more lenient or more severe relative to the other raters. Therefore, we compared the rater severity calibration values derived from co-scoring the European subjects from the AMPS training workshop to those derived from co-scoring the Japanese subjects. This comparison required that we not anchor (preset) the rater severity calibration values. To test for significant differences in rater severity between the calibration values from the two groups of subjects, we calculated the standardized difference ($z$) between the two estimates (Wright & Masters, 1982). Both sets of estimates were centered such that the mean severity of all six raters was zero.

Results

The FACETS computer program generated the person ability measures and goodness-of-fit statistics for the subjects (see Table 1). To answer the first question of whether Japanese persons fit the measurement model defined by European calibration values, the infit and outfit goodness-of-fit statistics were examined. All 10 subjects demonstrated goodness of fit to the measurement model for the motor scale, and all but one demonstrated goodness of fit for the process scale. Because the overall rate of acceptable goodness of fit was 95% across both scales, we concluded that the subjects fit the AMPS measurement model.

To answer the second question of whether raters from diverse cultural backgrounds demonstrate acceptable goodness of fit when rating the Japanese subjects, the infit and outfit goodness-of-fit statistics were examined (see Table 2). All six raters fit the measurement model. The proportion of individual misfitting ratings was less than 2.5% when $t = 3$. These results suggest that raters from diverse cultural backgrounds score consistently when they rate Japanese subjects.

Finally, to answer the third question of whether raters from diverse cultural backgrounds vary in severity when they rate Japanese subjects, we calculated the standardized difference ($z$) as described previously (see Table 3). The Japanese rater and one American rater varied significantly in rating severity. Relative to the other raters, the Japanese rater became more lenient on both scales and the American rater became more severe on the process scale.

Discussion

The cross-cultural validity of the AMPS was supported despite our finding that one subject did not demonstrate acceptable goodness of fit to the process scale. This study complements previous studies (Besvroe, 1995; Clawson, 1995; Dickerson & Fisher, 1995; Fisher et al., 1992; Magalhães et al., 1996) and offers further support of the AMPS as an assessment that can be used with persons from a wide variety of cultures. Therefore, subject response validity was upheld.

The tasks used in this study were familiar enough to the subjects to allow their appropriate use. Therefore, it seems that as long as the subject indicates familiarity with
the tasks and has had experience performing them, cultural differences can be accommodated in the existing AMPS scoring criteria. For example, the number or kinds of utensils often used differed among the subjects but did not interfere with valid evaluation. As will be discussed later, however, our results suggest that the rater must be somewhat familiar with the subjects’ culture, including how the relevant AMPS tasks are performed within that culture.

Because all raters demonstrated acceptable goodness of fit to the AMPS measurement model, regardless of their country of origin, the interrater reliabilities established previously for European raters (Fisher, 1995a) and rater response validity were supported. It appears that raters who have been trained and calibrated on the AMPS and have received some minimal amount of cross-cultural orientation can be reliable when scoring persons from other cultures as long as they are knowledgeable enough to ascertain what is inefficient behavior and what is accepted cultural deviance.

There was a related finding, however, that deserves additional discussion. The Japanese rater was significantly more lenient on both scales when judging Japanese subjects, whereas one of the American raters was significantly more severe on the process scale. There are several possible explanations. First, the Japanese rater may have been culturally biased. This bias could have occurred partly because some Japanese actions or methods may have been so familiar to this rater that she did not see them as illogical and inefficient by the AMPS criteria, and, therefore, she awarded higher scores than did the European raters. With similar logic, the American rater was perhaps biased toward European behaviors and did not separate from the scoring process what was inefficient behavior on the part of the subjects versus what was appropriate to Japanese culture. For example, several of the raters noted that the subjects turned on the burner and placed a pan on the stove “a long time before cooking anything in the pan.” The Japanese rater did not find this behavior to be unusual as it is common for Japanese people living in the United States to heat the pan for a long time before cooking because, as they state, “American electric burners are so slow to heat up.”

Second, although we conclude that the Japanese rater became significantly more lenient and the American rater became significantly more severe, this is not necessarily the case. Rater calibration values are calculated by the FACETS computer program (Linacre, 1988) such that the mean rater severity is zero. If most or all of the European raters became more severe when rating Jap-

Table 1
Subject Person Ability Measures and Goodness-of-Fit Statistics

<table>
<thead>
<tr>
<th>Subject (logits)</th>
<th>Infit</th>
<th>Outfit</th>
<th>Subject (logits)</th>
<th>Infit</th>
<th>Outfit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 3.26</td>
<td>1.2 0</td>
<td>0.9 0</td>
<td>1 1.74</td>
<td>1.0 0</td>
<td>1.0 0</td>
</tr>
<tr>
<td>2 3.28</td>
<td>1.1 0</td>
<td>0.8 0</td>
<td>2 1.66</td>
<td>0.8 -2</td>
<td>0.8 -1</td>
</tr>
<tr>
<td>3 3.07</td>
<td>1.2 0</td>
<td>1.0 0</td>
<td>3 1.72</td>
<td>1.0 0</td>
<td>1.0 0</td>
</tr>
<tr>
<td>4 2.98</td>
<td>1.1 0</td>
<td>1.0 0</td>
<td>4 1.86</td>
<td>0.9 0</td>
<td>1.3 1</td>
</tr>
<tr>
<td>5 2.89</td>
<td>1.0 0</td>
<td>0.9 0</td>
<td>5 1.77</td>
<td>0.9 0</td>
<td>1.0 0</td>
</tr>
<tr>
<td>6 3.85</td>
<td>0.8 0</td>
<td>0.9 0</td>
<td>6 2.62</td>
<td>0.9 0</td>
<td>1.1 0</td>
</tr>
<tr>
<td>7 3.45</td>
<td>1.3 0</td>
<td>0.9 0</td>
<td>7 2.86</td>
<td>1.1 0</td>
<td>1.2 1</td>
</tr>
<tr>
<td>8 3.67</td>
<td>1.1 0</td>
<td>0.8 0</td>
<td>8 2.13</td>
<td>1.1 0</td>
<td>1.1 0</td>
</tr>
<tr>
<td>9 3.37</td>
<td>1.1 0</td>
<td>0.9 0</td>
<td>9 2.02</td>
<td>1.3 1</td>
<td>1.6 2</td>
</tr>
<tr>
<td>10 2.76</td>
<td>1.1 0</td>
<td>1.1 0</td>
<td>10 1.61</td>
<td>1.0 0</td>
<td>1.1 0</td>
</tr>
</tbody>
</table>

Subject who misfit. Note. MnSq = mean square.

Table 2
Rater Goodness-of-Fit Statistics

<table>
<thead>
<tr>
<th>Rater</th>
<th>Infit</th>
<th>Outfit</th>
<th>Rater</th>
<th>Infit</th>
<th>Outfit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infit</td>
<td>MnSq t</td>
<td></td>
<td>Outfit</td>
<td>MnSq t</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>1.2 0</td>
<td>0.9 0</td>
<td>1.0 0</td>
<td>1.3 2</td>
<td></td>
</tr>
<tr>
<td>American 1</td>
<td>1.2 0</td>
<td>0.9 0</td>
<td>1.0 0</td>
<td>1.1 0</td>
<td></td>
</tr>
<tr>
<td>American 2</td>
<td>1.1 0</td>
<td>0.9 0</td>
<td>0.9 1</td>
<td>0.9 1</td>
<td></td>
</tr>
<tr>
<td>American 3</td>
<td>1.1 0</td>
<td>1.0 0</td>
<td>1.1 0</td>
<td>1.1 0</td>
<td></td>
</tr>
<tr>
<td>Swedish</td>
<td>1.0 0</td>
<td>0.8 0</td>
<td>0.9 0</td>
<td>0.9 0</td>
<td></td>
</tr>
<tr>
<td>British</td>
<td>1.0 0</td>
<td>0.8 0</td>
<td>1.0 0</td>
<td>1.1 0</td>
<td></td>
</tr>
</tbody>
</table>

Note. MnSq = mean square.
Table 3
Comparison of Rater Calibrations

<table>
<thead>
<tr>
<th></th>
<th>European</th>
<th>Japanese</th>
<th></th>
<th>European</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rater</td>
<td>Severity (logits)</td>
<td>SE</td>
<td>Rater</td>
<td>Severity (logits)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(logits)</td>
<td></td>
<td></td>
<td>(logits)</td>
</tr>
<tr>
<td>Japanese</td>
<td>-.07</td>
<td>.09</td>
<td>.41</td>
<td>.14</td>
<td>-.48</td>
</tr>
<tr>
<td>American</td>
<td>.04</td>
<td>.06</td>
<td>.08</td>
<td>.13</td>
<td>-.04</td>
</tr>
<tr>
<td></td>
<td>.00</td>
<td>.06</td>
<td>.20</td>
<td>.13</td>
<td>-.20</td>
</tr>
<tr>
<td>American3</td>
<td>-.13</td>
<td>.06</td>
<td>.15</td>
<td>.15</td>
<td>.02</td>
</tr>
<tr>
<td>Swedish</td>
<td>.12</td>
<td>.06</td>
<td>-.09</td>
<td>.15</td>
<td>.21</td>
</tr>
<tr>
<td>British</td>
<td>.02</td>
<td>.06</td>
<td>-.05</td>
<td>.15</td>
<td>.07</td>
</tr>
</tbody>
</table>

Raters who changed significantly in severity.

Japanese subjects, the Japanese rater would appear to become more lenient when, in fact, she may not have varied in overall severity at all. To determine whether the Japanese rater became more lenient or the European raters became more severe, we examined the raw data for the Japanese subjects and for a similar subset of European subjects drawn from the AMPS database (Fisher, 1995b). Among the European subjects, the majority had between 2 and 4 raw score points deducted from the total possible for the AMPS motor scale and between 4 and 8 points deducted from the total possible for the AMPS process scale. Compared with the Japanese data, 33% of the total raw motor scores and 75% of the total raw process scores assigned by the European raters were lower than commonly found among European subjects rated by European raters. In contrast, none of the Japanese rater's total raw motor scores and only 20% of the total raw process scores were lower than commonly found among European subjects rated by European raters. This finding suggests that there may have been a European bias present when the European raters rated the Japanese subjects.

The conclusion that the European raters became more strict was supported when we compared the mean of the person ability measures of the 672 European adults without disability (Fisher, 1995b) to that of our subjects. The mean person ability measures of the European subjects was 3.66 logits (SD = .5) for the motor scale and 2.15 logits (SD = .56) for the process scale (Fisher, 1995b). The mean of the person ability measures of the Japanese subjects was lower: 3.26 logits (SD = .33) for the motor scale and 1.90 logits (SD = .29) for the process scale. Because of the unequal sample sizes and lack of homogeneity of variance, we used two-tailed Welch's t tests (Shott, 1990) and found a significant difference on both: the motor scale, t(9) = 3.77, p < .05, and the process scale, t(10) = 2.67, p < .05.

Another possible explanation for the lower mean AMPS person ability measures for the Japanese subjects may be that the setting or environment may have put the subjects at a disadvantage. Previous studies have demonstrated that environmental familiarity affects a person's AMPS process skill ability measures (Nygard, Bernspang, Fisher, & Winblad, 1994; Park et al., 1994). The IADL task performances of persons living in the community tend to be better when they are in a familiar setting (Park et al., 1994). Although the Japanese subjects performed familiar and cultural-relevant tasks in their homes, they were not in a fully familiar setting. They all performed the tasks in American kitchens with American tools and equipment. Most of the subjects were living alone or with their families and, consequently, were not familiar with American tools and equipment. It is possible that the European raters tended to assign lower scores to the subjects when they used American-style objects unusually or ineffectively. The Japanese rater may not have lowered the subjects' scores because she shared the same lack of familiarity with American tools and equipment.

Finally, the Japanese subjects could have simply been less able than the European subjects, which seems unlikely. Most of the subjects were studying at an American university and were, therefore, probably very capable persons in their own culture.

Considered together, these findings support the possibility that some cultural bias could have affected the results. Even though the European raters were reliable and had received some training on Japanese cultural traits, they still may have been unable to recognize performance that was logical or efficient, not better or worse, but different from their own.

Limitations

One limitation of this study was that there was only one Japanese rater. As a result, we could not determine whether she only became more lenient relative to the European raters or whether this would be a common tendency among all Japanese raters. Another limitation was that our
subjects only performed European tasks in American kitchens with American tools and equipment. All these factors confounded our ability to differentiate between rater bias and familiarity of the environment and their effect on rater scoring and subject performance. Additionally, because of the timing of the study, there was no opportunity to further train the European raters about Japanese culture. Because of these limitations, we could not be certain that the finding that the Japanese rater and the one American rater varied significantly in their rating severity was caused by the European rater's limited familiarity with Japanese culture or the Japanese subject's limited familiarity with American kitchens. Finally, this study included only young Japanese subjects with no known disabilities. Thus, these findings should not be generalized to other populations.

Recommendations for Further Research

Although the results of this study supported the cross-cultural usefulness of the AMPS with Japanese persons, further research clearly is indicated. First, to clarify whether cultural unfamiliarity affects rater severity, we recommend further study with (a) a larger number of Japanese raters, (b) Japanese subjects in Japan performing European tasks in their homes, and (c) European raters more thoroughly trained to recognize variations in Japanese culture. Second, we recommend that there be further examination of the feasibility of adding Asian culture-specific tasks to the AMPS to provide a broader task choice. Third, cross-cultural interrater reliability studies with Asian persons performing Asian tasks are recommended to confirm the usefulness of the AMPS as a cross-cultural IADL assessment. Fourth, the previous three recommendations could be modified and applied to comparable studies of subcultural groups within any country that has a multicultural composition. Finally, to further the standardization of the AMPS for Japanese persons, we recommend that a larger sample be evaluated and that other populations, such as persons with disabilities, be included.

Application to Occupational Therapy

With increasing cultural diversity within the United States and other countries, this study provides additional support for the development of the AMPS as a cross-cultural instrument for the measurement of IADL ability (Bennett, 1995; Clawson, 1995; Dickerson & Fisher, 1995; Fisher et al., 1992; Magalhães et al., 1996). In particular, our results suggest that the AMPS will be valid and reliable when used with diverse clients as long as they are tested in familiar environments by raters familiar with their specific cultures. Although the exact reasons for our results can only be determined through further research, it is clear that cultural differences can affect IADL task performance or the validity of the scores. Therefore, whenever a rater observes and scores the performance of a client from a different culture, the rater must (a) ensure that the client is fully familiarized with the tasks and equipment to be used during the evaluation and (b) be familiar enough with what is considered logical, efficient, and usual by the client's cultural group. It is critical that raters are knowledgeable enough to recognize legitimate cultural variations in task performance. Studies of African-Americans, Mexican-Americans, and Cuban-Americans indicate that cross-cultural rater severity is not a factor within a country; unlike the results of this study, none of these studies found a significant difference in the AMPS IADL motor or process ability between African-American, Mexican-American, or Cuban-American and matched Anglo-American control groups (Bennett, 1995; Clawson, 1995; Dickerson & Fisher, 1995).

Occupational therapists can benefit from a cross-cultural instrument that has relatively few tasks that need to be performed by the client, allows the client to choose the task and control how it is done, and takes into consideration the rater's severity as well as the client's performance. This particular study represents an important development in the application of the AMPS to Japanese and other Asian populations.

Summary

Both the subjects and raters demonstrated expected response patterns as indicated by goodness of fit to the Rasch model. The results indicate that the AMPS is valid and reliable when applied to Japanese subjects. There is, therefore, support for the hypothesis that the AMPS can be used as a cross-cultural IADL assessment (Bennett, 1995; Clawson, 1995; Dickerson & Fisher, 1995; Fisher et al., 1992; Magalhães et al., 1996). Further study is needed to clarify the issue of possible cultural bias in rater severity.

Acknowledgments

We thank Birgitta Bernspang, DeMois, Glenn Stewart, Jean Cannella, MOT, and Susan Goldman, MS, for participating as raters. This study was completed in partial fulfillment of the first author's requirements for a master of science degree at Colorado State University. This study was supported in part by a grant from the National Institutes of Health, National Institute on Aging (Grant no. 1R01-AG12345-3).

References


