Volume 3, Number 3, 1999 ISSN 1090-655X

Journal of Outcome Measurement®

Dedicated to Health, Education, and Social Science

REHABILITATION FOUNDATION INC.
Est. 1993
Research & Education

JOURNAL OF OUTCOME MEASUREMENT®

Volume 3, Number 3 1999

Articles

Indexing/Abstracting Services: JOM is currently indexed in the *Current Index to Journals in Education* (ERIC), *Index Medicus*, and MEDLINE. The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

JOURNAL OF OUTCOME MEASUREMENT®, 3(3), 200–215

The Impact of Socio-cultural and Clinical Factors On Health-related Quality of Life Reports Among Hispanic and African-American Cancer Patients

George J. Wan
Merck & Co., Inc., West Point, PA
Michael A. Counte
Saint Louis University, St. Louis, MO
David F. Cella
Rush-Presbyterian-St. Luke's Medical Center, Chicago, IL
Lesbia Hernandez
University of Puerto Rico, San Juan, PR
Deborah B. McGuire
Emory University, Atlanta, GA
Shirley Deasy
Emory University and Grady Memorial Hospital, Atlanta, GA
Gail Shiomoto
Cook County Hospital, Chicago, IL
Elizabeth A. Hahn
Rush-Presbyterian-St. Luke's Medical Center, Chicago, IL

A hierarchical multiple linear regression approach (N = 761) was used to identify pertinent factors which influence health-related quality of life (HRQL) reports among Hispanic and African-American cancer patients. The independent variables include: performance status, disease site, disease stage, mode of administration, socio-economic status (SES), gender, age, living arrangement, race/ethnicity, religious affiliation, insurance status, and spiritual beliefs. The outcome measures, five subscales of HRQL (physical well-being, social well-being, satisfaction with treatment, emotional well-being, functional well-being) and overall HRQL (sum of the five subscales), were estimated using the Functional Assessment of Cancer Therapy (FACT) Scales. This study identified performance status and spiritual beliefs as consistent predictors of overall HRQL. This study also found no significant effects of SES, mode of administration, gender, age, living arrangement and insurance status on the reporting of overall HRQL. Spiritual beliefs and performance status are important determinants of HRQL across a diverse group of cancer patients.

Requests for reprints should be sent to George J. Wan, Ph.D., M.P.H., Merck & Co., Inc., P.O. Box 4, WP39-170, West Point, PA 19486. E-mail:george_wan@merck.com.
Support for this project was provided in part by Grant #5 R01 CA61679 from the National Cancer Institute, David F. Cella, Ph.D., Principal Investigator. The opinions stated in this manuscript are those of the authors and do not represent those of Merck & Co., Inc.

Introduction

Although health-related quality of life (HRQL) in cancer patients has been modeled in numerous studies, ironically few race or ethnic-based analyses have been reported. Thus, very little is known regarding the appraisal of quality of life among minority populations. Moreover, minority populations such as Hispanics and African-Americans are often even excluded from HRQL clinical trials due to differences in reporting style, acculturation, language and comprehension levels (Cella and Bonomi, 1996; Cella, et al., 1996). Hence, there is a clear need for greater participation of minority and disadvantaged populations in the development of HRQL assessments. The improvement of culturally sensitive questionnaires is also necessary for accurately measuring HRQL across different ethnic groups. Thus, the primary purpose of this paper was to examine the impact of selected clinical and socio-cultural variables on Hispanic and African-American cancer patients' reports of HRQL.

HRQL is defined as the gap between an individual's expectations and actual experience (Calman, 1984; Calman, 1987; Cella, et al., 1993). Thus, a patient will likely report a higher level of HRQL if their actual functioning is better than or as expected. In contrast, an individual will likely report a lower rating of HRQL if his or her personal experience is worse than expected (Calman, 1984; Calman, 1987).

HRQL is typically conceptualized as a multidimensional construct with four distinct domains: physical, social, emotional and functional well-being (Cella, et al., 1993; Hays and Stewart, 1990; Schipper, et al., 1984; Stewart, et al., 1981; Yancik, et al., 1987). Physical well-being refers to the presence of physical symptoms related with therapy such as fatigue, nausea, pain or treatment side-effects. Social well-being describes an individual's relationship with friends and family members. Emotional well-being pertains to the presence of psychological distress such as nervousness, depression or anxiety. Functional well-being includes the ability to perform daily activities at work or at home.

It is also important to note that since HRQL is evaluated by the individual, it is inherently subjective. Studies in mostly Caucasian patients have found that reporting of HRQL can be affected by one's age (Cassileth, et al., 1984; Ganz, et al., 1985; Rodin, 1986; Wan, et al., 1997a) or gender (Ferrell, et al., 1995; Ganz, 1993; Myken, et al., 1995; Shag and Heinrich, 1988). Specifically, older adults tend to report better HRQL

than younger adults and women tend to report worse HRQL than men. Possible explanations for these differences include variations in their personal expectations. For instance, Wan, et al., (1997a) found lower physical and emotional well-being scores among younger adults. In addition, they found that females reported lower scores in the areas of emotional and functional well-being compared to men. Several studies in mostly Caucasian patients have evaluated the relationship between performance status (activity level rating), spiritual beliefs (Fitchett, et al., 1996) and HRQL (Cassileth, et al., 1984; Ganz, et al., 1985; Cella, et al., 1987; Cella, et al., 1993; Wan, et al., 1997a). Other studies have examined the association between disease site, disease stage and HRQL (Stewart and Van Dam, 1983; Llewellyn-Thomas, et al., 1984; Aaronson, et al., 1986; Aaronson and Beckman, 1987; Wan, et al., 1997a).

It is uncertain whether ethnic HRQL differences actually exist. Hispanic HRQL responses may be influenced by Hispanic values such as "simpatia" (need for harmonious social arrangements), familialism and power distance (respect for social power) (Marin and Marin, 1991). African-American HRQL responses may be affected by African-American cultural values that support the strong role of church, familialism, unity, collective community responsibility and self-determination (Billingsley, 1968; Brisbane and Womble, 1992). More cross-cultural research is needed to explore these differences within and across minority populations.

Hypotheses

Prior studies have found that activity level (Cella, et al., 1987; Cella, et al., 1993; Wan, et al., 1997a) is strongly related to HRQL. However, relationships between performance status, spiritual beliefs and HRQL among African-Americans and Hispanics are largely unknown. Thus, these factors need to be further examined. No specific hypotheses were offered regarding their association with HRQL.

Based on a previous study by Wan, et al., 1997a, it was expected that older patients and those living with others would indicate higher overall HRQL. The associations between insurance status, disease stage and HRQL remain unclear. Numerous empirical studies have established a relationship between being uninsured and reporting poorer health (Freeman, et al., 1987; Freeman, et al., 1990; Hahn and Flood, 1995; Herndershot, 1988; Monheit, et al., 1985). However, few reported studies have examined the relationship between insurance status and HRQL.

One exception is Wan, et al., (1997a) who reported higher levels of satisfaction with treatment, emotional well-being, functional well-being and overall HRQL among those cancer patients enrolled in managed care plans. They also found lower functional well-being among more severely staged cancer patients (i.e., patients with more advanced disease).

Methods

Subjects

This study is cross-sectional in design. A total of 761 Hispanic and African-American individuals were enrolled from each of the following Eastern Cooperative Oncology Group (ECOG) institutions from May 1994 to November 1995: Grady Memorial Hospital Minority Community Clinical Oncology Program (CCOP)/Emory University Hospital, Atlanta (N=161), San Juan Minority CCOP, Puerto Rico (N=375) and Rush-Presbyterian-St. Luke's Medical Center/Cook County Hospital, Chicago (N=212). Thirteen cases were obtained from other sites. Eligible participants met the following criteria: 17 years of age, diagnosed with breast cancer (39.3%), lung cancer (19.2%), colon cancer (21.9%) or head/neck cancer (19.6%). To ensure sufficient experience with cancer treatment, respondents also had to have completed a minimum of two cycles of chemotherapy treatment or 10 radiation therapy treatments.

Independent Variables

Independent variables were classified into three categories: clinical, mode of administration and socio-cultural characteristics. Reference cell coding (Kleinbaum, et al., 1988) was used for all dummy variables ($k - 1$ for dummy variables, where k = categories). Thus, reference groups were assigned for some of the dummy variables that were created. Binary variables with the coding no = 0; yes = 1 were created for disease sites, mode of administration, religion and insurance coverage. Clinical measures included performance status, disease site and disease stage. Performance status rating (PSR) is a patient's appraisal of activity level. It was originally developed by the ECOG (Zubrod, et al., 1960). The range of scores is from 0 to 4, where 0 = normal activity and 4 = bedridden. Due to their small number, 11 patients who were rated "4" on the PSR were included with individuals rated "3" (3 = requiring bed rest greater than 50% of the waking day). Binary variables were created for the following disease sites: breast, colon, lung and head/neck cancer as the reference group.

Disease stage was determined by the patient's oncologist and then later corroborated by a nurse reviewer (in Puerto Rico the diagnosis and extent of disease were extracted from the medical record by the interviewer and then later corroborated by the oncologist) . Staging is based on the National Cancer Institute (NCI) criteria which include in situ (noninvasive malignancy), local (invasive cancer confined to the site of origin), regional (spread by direct extension to adjacent organs or to regional lymph nodes) and distant metastases (spread to distant organs or lymph nodes) (Riley, et al., 1994).

Dummy variables were created for mode of administration [High literacy, Randomized to Interview (RI), High literacy, Randomized to Self Administration (RS), Low literacy Assigned to Interview (AI) as the reference group]. Patients who scored at a reading level of 6th grade or higher on the Woodcock Language Proficiency Battery-Passage Comprehension (WLPB-PC) subtest (Woodcock, 1991) were randomly assigned to either an interview (N = 263; 38%) or self-administration (N = 239; 35%) and participants who scored below the 6th grade reading level were assigned to be interviewed (N = 183; 27%). Low literacy was operationalized as a reading level below the 6th grade. Socio-cultural factors include: socio-economic status (SES), gender, age, living arrangement, religion, insurance status and spiritual beliefs. SES was measured with the Hollingshead Two-Factor Index of Social Position (SES), a five level ordinal scale. It is based on the sum of the occupational factor weighted by five and the educational factor weighted by three (Hollingshead, 1975). The educational and occupational factors are further described in the Hollingshead (1975) handbook. Gender (Male = 0; Female = 1), living arrangement (Lives alone = 0; Lives with other (s) = 1), race/ethnicity (African-American = 0; Hispanic = 1), religion (Catholic vs. Protestant vs. No Affiliation, Other Religion as the reference group) and insurance coverage (Medicaid vs. Medicare only vs. Medicare plus supplemental vs. Private Insurance vs. No Insurance, Other Insurance as reference group) were coded as dummy variables. The Other Insurance category includes beneficiaries of disability insurance (N=3), enrollees in HMOs or other limited providers (N=20) and other insurance carriers (N=56). Spiritual beliefs were measured using a content-valid set of 12 items developed within the Functional Assessment of Cancer Therapy (FACT) Quality of Life Measurement System . An aggregate index was computed by summing responses to the 12 items which measure an individual's sense of meaning or purpose in life (i.e., "I feel a sense of

purpose in my life"), strength of faith (i.e., "I feel strength in my faith") and optimism/self-efficacy (i.e., "I know that whatever happens with my illness, things will be okay"). Negatively worded items were reverse scored, so that a higher score indicates stronger spiritual beliefs. Since there is no normative standard for assessing spiritual beliefs, content validity was demonstrated by expert panel assessment and by review of the literature. Internal consistency using Cronbach's alpha was 0.80 in this sample of cancer patients. A brief psychometric analysis of this FACT subscale has been presented (Fitchett, et al., 1996) and a manuscript is in preparation.

Dependent Variables

Five outcome HRQL scores and an overall HRQL index were estimated using the Functional Assessment of Cancer Therapy-General (FACT-G) Quality of Life Measurement System (Cella, 1994). The FACT-G is a reliable and valid instrument for evaluating HRQL in cancer and HIV patients (Cella, et al., 1993; Cella and Bonomi, 1996). The FACT-G has been adapted for use with Spanish-speaking cancer patients, including those with low literacy (Bonomi, et al., 1996; Cella, et al., 1998). The Spanish language version of the FACT-G was translated, reviewed, revised, and then pretested with 92 Spanish-speaking patients from the Central U.S. and Puerto Rico (Cella, et al., 1998). Furthermore, semantic, content, and partial technical equivalence were established using an iterative forward-backward-forward translation sequence, followed by an expert bilingual panel review, and pretesting interviews with 92 Spanish-speaking patients (Cella, et al., 1998). The instrument contains 34 items (only 28 items are scored) in a five point Likert type format (Cella, 1994). An overall HRQL index was calculated by summing the five subscales of HRQL (physical well-being, social well-being, emotional well-being, functional well-being, relationship with doctor). Negatively phrased items were reversed scored, with a high rating reflecting a high score.

Statistical Procedures

SPSS V6.0 for Windows (Norusis, 1993) was the software package used for all statistical procedures. Frequency distributions of the sample are shown in Table 1. Correlation coefficients and collinearity diagnostics were performed in order to detect the presence of any serious multicollinearity problems among the independent variables. After ex-

Table 1

*Sample Characteristics**

Variable	Number	Percent (%) or Range
Performance Status Rating		
0=Normal activity	253	33.6%
1=Some symptoms, but I can walk and I don't spend extra time in bed	248	32.9%
2=Need some time in bed, but it is less than 50% of normal daytime	175	23.2%
3=Need to be in bed greater than 50% of normal day**	78	10.3%
Disease Stage		
1=In Situ	174	23.0%
2=Local	213	28.2%
3=Regional	229	30.3%
4=Distant	140	18.5%
Survey Characteristics		
1=Randomized to Interview (RI)	263	38.4%
2=Randomized to Self-Admin. (RS)	239	34.9%
3=Assigned to Interview (AI) (ref. group)	183	26.7%
Hollingshead Two-Factor Index (SES)	M=2.3; SD=1.1	1-5
I=Lower Class (8 to 19)	186	24.8%
II=Working Class (20 to 29)	304	40.5%
III=Lower-Middle Class (30 to 39)	127	16.9%
IV=Middle Class (40 to 54)	98	13.1%
V=Upper Class (55 to 66)	35	4.7%
Gender		
0=Male	313	41.1%
1=Female	448	58.9%
Age (in years)	M=57.2; SD=12.4	17-90
Living Arrangement		
0=Lives Alone	123	16.2%
1=Lives with Other (s)	638	83.8%
Race/Ethnicity		
0=African-American	320	42.0%
1=Hispanic	441	58.0%
Religion		
Catholic	333	43.8%
Protestant	207	27.2%
Other (reference group)	185	24.3%
No Affiliation	36	4.7%
Insurance Status		
Medicaid	208	28.2%
Medicare	96	13.0%
Medicare+supplemental	113	15.3%
Private	137	18.6%
No Insurance	104	14.1%
Other Insurance (reference group)***	79	10.7%

* Some categories do not add up to 761, because of missing data.

** 11 patients who were rated "4" on the PSR ("Unable to get out of bed") were included with patients rated "3".

*** Includes disability insurance (N = 3), HMO or other limited provider plan (N = 20) and other insurance (N = 56).

M = mean; SD = standard deviation.

amining the pairwise coefficients of simple correlations among the independent variables, multicollinearity did not appear to be a problem. In addition, variance inflation factors (VIF) for each of the independent variables entered into the six regression equations indicate the absence of serious multicollinearity problems (the VIF computed for the independent variables were all less than 10). Hierarchical multiple linear regression was used in order to examine the relationship between the independent variables and the six outcome measures of HRQL. The predictor variables were entered into three blocks using the forced-entry method. The order of entry was as follows: 1) clinical (PSR, disease site, extent of disease), 2) mode of administration (RI, RS) and 3) socio-cultural variables (SES, gender, age, living arrangement, race/ethnicity, religion, insurance status, spiritual beliefs). Due to space constraints, only the standardized regression coefficients are reported. Unstandardized regression coefficients do not suggest any different conclusions and are available upon request from the primary author.

Results

Table 1 summarizes the attributes of the sample. Approximately 67% of the participants indicated being fully ambulatory with or without symptoms (PSR = 0 or 1). The mean age of the respondents was 57 years old. The cancer patients were predominately female (59%), Hispanic (58%), Catholic (44%) and living with other(s) (84%). Sixty-five percent of the participants were classified as either working or lower class (SES categories). Furthermore, 28% were enrolled in Medicaid, 13% were Medicare recipients, 19% had private insurance and 14% did not have any insurance.

Multivariate Regression Analyses of HRQL Dimensions

The explanatory variables explained 40% of the variation in patients' reports of physical well-being, 22% of the variation in social well-being, 10% of the variation in the satisfaction with treatment, 32% of the variation in emotional well-being and 42% of the variation in functional well-being. All five models were significant at the $p < .001$ level (Table 2).

Clinical variables, namely performance status, lung cancer diagnosis and extent of disease, were significant determinants of certain subscales of HRQL. Participants with poorer performance status reported lower physical well-being, emotional well-being and functional well-being. Lung

Table 2

Summary of Multivariate Regression Analyses

Variables	Physical WB	Social WB	RWD	Emotional WB	Functional WB	Overall HRQL
Clinical Factors						
Performance Status	**-.489 (.000)**	-.016 (.678)	.049 (.240)	**-.149 (.000)**	**-.300 (.000)**	**-.341 (.000)**
Breast	.022 (.696)	.020 (.755)	-.003 (.966)	-.035 (.562)	-.053 (.343)	-.016 (.751)
Colon	-.050 (.236)	.078 (.107)	.033 (.523)	.006 (.897)	-.048 (.256)	-.010 (.801)
Lung	-.069 (.098)	.013 (.775)	-.017 (.737)	-.001 (.982)	**-.118 (.004)**	-.071 (.063)
Disease Stage	**-.136 (.000)**	-.039 (.317)	-.004 (.927)	-.076 (.036)	-.044 (.194)	**-.098 (.002)**
Mode of Administration						
RI	-.003 (.951)	.021 (.664)	-.085 (.101)	-.068 (.132)	-.048 (.249)	-.038 (.327)
RS	.049 (.256)	-.078 (.110)	-.031 (.554)	-.076 (.098)	-.061 (.149)	-.049 (.208)
Socio-cultural Factors						
SES	.059 (.115)	.062 (.146)	.016 (.720)	**.105 (.008)**	-.005 (.897)	.066 (.052)
Gender	.004 (.936)	.052 (.324)	-.079 (.165)	-.016 (.748)	.081 (.076)	.040 (.343)
Age	**.120 (.001)**	-.061 (.143)	-.040 (.369)	**.131 (.001)**	.028 (.431)	.062 (.060)
Living Arrangement	-.035 (.285)	**.163 (.000)**	-.042 (.299)	-.014 (.698)	.043 (.183)	.052 (.085)
Race/Ethnicity	.074 (.131)	-.002 (.966)	**-.130 (.031)**	-.032 (.537)	-.077 (.111)	-.029 (.515)
Catholic	.083 (.145)	.091 (.163)	.078 (.264)	.045 (.458)	.097 (.086)	**.120 (.021)**
Protestant	-.006 (.893)	.051 (.310)	.055 (.302)	.040 (.390)	.007 (.868)	.036 (.367)
No Affiliation	.035 (.321)	.014 (.730)	-.079 (.068)	.052 (.171)	-.017 (.629)	.014 (.673)
Medicaid	.063 (.236)	-.036 (.556)	-.039 (.551)	**.141 (.013)**	-.024 (.650)	.030 (.530)
Medicare	.071 (.122)	.006 (.904)	.051 (.368)	.038 (.437)	.043 (.345)	.064 (.124)
Medicare+suppl	.046 (.333)	.042 (.441)	.014 (.812)	**.107 (.034)**	-.026 (.585)	.045 (.300)
Private Insurance	.016 (.742)	.013 (.820)	.074 (.228)	.058 (.277)	.054 (.272)	.054 (.238)
No Insurance	.068 (.157)	-.039 (.474)	-.018 (.757)	.034 (.513)	-.032 (.503)	.007 (.865)
Spiritual Beliefs	**.195 (.000)**	**.369 (.000)**	**.228 (.000)**	**.485 (.000)**	**.455 (.000)**	**.507 (.000)**
Intercept	**13.9 (.000)**	**8.63 (.000)**	**5.30 (.000)**	**3.81 (.003)**	1.06 (.646)	**32.8 (.000)**
R^2	.40	.219	.098	.321	.419	.508
Adj R^2	.38	.192	.066	.298	.399	.490
Model F	**19.1 (.000)**	**7.99 (.000)**	**3.10 (.000)**	**13.5 (.000)**	**20.5 (.000)**	**29.3 (.000)**

Note: P-values are shown in parentheses next to each standardized regression coefficient; Due to space constraints, only the standardized regression coefficients were reported; Data on the unstandardized regression coefficients are available upon request from the author; WB = Well-Being; RWD = Relationship with Doctor.

cancer patients rated lower levels of functional well-being compared with other disease sites. Individuals with more advanced disease indicated poorer physical well-being and emotional well-being. Mode of administration was an insignificant predictor of all five dimensions of HRQL.

Although no statistically significant differences were found for gender, other explanatory variables were associated with specific HRQL subscales. Individuals with higher SES reported higher emotional well-being. Younger adults indicated poorer physical and emotional well-being compared to older adults. Those who live with other(s) reported higher levels of social well-being. Hispanics reported lower levels of functional well-being and satisfaction with treatment than African-Americans. Respondents who indicated stronger spiritual beliefs had higher scores on all five subscales of HRQL.

Regression Analysis of Overall HRQL

The complete set of predictor variables explained 51% of the total variance in overall HRQL. The model was significant at the $p < .001$ level. Individuals with worse ratings of activity level and advanced disease stage indicated lower overall HRQL. In addition, Catholics and respondents with stronger spiritual beliefs reported higher HRQL. Stepwise regression was used to determine the total proportion of variance explained by the statistically significant predictors of overall HRQL in the previous regression analysis. PSR, extent of disease, Catholic affiliation and strength of spiritual beliefs explained 50% of the variability in patients' rating of overall HRQL.

Discussion and Conclusion

There is a growing need to evaluate cultural, ethnic and religious variables associated with an individual's perception of HRQL (Warnecke, et al., 1996), especially since minority populations are often times excluded from HRQL assessment studies. Moreover, the inclusion of patients from diverse cultural and socioeconomic backgrounds is necessary to achieve cultural equivalence in the assessment of HRQL in clinical trials (Baker, et al., 1996; Warnecke, et al., 1996).

Despite the need for more cross-cultural HRQL studies, HRQL among culturally diverse populations have not been widely reported. The purpose of this study was to investigate the relationship between select clinical, method of administration, socio-cultural variables and the reporting of HRQL among Hispanic and African-American cancer patients.

Across the ethnic subgroups, individuals with stronger spiritual beliefs consistently report higher HRQL including its component dimensions. One explanation is that stronger spiritual beliefs may enhance emotional well-being and subsequently improve overall HRQL. Another possible explanation is that stronger spiritual beliefs may be inversely related to anxiety. Kaczorowski (1989) substantiates this hypothesis by noting that cancer patients with stronger spiritual well-being have lower levels of anxiety. Spiritual beliefs may affect HRQL by decreasing the disparity between an individual's personal expectations and actual functioning. Finally, spiritual beliefs may also reflect stronger personal control (i.e., self-esteem, anxiety, self-efficacy). This could account for the strong correlation between spiritual beliefs and HRQL.

Observed relationships between performance status (Cella, et al., 1987; Cella, et al., 1993; Wan, et al., 1997a), spiritual beliefs (Fitchett, et al., 1996) and overall HRQL support previous findings. Both clinical factors (PSR, disease site, disease stage) and cultural factors (spiritual beliefs) are important predictors of HRQL subscales. In contrast, survey characteristics (RI vs. RS), SES, insurance status, gender, age and living arrangement were not significantly related to overall HRQL.

The findings are also consistent with previous results that indicate lower social well-being among persons living alone and lower physical and emotional well-being among younger adults (Wan, et al., 1997a). This study is limited because it evaluates an individual's rating of HRQL at only one single point in time. Future studies should examine a patient's report of HRQL over multiple time periods since HRQL may change over time. Additionally, these findings may not be generalizable to the general population since the sample is based on African-American and Hispanic (mainly Puerto Rican) reports of HRQL. After adjusting for clinical and socio-cultural factors in the regression analyses, this study found that Hispanics indicated lower ratings of the doctor/patient relationship and functional well-being compared to African-Americans. One of the reasons why Hispanics have reported lower satisfaction with treatment scores may be due to the institutional setting where the study was carried out (i.e., 375 of the 441 Hispanic patients were admitted to hospitals which serve a predominately medically indigent population). It would be of interest to evaluate satisfaction with treatment and compare different health care institutions in Puerto Rico in a future study. Other possible explanations for these disparities could be attributed to the differences in the re-

porting of performance status, SES, disease site, religious affiliation and living arrangement among Hispanics and African-Americans. Lower SES found among different ethnic groups might be explained by differences related to institutional settings (i.e., individuals with higher SES likely prefer private physician offices and private hospitals). Future studies should examine differences in the appraisal of HRQL among Hispanics and African-Americans born in the U.S. versus those born elsewhere. Further, "Hispanic" is a broad term used to classify persons of Mexican, Puerto Rican, Cuban, Central American, South American or other Spanish cultural descent. Differences and similarities among Hispanics have been previously described in the literature (Molina, et al., 1994) and these Hispanic subgroups should be considered in future studies on HRQL.

In conclusion, the findings indicate that cultural factors such as spiritual beliefs and clinical measures such as performance status are important determinants of HRQL reports across a diverse group of cancer patients. These findings are consistent with other studies which evaluate the relationship between performance status, spiritual beliefs and HRQL in predominately Caucasian patients (Cassileth, et al., 1984; Ganz, et al., 1985; Cella, et al., 1987; Cella, et al., 1993; Fitchett, et al., 1996; Wan, et al., 1997a). Although spiritual beliefs and HRQL are distinctly different concepts it is possible that some of the items which empirically measure HRQL are related to the items which measure spiritual beliefs. Further psychometric analysis is necessary to determine if such item-content bias truly exists.

Clinicians and researchers need to recognize the importance of other psychosocial characteristics which may influence Hispanic and African-American cancer patients' appraisal of HRQL. These factors can include personal expectations (Calman, 1984; Calman, 1987) or coping strategies of cancer survivors (Halstead & Fernsler, 1994; Jenkins and Pargament, 1988; Lazarus, 1980). Perhaps the treatment of cancer patients should also focus on the importance of pastoral care as well as nondenominational religious services addressing spiritual concerns which may help cancer patients cope with their illness and improve quality of life.

References

Aaronson, N. K. and Beckman, J. (1987). *Quality of Life of Cancer Patients, Monograph Series of the European Organization for Research on Treatment of Cancer (EORTC), Vol., 17.* Raven Press, New York.

Aaronson, N. K., DaSilva, F. C., Yoshida, O., Van Dam, F. S. A. M., Fossa, S. D., Miyakawa, M., Raghaven, D., Riedl, H., Robinson, M. R. G., and Worden, J. W. (1986). Quality of life assessment in bladder cancer clinical trials: Conceptual, methodological and practical issues. In: *Developments in Bladder Cancer*, Alan R.Liss, Inc., New York.

Baker, F., Jodrey, D., Zabora, J., Douglas, C. and Fernandez-Kelly, P. (1996). Empirically selected instruments for measuring quality-of-life dimensions in culturally diverse populations. *Journal of the National Cancer Institute Monographs, 20*, 39-47.

Billingsley, A. (1968). *Black Families in White America*, Prentice Hall, Englewood Cliffs, NJ.

Bonomi, A. E., Cella, D. F., Hahn, E. A., Bjordal, K., Sperner, B., Gangeri, L., Bergman, B., Willems, J., Hanquet, P., and Zittoun, R. (1996). Multilingual translation of the Functional Assessment of Cancer Therapy (FACT) quality of life measurement system. *Quality of Life Research, 5*(3), 309-320.

Brisbane, F. L., and Womble, M. (1992). *Working with African Americans: The Professional's Handbook*, HRDI International Press, Chicago, IL.

Calman, K. C. (1984). Quality of life in cancer patients—an hypothesis. *Journal of Medical Ethics, 10*, 124-127.

Calman, K. C. (1987). Definitions and dimensions of quality of life. In Aaronson, N.K., and Beckman, J. (eds.), *The Quality of Life of Cancer Patients*, Raven Press, New York, pp. 1-10.

Cassileth, B. R., Lusk, E. J., Strouse, T. B., Miller, D. S., Brown, L. L., Cross, P. A., and Tenaglia, A. N. (1984). Psychosocial status in chronic disease: A comparative analysis of six diagnostic groups. *New England Journal of Medicine, 311*, 506-511.

Cella, D. F. (1994). *Manual for the Functional Assessment of Cancer Therapy (FACT) and Functional Assessment of HIV Infection (FAHI) Scales (Version 3),* Rush-Presbyterian-St. Luke's Medical Center, Chicago, IL.

Cella, D. F., and Bonomi, A. E. (1996). The Functional Assessment of Cancer Therapy (FACT) and Functional Assessment of HIV Infection (FAHI) Quality of Life Measurement System. In Spilker, B. (ed.), *Quality of Life and Pharmacoeconomics in Clinical Trials*, Raven Press, New York.

Cella, D. F., Hernandez, L., Bonomi, A. E., Corona, M., Vaquero, M., Shiomoto, G., and Baez, L. (1998). Spanish language translation and initial validation of the Functional Assessment of Cancer Therapy quality-of-life instrument. *Medical Care, 36*(9), 1407-1418.

Cella, D. F., Tulsky, D. S., Gray, G., Sarafiar, B., Lloyd, S., Linn, E., Bonomi, A., Silberman, M., Yellen, S. B., Winicour, P., Brannon, J., Eckberg, K., Purl, S., Blendowski, C., Goodman, M., Barnicle, M., Stewart, I., McHale, M., Bonomi,

P., Kaplan, E., Taylor, S., Thomas, C., and Harris, J. (1993). The Functional Assessment of Cancer Therapy Scale: Development and validation of the general measure. *Journal of Clinical Oncology, 11*(3), 570-579.

Cella, D. F., Lloyd, S. R., and Wright, B. D. (1996). Cross-cultural instrument equating: Current research and future directions. In Spilker, B. (ed.), *Quality of Life and Pharmacoeconomics in Clinical Trials*, Raven Press, New York.

Cella, D. F., Orofiamma, B., Holland J. C., Silberfarb, P. M., Tross, S., Feldstein, M., Orav, E. J., Perry, M., Maurer, L. H., Comis, R., and Green, M. (1987). Relationship of psychological distress, extent of disease, and performance status in patients with lung cancer. *Cancer, 60*, 1661-1667.

Ferrell, B. R., Dow, K. H., Leigh, S., Ly, J., and Gulesekaram, P. (1995). Quality of life in long-term cancer survivors. *Oncology Nursing Forum, 22*(6): 915-922.

Fitchett, G., Cella, D. F., and Peterman, A. (1996). Spiritual beliefs and quality of life in cancer and HIV patients. *Paper Presentation at the Society for Scientific Study of Religion: Annual Meeting*, Nashville, TN.

Freeman, H. E., Aiken, L. H., Blendon, R. J., and Corey, C. R. (1990). Uninsured working-age adults: Characteristics and consequences. *Health Services Research, 24*(6), 811-823.

Freeman, H. E., Blendon, R. J., Aiken, L. H., Sudman, S., Mullinix, C. F., and Corey, C. R. (1987). Americans report on their access to health care. *Health Affairs, 6*(1), 6-18.

Ganz, P. A. (1993). Age and gender as factors in cancer therapy. *Clinical Geriatric Medicine, 9*, 145-155.

Ganz, P. A., Schag, C. A., and Heinrich, R. L. (1985). The psychosocial impact of cancer on the elderly: A comparison with younger patients. *Journal of the American Geriatrics Society, 33*, 429-435.

Hahn, B., and Flood, A. B. (1995). No insurance, public insurance, and private insurance: Do these options contribute to differences in general health? *Journal of Health Care for the Poor and Underserved, 6*(1), 41-59.

Halstead, M. T., and Fernsler, J. I. (1994). Coping strategies of long-term cancer survivors. *Cancer Nursing, 17*, 94-100.

Hays, R. D., and Stewart, A. L. (1990). The structure of self-reported health in chronic disease patients. *Psychological Assessment: American Journal of Consultative Clinical Psychology, 2*(1), 22-30.

Hendershot, G. E. (1988). Health status and medical utilization. *Health Affairs, 7*(1), 114-121.

Hollingshead, A. B. (1975). *Four Factor Index of Social Status*. Yale University, New Haven, CT.

Jenkins, R. A., and Pargament, K. I. (1988). Cognitive appraisals in cancer patients. *Social Science and Medicine, 26*, 625-633.

Kaczorowski, J. M. (1989). Spiritual well-being and anxiety in adults diagnosed with cancer. *Hospice Journal, 5*, 105-116.

Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. (1988). *Applied Regression Analysis and Other Multivariable Methods.* Duxbury Press, Belmont, California.

Lazarus, R. S. (1980). The stress and coping paradigm. In Bond, L.A. , and Rosen, J.C. (eds.), *Competence and Coping during Adulthood*, University Press of New England, Hanover, N.H.

Llewellyn-Thomas, H. A., Sutherland, H. J., Hoggs, S. A., Ciampi, A., Harwood, A. R., Keane, T. J., Till, J. E., and Boyd, N. F. (1984). Linear analogue self assessment of voice quality in laryngeal cancer. *Journal of Chronic Disease, 37*, 917-924.

Marin, G., and Marin, B. V. O. (1991). *Research with Hispanic Patients*. Applied Social Research Methods Series (Vol 23), Sage, Newberry Park, CA.

Molina, C. W., Zambrana, R. E., and Aguirre-Molina, M. (1994). The influence of culture, class and environment on health care. In Molina, C.W., and Aguirre-Molina, M. (eds.), *Latino Health in the U.S.*, American Public Health Association, Washington, D.C.

Monheit, A. C., Hagan, M. M., Berk, M. L., and Farley, P. J. (1985). The employed uninsured and the role of public policy. *Inquiry, 22*(4), 348-364.

Myken, P., Larsson, P., Larsson, S., Berggren, H., and Caidahl, K. (1995). Similar quality of life after heart valve replacement with mechanical or bioprosthetic valves. *Journal of Heart Valve Diseases, 4*(4), 339-345.

Norusis, M. J. (1993). *SPSS for Windows Base System User's Guide Release 6.0.*, SPSS Inc., Chicago, IL.

Riley, G. F., Potosky, A. L., Lubitz, J. D., and Brown, M. L. (1994). Stage of cancer diagnosis for Medicare HMO and fee-for-service enrollees. *American Journal of Public Health, 84*(10), 1598-1604.

Rodin, J. (1986). Aging and health: Effects of the sense of control. *Science, 233*, 1271-1276.

Schag, C. A., and Heinrich, R. L. (1988). *Cancer Rehabilitation Evaluation System (CARES) Manual*, CARES Consultants, Los Angeles, CA.

Schipper, H., Clinch, J., McMurray, A., and Levitt, M. (1984). Measuring the quality of life of cancer patients: The Functional Living Index-Cancer: Development and validation. *Journal of Clinical Oncology, 2*, 472-483.

Stewart, A. L., and Van Dam, F. S. A. M. (1983). Quality of life questionnaire for use in studying chemotherapy and radiotherapy treatments in lung cancer

patients. *Proceedings from the 4th Workshop EORTC Study Group on Quality of Life.* Birmingham, England.

Stewart, A. L., Ware, J. E., and Brook, R. H. (1981). Advances in the measurement of functional status: Construction of aggregate indexes. *Medical Care, 19*, 473-488.

Wan, G. J., Counte, M. A., and Cella, D. F. (1997a). The influence of personal expectations on cancer patients' reports of health-related quality of life. *Psycho-Oncology, 6*, 1-11.

Wan, G. J., Counte, M. A., and Cella, D. F. (1997b). A framework for organizing health-related quality of life research. *Journal of Rehabilitation Outcomes Measurement, 1*(2), 31-37.

Warnecke, R. B., Ferrans, C. E., Johnson, T. P., Chapa-Resendez, G., O'Rourke, D. P., Chavez, N., Dudas, S., Smith, E. D., Martinez Schallmoser, L., Hand, R. P., and Lad, S. T. (1996). Measuring quality of life in culturally diverse populations. *Journal of the National Cancer Institute Monographs, 20*, 29-38.

Woodcook, R. W. (1991). *Examiner's Manual: Woodcock Language Proficiency Battery-Revised.* DLM Publishers, Allen, Texas.

Yancik, R., Edwards, B. K., and Yates, J. W. (1987). *Quality of Life Assessment: A Pilot Study Report*, National Cancer Institute, Division of Cancer Prevention and Control, Washington, DC.

Zubrod, C. G., Schneiderman, M., Frei, E., Brindley, C., Gold, G. L., Shnider, B., Orieto, C., Gorman, J., Jones, R., Jonsson, V., Colsky, J., Chalmers, T., Ferguson, B., Dederick, M., Holland, J., Selawry, O., Regelson, W., Lasagna, L., and Owens, A. H. (1960). Appraisal of methods for the study of chemotherapy of cancer in man: Comparative therapeutic trial of nitrogen mustard and triethylene thiophosphoramide. *Journal of Chronic Disease, 11*, 7-33.

JOURNAL OF OUTCOME MEASUREMENT®, *3*(3), 216–232

Creating Performance Categories from Continuous Motor Skill Data Using a Rasch Measurement Model

Beth Hands
University of Western Australia

Barry Sheridan
RUMM Laboratory

Dawne Larkin
University of Western Australia

This paper reports the use of the Extended Logistic Model (ELM) of Rasch (Andrich, 1988), based on Item Response Theory, to validate the reduction of continuous motor skill data to categories of performance. The data were gathered from the performances of 5 and 6 year old children on 24 fundamental movement skills and involved different measurement units such as seconds, centimetres, scores and counts. In order to compare results across all skills the data were collapsed into discrete sets of categories. Several alternative cut-off locations based on normative data were considered. A feature of the ELM is that it can account for correct scoring of the response categories, but only if the threshold estimates derived from the data by the measurement model are correctly ordered in a hierarchical fashion, from lowest to highest. Should this be the case, a valid scoring function has been established. In this study, the data were successfully reduced to three categories based on the 15th and 85th percentile allowing further analysis to proceed.

Requests for reprints should be sent to Beth Hands, Department of Human Movement and Exercise Science, University of Western Australia, NEDLANDS WA 6907, e-mail: bhands@cygnus.uwa.edu.au

Introduction

There are a number of occasions when it is useful to categorize continuous data. In most cases, a cut score is used to dichotomize data for a specific purpose, such as deciding a person's eligibility for a particular intervention program, or whether a student has mastered the learning outcomes of a teaching unit. Another reason to categorize data occurs when the items in a data set have been measured in different units and performance comparisons between items is necessary. This situation often arises when assessing motor skill performances. Units such as centimeters (for the distance jumped), seconds (the time to run 30 meters or balance on one leg), scores (a throw for accuracy), and counts (number of successful catches) may be involved. The raw scores derived from such skill performances, however, cannot be compared directly. Performance comparisons across items are important when seeking to measure a single underlying construct, for example motor ability, which is demonstrated through a number of different skills. One solution is to standardize the scores by conceiving a continuum of competence for each motor skill and dividing each continuum into a sequence of categories (for example, 1, 2 or 3), similar to the typical Likert format, for each item. These categories represent levels of performance on each skill; for example inefficient, proficient and advanced. This particular solution is appealing because it enables the data to be analyzed with a measurement model capable of validating the categorization procedure (Andrich, 1988).

Determining the cut score to divide data into two or more categories has been the topic of debate for many years (Berk, 1996; Dwyer, 1996; Glass, 1978; Shepard, 1980). A number of techniques have been used to establish cut scores in criterion testing, for example Berk (1996) identified nearly 50 methods in the literature. However, all approaches depend principally on human judgement. As a consequence, the creation of cut off locations between categories is often quite arbitrary (Dwyer, 1996). Glass (1978, p. 258) considered every attempt at standard setting to be either "blatantly arbitrary or derives from a set of arbitrary premises". Burns and Harrison (1978) concluded that no technology existed that could satisfactorily guide performance standard setting. Some classification error was unavoidable given the necessary human element. Since the appearance of those papers, others have argued that standard setting is acceptable provided sufficient information, expertise and logic are applied by the judges (Dwyer, 1996; Shepard, 1980), although a sound, valid strategy has yet to be determined.

Various methods have been proposed and many reviews have been written regarding the most appropriate cut score procedure (Glass, 1978; Hambleton and Eignor, 1979; Shepard, 1980). In most cases, the discussion has centered on the issue of mastery versus non-mastery that involves simply dichotomizing the data. Recent assessment requirements in education in the U.S.A. now require multiple cut-scores, consequently some new approaches are being investigated (Berk, 1996). Some strategies use minimum competency levels derived from other criterion scores, decision theory, operations research methods, needs assessment, or contrasting groups (Glass, 1978; Scriven, 1978; Shepard, 1980). Judges setting the cut scores may be asked to consider the characteristics of the test sample, the type of test questions or some aspects of performance (Dwyer, 1996). A more recent review by Berk (1995) has highlighted methods such as judgemental policy capturing, extended Angoff procedure, and the dominant profile method. A special feature of these approaches is the extensive training of judges involved. When multiple cut-scores are required, procedures such as behavioral anchoring or contrasting-groups have been used with written papers (Berk, 1996).

Some cut score methods may be guided by normative information such as means, standard deviations or percentiles. Both Linn (1978) and Shepard (1980) argue that failing to consider such information may result in unrealistic or inappropriate standards whereas using normative data ensures the standards are reasonable and realistic across the population.

With most methods, however, any misclassification of responses into categories is generally not identified, or worse, ignored, although in many cases the success or failure of the cut-off points may be statistically evaluated after the event (Hambleton, 1978). Generally, such post hoc evaluation offers little guidance as to more suitable locations. Further, different cut score methods often result in different outcomes. For example, when a group of well-trained judges created cut scores from the same set of data using three different procedures, judgemental policy capturing, extended Angoff procedure, and dominant profile method, each procedure resulted in a different cut score (Plake, 1995). The literature has, to date, ignored these issues and most current approaches lack evidence of validity and psychometric sophistication.

The Extended Logistic Model of Rasch

One method of validating a series of cut score locations with continuous data is available from Item Response Theory (IRT) using the Ex-

tended Logistic Model (ELM) of Rasch (Andrich, 1988). Some other IRT models have been used to detect aberrant cut scores, unusual person responses (Meijer, Muijtjens and van derVleuten, 1996) or scale test scores (Berk, 1996). The ELM model, however is particularly suited to validating multiple cut scores as it is designed for assessing the psychometric properties of items by capturing information relevant to category order. By estimating threshold values linked to the cut off locations, this model provides empirical evidence as an integral part of the measurement process and so takes the guesswork out of the procedure.

Threshold Estimates

Item thresholds are conceptualized as a set of boundaries located between the response categories of an item and specify the change in probability of a response occurring in one or the other of the two categories separated by the threshold, for example, between 'low' and 'very low'. As the difficulty of an item increases, the probability of a response in a higher category decreases for a person of a given ability. For a person possessing a low ability, the most probable response to an item of high difficulty would be in the first category. A person possessing greater ability would be expected to score in a higher order category for the same item, so that one or more thresholds will be exceeded. The logic of this situation means that a person of greater ability would receive scores in higher placed categories for items of increasingly lower difficulty values, than would be the case for a person of lesser ability. If this consistency is observed across all person response patterns for all items, then the Guttman requirement (Guttman, 1954) for objective measurement will prevail and the item thresholds will be in an ordered sequence.

A special feature of the ELM, then, is the facility to account for correct scoring of the response categories, but only if the threshold estimates derived from the data by the measurement model are correctly ordered in a hierarchical fashion, from lowest to highest (Andrich, 1985). The threshold estimates appear in an ordered sequence from low to high when the item is behaving according to the intended, original scoring pattern as displayed in Figure 1. These category characteristic curves display visually the meaning of the threshold locations relative to adjacent categories. If thresholds are disordered, however, then the probability of a score in at least one category will never be greater than the probability of a score in at least one other category. The scoring of the categories is, therefore, not acting as expected, and the data do not fit the construction of the model.

This situation is illustrated in Figure 2 where the probabilities of success in both Category 1 and Category 2 are never the most likely at any point along the Person Ability continuum. Andrich (1982) noted that only if the thresholds were ordered was the model distribution strictly unimodal.

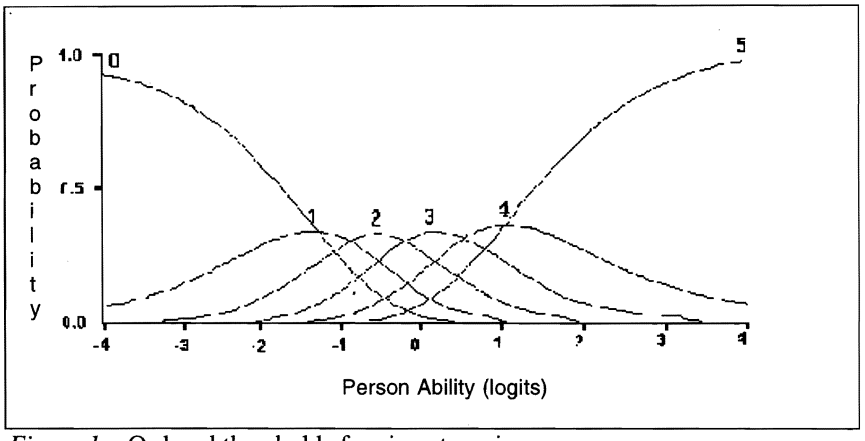

Figure 1. Ordered thresholds for six categories.

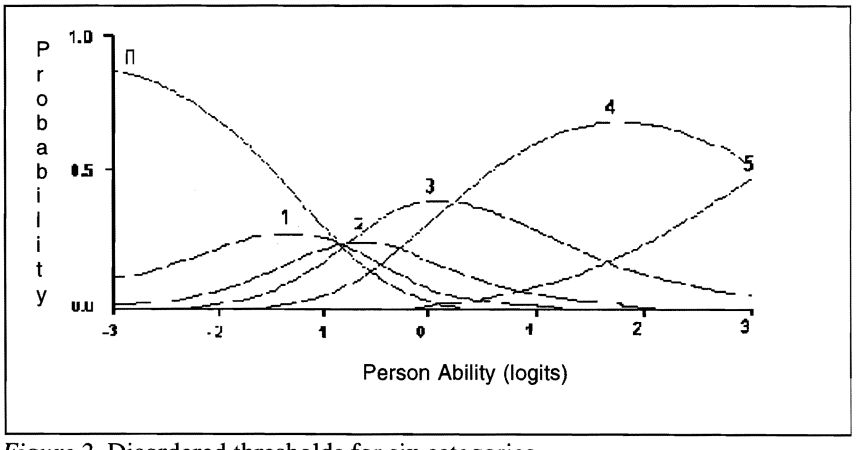

Figure 2. Disordered thresholds for six categories.

Once a set of categories have been determined the threshold estimates, which arise as a consequence of this action, provide evidence of the validity of the cut scores to create the scoring function mechanism. When threshold estimates are disordered, one reason could be the incorrect location of one or more category cut-off point. Conversely, if the thresholds are ordered then the mechanism used to create the scoring function has a logical base within the overall context of the measurement structure.

The purpose of this paper is to report the application of the Extended Logistic Model to validate the reduction of continuous motor skill data to categories of performance. This procedure was important because it enabled the data collected from the performances of five and six year old children on 24 fundamental movement skills to be analyzed using IRT. In so doing, a measure of motor ability was constructed. Four different measurement units were involved - seconds, centimeters, counts and a score. By exploring a number of alternative cut-off points based on normative data, and consulting the threshold estimates that are derived by the model, a valid scoring function can be established. This procedure is in contrast to the more arbitrary approaches used to date, for establishing cut scores when reducing continuous motor skill data to categories. Under these circumstances, the validity of the cut score locations have not been statistically evaluated.

Cut Score Methods Used in Motor Skill Assessments

The measurement of motor skills often involves sets of raw scores expressed in a number of different units. In most cases, normative data have provided the source of information for establishing cut scores so that performance data could be compared and easily interpreted. Means and standard deviations guided Calder (1979) and Larkin and Revie (1994) when they created categories of performance from continuous data in order to interpret motor screening tests. Test performances across a number of different skills were categorized into five levels - very low, low, medium, high, and very high. Ulrich (1984) used percentiles to represent two different mastery points. The 70^{th} and 85^{th} percentiles of the total score for the Objectives-based Motor Skill Assessment instrument, a criterion referenced test, both successfully and consistently classified students as masters or non-masters. Other tests of motor performance have also used percentiles to determine degree of motor impairment. Henderson and Sugden (1992) used the 5^{th} and 15^{th} percentiles as significant cut off points between performance scores on the Movement ABC and Gubbay (1975) employed the 5^{th}, 10^{th}, 50^{th}, 90^{th}, and 95^{th} percentiles.

Information based on IRT analysis has been used to convert item raw scores to item point scores (or category cut scores) in the Bruininks-Oseretsky Test of Motor Proficiency (Bruininks, 1978). Bruininks initially assigned raw scores to point scores to ensure a reasonable range of scores that discriminated between performances. This procedure was refined by using an IRT analysis procedure developed by Woodcock and Dahl (1971) based on W-scale values (as cited in Bruininks, 1978) and

was followed by a visual inspection of the difficulty estimates for each cut score to check for irregularities in spacing between locations. While adjustments from the raw score to category conversions were made to establish greater uniformity, this procedure, although innovative at the time, still relied heavily on human judgement.

Method

Sample

The sample comprised of 332 pre-primary and primary school children from four randomly selected schools in Perth, Western Australia. The boys ($n = 168$) and girls ($n = 164$) ranged in age from 54 to 82 months ($M = 66.5$ months, $SD = 7.3$ mths).

Data Collection

The children performed 24 motor skills during two testing sessions, no more than one week apart. The quantitatively scored items included a range of locomotor, object control and balance tasks typical of children aged 5 and 6 years (Hands, Larkin, and Sheridan, 1997; Wickstrom, 1987; Williams, 1983). The children were tested in their own school environment and the testing procedure was identical for all children.

Data Reduction

To facilitate a mechanism for creating categories, two strategies based on normative information were considered. The data were divided into discrete categories based on the mean and standard deviation for each item in the first instance, and on percentiles in the second instance. With each procedure, the categorized data were analyzed using the ELM and the threshold estimates for each item examined. The sequence order of the threshold estimates then provided a basis for assessing the validity of the cut score procedure involved. By anchoring these data to the measurement process itself, the threshold strategy provides a more informed outcome than that based largely on judgement, as in the past.

Results

Cut Scores Derived From Item Mean and Standard Deviation

With the mean and standard deviation method for deriving item categories, six performance categories were created initially by using three standard deviations either side of the mean. Cut off points were located at

the mean, one standard deviation each side of the mean, and at two standard deviations each side of the mean. With the 30 meter run, for example, a mean of 8.11 secs and standard deviation of .89 produced cut-off values of 6.1 secs, 7.1 secs, 8.1 secs, 9.1 secs, and 10.1 secs.

The success of any cut score procedure is contingent on the meaningful separation of information across all categories. Allied to this is the notion that the greater the number of categories created for each item, the greater the potential for increasing the precision of the measurement outcomes which in turn leads to a finer discrimination between individual performances. It was apparent in this present study that a number of items had skewed distributions and/or large standard deviations, while other items with a high percentage of zero scores created a scoring problem. For these latter items, such as hop, skip, or gallop, children unable to

Table 1

Threshold Estimates for 24 Items Using Means and Standard Deviations

	Thresholds (logits)				
Item	1	2	3	4	5
Balance Left Foot*	-1.158	0.829	3.08	-2.751	
Balance Right Foot*	-1.287	1.278	2.621	-2.613	
Distance Hop Left Foot*	-.0969	-1.383	0.253	2.099	
Distance Hop Right Foot	-1.465	-0.907	0.427	1.945	
Standing Broad Jump	-3.11	-1.779	-0.114	1.665	3.338
Leap	-2.433	-1.634	-0.075	1.559	2.582
Hop Left Foot*	-0.9	-0.991	-1.511	-0.52	2.123
Hop Right Foot*	0.661	-0.765	-1.219	-0.454	1.777
Bounce Large Ball	-0.548	-1.002	-0.483	0.519	1.514
Bounce Small Ball-Left Hand*	-0.279	-0.54	-0.366	0.174	1.01
Bounce Small Ball-Right Hand	-1.2	-0.499	-0.055	0.445	1.309
Shuttle Run	-1.545	-0.911	-0.129	2.585	
30 Metre Run	-3.295	-2.095	-0.665	1.43	4.626
Target Throw	-0.963	-0.658	1.62		
Vertical Jump	-2.614	-1.486	-0.239	1.247	3.092
Catch Large Ball*	1.709	-1.572	-0.137		
Catch Small Ball	-0.619	0.619			
2 Arm Strike	-1.019	-0.481	1.501		
Skip*	1.301	-0.906	-1.511	-0.605	1.721
Slide*	0.658	-1.045	-1.377	-0.332	2.096
Gallop*	0.08	-0.761	-1.008	-0.248	1.937
Distance Throw	-3.695	-1.132	0.613	1.705	2.469
Distance Kick*	-0.069	0.286	-0.217		
Tee Ball Hit*	-1.661	1.06	0.601		

*Disordered thresholds

perform the task were given a score of zero. This important information, however, was not congruent with the scoring method for these items, where the lower score indicated the better performance. The variable nature of the scoring function for these items necessitated the adoption of an alternative categorization strategy. This modification was achieved, in some cases by halving or doubling the standard deviations for each item, and in others by simply reducing the number of categories. As a result, the data for 13 of the 24 items were organized into six categories, as a normal distribution prevailed. Of the remainder, 5 items were divided into 5 categories, another 3 items were divided into 4 categories and 1 item into 3 categories. Once all the items were allocated their respective number of categories, the data were scored using the familiar Likert method and the resultant scores submitted for analysis using the RUMM computer pro-

Table 2

Threshold Estimates for 24 Items Using 10th, 25th, 50th, 75th, and 90th Percentiles

	Thresholds (logits)				
Item	1	2	3	4	5
Balance Left Foot	-.554	-.417	.407	.564	
Balance Right Foot	-.743	-.251	.442	.551	
Distance Hop Left Foot	-1.113	-.799	.004	.803	1.105
Distance Hop Right Foot	-1.070	-.826	.044	.870	.982
Standing Broad Jump	-.956	-.816	.034	.851	.887
Leap	-.968	-.780	-.070	.710	1.108
Hop Left Foot*	-.107	-.683	-.732	-.049	1.572
Hop Right Foot*	-.046	-1.024	-.798	.226	1.643
Bounce Large Ball	-1.378	-.559	-.143	.416	1.664
Bounce Small Ball-Left Hand*	-.209	-1.182	-.236	.945	.682
Bounce Small Ball-Right Hand	-.983	-.765	-.019	.746	1.020
Shuttle Run	-.982	-.794	-.003	.791	.987
30 Metre Run	-1.063	-.817	.044	.861	.975
Target Throw*	.744	-1.189	.445		
Vertical Jump *	-.806	-.821	.057	.878	.692
Catch Large Ball*	.261	-.765	.504		
Catch Small Ball	-.570	.570			
2 Arm Strike*	-.350	-.602	.952		
Skip*	.566	-.537	-.504	.474	
Slide*	-.429	-.617	-.437	.179	1.304
Gallop	-1.881	-.368	-.022	.346	1.925
Distance Throw*	-.740	-.810	-.084	.726	.909
Distance Kick*	-.076	-.717	.277	.516	
Tee Ball Hit*	-.519	-.762	-.108	.655	.735

*Disordered thresholds

gram (Andrich, Sheridan, and Luo, 1997). As Table 1 reveals, 12 items exhibit disordered threshold values which indicate that the scoring mechanism derived for the response categories was not working as expected.

Cut Scores Derived from Percentiles

With the second categorization method, two sets of percentiles were considered for the cut off values by employing a procedure similar to that described in the previous section.

10th, 25th, 50th, 75th, and 90th percentiles. Using the same example as above for the 30 meter run, six categories were now defined by cut scores located at 7.0 secs, 7.4 secs, 8.0 secs, 8.7 secs, 9.2 secs, and 9.3 secs. As for the previous method, it was not possible to create six categories for all items. The data for 16 items were collapsed into 6 categories, 4 items

Table 3

Centralized Threshold Estimates for 24 Items using 15th and 85th Percentiles for the Total Sample

	Thresholds (logits)	
Item	1	2
Balance Left Foot	-1.649	1.649
Balance Right Foot	-1.469	1.469
Distance Hop Left Foot	-2.086	2.086
Distance Hop Right Foot	-1.945	1.945
Standing Broad Jump	-1.992	1.992
Leap	-1.824	1.824
Hop Left Foot	-2.482	2.482
Hop Right Foot	-2.514	2.514
Bounce Large Ball	-1.744	1.744
Bounce Small Ball-Left Hand	-1.698	1.698
Bounce Small Ball-Right Hand	-1.868	1.868
Shuttle Run	-1.923	-1.923
30 Metre Run	-1.779	1.779
Target Throw	-.694	.694
Vertical Jump	-1.880	1.880
Catch Large Ball	-.994	.994
Catch Small Ball	.000	
2 Arm Strike	-1.476	1.476
Skip	-1.692	1.692
Slide	-2.203	2.203
Gallop	-1.856	1.856
Distance Throw	-2.076	2.076
Distance Kick	-1.250	1.250
Tee Ball Hit	-1.929	1.929

were divided into 5 categories, 3 items were divided into 4 categories and 1 item into 3 categories. Once again, threshold disorder resulted for 12 items (Table 2). This situation could have arisen, in part, from a lack of precision in distinguishing between some percentile scores around the median and may have led to an incorrect classification of some performances. On the other hand, percentiles are often useful in distinguishing between performances at the more extreme areas of the ability range. As a consequence of these outcomes, it was decided to create fewer categories using the more extreme 15th and 85th percentiles.

15th and 85th Percentiles. The 15th percentile was chosen as it has been found to identify the approximate percentage of children with movement difficulties in the population (Larkin and Hoare, 1991) and was also used as the borderline or cut off value for identifying such children using the Movement ABC (Henderson and Sugden, 1992). In contrast to this, the 85th percentile has been used to reflect skill mastery (Ulrich, 1984). Further,

Table 4

Category Scores Based on 15th and 85th Percentiles

		Category		
Item	Unit	0	1	2
Balance Left Foot	secs	≤ 4	4.1 - 29.9	30
Balance Right Foot	secs	≤5	5.1	30
Distance Hop Left Foot	secs	≤37	38-79	≥80
Distance Hop Right Foot	secs	≤40	41-82	≥83
Standing Broad Jump	cm	≤79	80-11	≥112
Leap	cm	≤68	69-91	≥92
Hop Left Foot	secs	0	≥1.9	≤1.8
Hop Right Foot	secs	0	≥1.9	≤1.8
Bounce Large Ball	count	≤5	6 - 18	≥19
Bounce Small Ball-Left Hand	count	0	1 - 9	≥10
Bounce Small Ball-Right Hand	count	≤1	2 - 13	≥14
Shuttle Run	secs	≥12.0	9.5 -11.9	≤9.4
30 Metre Run	secs	≥9	7.3 - 8.9	≤7.2
Target Throw	score	0	1 - 4	≥5
Vertical Jump	cm	≤14	14.5 - 22	≥22.5
Catch Large Ball	count	0	1 - 9	10
Catch Small Ball	count	0	1 - 10	
2 Arm Strike	count	0	1 - 8	≥9
Skip	secs	0	4.3	≤4.2
Slide	secs	0 + ≥7.5	4.7 - 7.4	≤4.5
Gallop	secs	0 + ≥6.0	4.1 - 5.9	≤4.0
Distance Throw	cm	≤394	395 - 987	≥988
Distance Kick	cm	0	1 - 413	≥414
Tee Ball Hit	cm	≤80	81 - 669	≥670

maximum discrimination occurs at the more extreme levels of sets of scores. One item, Catch Small Ball, was scored dichotomously as this classification was found to reflect more accurately the children's performance where they either caught most of the 10 throws or caught none at all.

When the scoring function was based on the 15th and 85th percentiles (Table 4) all thresholds were now found to be both ordered and well spaced (Table 3). Having established a sound scoring function, the meaning of the variable of interest, in this case, motor ability, could now be explored. This can be achieved by noting the order of the items along the measurement continuum and examining the nature of the easier and harder items in the context of the meaning of the original conceptual framework guiding the study. Once such meaning has been established, then each person's score will become meaningful based on the location of their ability relative to the items on the continuum. The results of this analysis are reported elsewhere (Hands, 1998).

Discussion

Techniques for dividing continuous data sets into finite numbers of categories have, in general, relied essentially on human judgement with little empirical evidence provided to substantiate the scoring mechanism adopted. In the present paper, two methods of data reduction were presented and the resultant scoring functions assessed empirically using the criterion of threshold order between the categories. In this technique, which is fundamental to the measurement process, threshold values that appear in an ordered sequence across the categories provide strong evidence that the scoring function is working as expected. Andrich (1982, p. 121) suggested that in the social and biological sciences, the rating model be used "...when formal measurements cannot be made. Application of the ELM to such data can, therefore, provide a check on the quality of rating mechanism and help place the results of ratings on the same level as that of usual measurement".

The methods employed in the present study provided valuable outcomes as regards threshold order for the number of categories created. With categories created based on either the mean and standard deviations of each item or five different percentiles, a significant number of items exhibited threshold disorder. These outcomes indicated problems with the scoring mechanism. The reduction of the data into three performance categories based on the 15th and 85th percentiles was successful. One disadvantage of this outcome is the loss of sensitivity in detecting perfor-

mance differences. This was especially true for the second or middle category that represented a wide range of proficiency. However, given the skewed nature of the data for some items and the various item scoring strategies involved (for example, "0" was recorded for failure to perform a task) some loss of information was unavoidable. Ideally, a greater number of categories would have allowed greater precision of measurement. However, the data for only 13 items were distributed across a normal curve and able to be categorized into six categories based on the mean and standard deviation of the raw data. Of those 13 items, the threshold estimates for 6 items were disordered (Table 1). Similarly, using the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles, it was possible to create six categories for 16 of the 24 items. However, as shown in Table 2, the threshold estimates for 7 of those items were disordered. Some of these items may have been too easy or too hard for the sample and therefore attracted skewed results, however when this procedure is undertaken to explore a construct, there is a trade-off situation present at all times. Although more categories would result potentially in a higher degree of precision, this would occur only if the difference between each category was meaningful and the allocation of responses to a particular category followed a logical pattern based on the interaction between a person's ability and item difficulties across all items. This pattern would be unattainable if ambiguity existed between adjacent categories such that responses could occur on a more random basis. The allocation of responses to one category in preference to another would be difficult. Under these conditions, the number of categories would have to be reduced until a consistent pattern evolves. On the other hand, three categories representing differing levels or performance such as basic, average and advanced are frequently used in a number of fields.

Other applications

In many cases, data simply needs to be divided into two categories representing mastery or non-mastery, and even though the data may be analyzed using the ELM, the cut score technique presented in this paper is not available. This is because no threshold estimates are derived from analysis of dichotomous data. In this situation, the cut score for each item may be still independently determined. This is the simplest case of category allocation. For example, the category of mastery may be allocated should a child catch 6 out of the 10 throws. In this case a count of 6 becomes the cut score. Validity of the cut score is derived from an exami-

nation of the item difficulties and the overall fit of the items to the measurement model after the data has been analyzed using the ELM. Placing the cut scores at different locations, for example 5 out of 10 catches, or 8 out of 10 catches, effectively alters the difficulty of an item and invariably its location along the measurement continuum relative to the other items. Provided the cut score locations have a sound theoretical basis, an examination of the outcome of the data analysis will provide empirical evidence for the validity of the cut scores.

This same procedure could also be used to determine the cut score between mastery and non-mastery for person ability based on an overall test performance. An examination of the location and content of all test items along the measurement continuum of item difficulty would enable a mechanism for determining overall mastery of the variable of interest. For example, in this present study, an ability score could be identified which classifies a child as a proficient or non-proficient mover. Clearly any demarcation specifying mastery from non-mastery, although based on a sound theoretical position, will always be relative.

In this study, only normative information was used to create the categories. However the validity of cut scores based on other methods, some of which were mentioned earlier, could also be evaluated using this approach.

Limitations

This innovative procedure is one method of validating any given cut score allocated to an outcomes continuum, however further investigation is warranted as the importance of some issues to the result presented in this study need to be considered.

Threshold disorder, for example, can be influenced by the degree to which person abilities and item difficulties overlap. When a mistarget arises in this way between the set of persons and item estimates, insufficient variability, often in the extreme categories, in the data can mask the true nature of the underlying structure or latent variable being measured. That is, threshold disorder may not, in fact, be due to incorrect category scoring but rather to a lack of representative data in some categories.

While this approach offers a more sophisticated and more objective method of setting cut scores, human judgement is still involved in deciding what normative information (for example mean, percentiles) will be used in the first instance. However, the advantage of employing the ELM is that information derived from threshold estimates provides a mecha-

nism for evaluating cut scores in a meaningful way so that adjustments, based on human judgement in the context of reasoned theory, can be made prior to further analysis or decision making.

Conclusion

The use of threshold estimates to validate the location of category cut scores is an innovative contribution to measurement practice. The ELM involved in this process is a valuable tool to guide the building of meaningful variables or constructs whose measure is required in varying units. Motor performance data collected from children were variously scored in centimeters, seconds, scores and counts. To compare performances across items, therefore, it was necessary to collapse the range of raw data to a finite number of categories. Whereas previously, the establishment of cut-off points for creating such categories has been based mainly on human judgement, the ELM can account for a correct scoring function if the threshold estimates are ordered. Where threshold estimates for a given item were disordered, it was assumed that one or more category cut-off points were in the incorrect location. Threshold estimates were employed to evaluate a number of strategies based on normative data, such as means and standard deviations, and percentiles to determine cut-off locations. In this example, percentiles were found to be the most appropriate source of information although this may differ for other situations.

References

Andrich, D. (1982). Using latent trait measurement to analyze attitudinal data: a synthesis of viewpoints. In D. Spearitt (Ed.), *The improvement of measurement in education and psychology* (pp. 89-126). Melbourne, VIC: ACER.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In B. Brandon-Tuma (Ed.), *Sociological Methodology* (pp. 33-80). San Francisco, CA: Jossey-Bass.

Andrich, D. (1988). A general form of Rasch's extended logistic for partial credit scoring. *Applied Measurement in Education, 1*(4), 363-378.

Andrich, D., Sheridan, B., and Luo, G. (1997). RUMM: A Windows program for analyzing item response data according to Rasch Unidimensional Measurement Models (Version 2.8) [Computer Software]. Perth, Western Australia: RUMM Laboratory.

Berk, R. A. (1995). Something old, something new, something borrowed, a lot to do! *Applied Measurement in Education, 8*(1), 99-109.

Berk, R. A. (1996). Standard setting: The next generation (Where few psychometricians have gone before!). *Applied Measurement in Education*, *9*(3), 215-235.

Bruininks, R. H. (1978). *Bruininks-Oseretsky Test of Motor Proficiency*. Circle Pines, MN: American Guidance Service.

Burns, Y., and Harrison, J. A. (1978). The role of physiotherapy in the assessment and treatment of neurological, sensory, motor problems in children. *Australian Journal of Remedial Education*, *10*(1), 8-11.

Calder, J. (1979). *The Queensland motor performance screening test for young children*. Brisbane, QLD: University of Queensland.

Dwyer, C. (1996). Cut scores and testing: statistics, judgement, truth and error. *Psychological Assessment*, *8*(4), 360-362.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, *15*(4), 237-261.

Gubbay, S. S. (1975). *The clumsy child: a study of developmental apraxic and agnosic ataxia*. London: W. B. Saunders.

Guttman, L. (1954). The principal components of scalable attitudes. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 24-68). New York, PA: Free Press.

Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*, *15*(4), 277-289.

Hambleton, R. K., and Eignor, D. R. (1979). Competency test development, validation, and standard setting. In R. Jaeger and C. Tittle (Eds.), Minimum competency testing. Berkeley, CA: McCutchan.

Hands, B. (1998). *Employing the Rasch Model to Measure Motor Ability in Young Children*. Unpublished doctoral thesis, University of Western Australia, Australia.

Hands, B., Larkin, D., and Sheridan, B. (1997). Rasch measurement applied to young children. *The Australian Educational and Developmental Psychologist*, *14*(1), 11-22.

Henderson, S. E., and Sugden, D. A. (1992). *Movement Assessment Battery for Children*. Kent, UK: The Psychological Corporation.

Larkin, D., and Hoare, D. (1991). *Out of step: Coordinating kids' movement*. Nedlands, WA: Active Life Foundation.

Larkin, D., and Revie, G. (1994). *Stay in Step: a gross motor screening test for children K-2*. Sydney, NSW: Authors.

Linn, R. L. (1978). Demands, cautions, and suggestions for setting standards. *Journal of Educational Measurement*, *15*(4), 301-308.

Meijer, R. R., Muijtjens, A. M. M., and Van der Vleuten, C. P. M. (1996). Nonparametric person-fit research: Some theoretical issues and an empirical example. *Applied Measurement in Education, 9*(1), 77-89.

Plake, B. S. (1995). An integration and reprise: what we think we have learned. *Applied Measurement in Education, 8*(1), 85-92.

Scriven, M. (1978). How to anchor standards. *Journal of Educational Measurement, 15*(4), 273-275.

Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4*(4), 447-467.

Ulrich, D. A. (1984). The reliability of classification decisions made with the objectives-based motor skill assessment instrument. *Adapted Physical Activity Quarterly, 1*, 52-60.

Wickstrom, R. L. (1987). Observations on motor pattern development in skipping. In J. E. Clark and J. H. Humphrey (Eds.), *Advances in motor development research*, 1 (pp. 49-60). New York: AMS Press.

Williams, H. G. (1983). *Perceptual and motor development.* Englewood Cliffs, NJ: Prentice-Hall.

JOURNAL OF OUTCOME MEASUREMENT®, *3*(3), 233–247

Detecting Differential Item Functioning with Five Standardized Item-Fit Indices in the Rasch Model

Hyunsoo Seol
Chung-Ang University

This study examined five Rasch-model-based item-fit indices: unweighted and weighted standardized indices (denoted UWz and Wz), standardized likelihood index (denoted Lz), and Extended Caution Indices (denoted ECI2z and ECI4z), in terms of their distributional properties and the power of detecting item bias or Differential Item Functioning (DIF). The results indicated that although these five standardized item-fit indices did not depart significantly from a normal distribution, it appeared that the Type I error rates were not reasonable. For the power of five standardized item-fit indices to detect DIF, the results showed that all indices did perform poorly across various conditions. These findings lead to the conclusion that all indices used in this study are inadequate fit measures for detecting DIF.

Requests for reprints should be sent to Hyunsoo Seol, National Examination Planning Division, Korea Institute of Curriculum and Evaluation, 15-1 Chungdam 2-Dong, Kangnam-Ku, Seoul, Korea (135-102), e-mail: seol@kice.re.kr

Over the years, a number of studies have been conducted to investigate person's or item's aberrant response patterns (e.g., Thurstone and Chave, 1929; Mosier, 1940; Glaser, 1949, 1952; Wright, 1977; Wright and Stone, 1979; Drasgow, 1982; Levine and Drasgow, 1982; Drasgow and Levine, 1986; Tatsuoka and Tatsuoka, 1982a; Harnisch and Tatsuoka, 1983; Smith, 1982, 1985, 1991b). Recently, Rob R. Meijer and others (1996), in a special issue of *Applied Measurement in Education*, explore a number of issues concerning person-fit analysis. Identifying and diagnosing aberrant response patterns can provide much information such as test anxiety, cultural bias (van der Flier, 1977, 1982), group differences with regard to sex or race (Donlon and Rindler, 1979; Harnisch and Linn, 1981b), school differences identifying schools with curricula that do not match the test content (Harnisch and Linn, 1981a), and sporadic study habits (Blix and Dinero, 1985).

There are the indices which are based on Item Response Theory (IRT). These indices consist of "fit statistics" (denoted UWz and Wz) as suggested by Wright and Panchapakesan (1969), "likelihood" (denoted Lz) index as introduced by Levine and Rubin (1979), and "Extended Caution Indices" (denoted ECI2z and ECI4z) as developed by Tatsuoka and Linn (1983). These indices measure the goodness of fit between individual/item response pattern and IRT model in which response pattern with a low probability given IRT models will be classified as aberrant. This study mainly focuses on the 'standardized' fit indices to overcome the problem of the dependence of ability or difficulty distribution (e.g., Smith, 1986, 1988, 1991a; Smith, Schumaker, and Busch, 1995; Drasgow, Levine, and Williams, 1985; Drasgow, Levine, and McLaughlin,1987; Tatsuoka and Tatsuoka, 1982b; Tatsuoka, 1984).

As indicated by Reise (1990, p. 129) and Rost and von Davier (1994, p. 171), person-fit indices can be applied to item-fit indices because the relation between person- and item-fit indices is a symmetry. The formulae of five item-fit indices are stated in the appendix. Much of the research to date,however, has been focused on person-fit analysis. Little research has been done to examine distributional properties and relative power of item-fit indices. Even though Smith (1991a, 1994a, 1994b, 1995) conducted several studies concerning the distributional properties and the power of item-fit indices, his studies were based on the Rasch-based total and between item-fit indices.

Actually, three (Lz, ECI2z, ECI4z) of five indices investigated in this study were not developed for the Rasch model. Drasgow and Levine (1986),

however, noted that although standardized likelihood index has been studied widely for the three-parameter model, this index can be used in "any of the logistic, normal ogive, or other parameter models" (p. 60). Further, with respect to the extended caution indices, Tatsuoka and Linn (1983) showed certain correspondence between the student-problem curve (S-P curve) and the test response curve (TRC)/group response curve (GRC). In the study, they found that the person response curve (PRC), for the Rasch model, had shown monotonically decreasing functions along the difficulty level. However, for the two-parameter and three-parameter model, the PRC did not appeared to be represented by smooth, monotonically decreasing curve (see p. 88). This is a major advantage of the Rasch model because "the ECIs may be viewed as linear transformation of the covariance or correlation between a person's response pattern and a theoretical curve" (p. 95). Although Tatsuoka and Linn (1983) described 5 ECIs, the present study considers two standardized ECIs (ECI2z and ECI4z) because Tatsuoka (1984) suggested that " . . . ECI4 and ECI6 have identical standardized forms . . . the relationship between ECI1 vs ECI2 . . . correlate very highly . . . we drop ECI1 . . . and recognize ECI2 and ECI4 . . . as representive indices among the family of ECI indices" (p. 104).

It is expected that different IRT models influence both distribution and power of the indices because different IRT models might produce different item and person parameter estimates, which are used to calculate fit indices. Accordingly, the findings appeared to be confused and often produced quite contradictory results due to the different model or conditions (see Rudner, 1983, p. 218; Drasgow, 1982; Harnisch and Tatsuoka, 1983, p. 115). Further, under the same conditions, no one did attempt simultaneously to examine distributional properties and the power of five indices (UWz, Wz, Lz, ECI2z, and ECI4z) for item bias or differential item functioning (DIF) analysis until now.

To test the power of five standardized item-fit indices, various measurement disturbances (e.g., random guessing, start-up, item bias . . .) can be simulated and tested (see Smith, 1991a, p. 560). This study,however, focuses on the power of item-fit indices for DIF because identifying biased items are the major solution of the problem of how to measure person fairly of their sex, race, or cultural background. No attempts are made to test several different types of measurement disturbance. Masters (1988, p. 17) stated that "If an item's estimated difficulty is significantly greater when calibrated on one subgroup than when calibrated on the other, then

that item is considered 'biased' with respect to those two groups (see Holland, 1986; Lord, 1980; Wright and Stone, 1979)". Scheuneman (1975) also noted that "an item is considered unbiased if for persons with the same ability in the area being measured, the probability of a correct response on the item is the same regardless of the population group membership of the individual". Smith (1992, p. 86) operationally defined bias as "statistically different item difficulty estimates for the same item in subpopulations of interest". Based on the definitions of item bias or DIF described above, it is possible to generate responses with DIF. Therefore, the purpose of this paper is to investigate the distributional properties of five standardized item-fit indices as a descriptive statistic and to provide relative power of detecting DIF with five standardized item-fit indices under the Rasch model.

Procedures

First, to conduct simulation study person/item generation program called SIMTEST (Luppescu, 1993) was used to create random response patterns. And then, to compute the probabilities of observed scores, the ability and difficulty parameters were estimated by BIGSTEPS (Linacre and Wright, 1993) for an each examinee and an item. Second, the IRT parameters were used to calculate unweighted and weighted standardized item-fit indices (denoted UWz and Wz) using BIGSTEPS computer program, standardized likelihood item-fit index (denoted Lz), standardized extended caution indices (denoted ECI2z and ECI4z) using IFIT computer program (Seol, 1997). The specific research questions, however, require the specific sequence of computations and analyses. These are presented along with specific research questions. Further, this study is based on a single series of simulated data. It may be necessary to conduct more rigorous replication studies to generalize the results. However, this study did not attempt to develop a replication design because Smith, Schumacker, and Bush (1994c) showed that the number of replications was not a major factor controlling outcomes of Rasch simulation studies.

Results

The distributions of five standardized item-fit indices

In order to test hypothesis that the distributions of five standardized item-fit indices have unit normal distribution with a mean of zero and a standard deviation of one when data fits the model (e.g., Smith, 1994a;

Drasgow et al., 1985; Tatsuoka, 1984), the responses of 100 items and 1000 persons were generated to fit the model using SIMTEST computer program. Item difficulties have a uniform distribution ranging from - 3 to + 3 logits. The abilities used to generate responses were normally distributed with a mean of zero and a standard deviations of one. Five standardized item-fit indices were then calculated for each item. The means, standard deviations, and Type I error rates were computed for each index. With regard to the Type I error rate, a value of greater than |2| was used as the critical value for five standardized item-fit indices. The following analyses were carried out using SPSS programs: (a) to check normalities of fit indices one sample Kolmogorv-Smirnov test was conducted. (b) the intercorrelations among five fit indices along with the difficulty levels were reported. The results of the first set of simulations are presented in Table 1, 2, and 3.

Table 1

A Descriptive Analysis of five Standardized Item-Fit Indices

Fit Index	Mean	Standard Deviation	Type I error rate	K-S Test
UWz	-.1	.8	1	.79
Wz	.0	.7	1	.74
Lz	.0	.7	1	.65
ECI2z	.0	.9	2	.49
ECI4z	.0	1.0	4	.54

Note: 1. N = 100 items, 1000 persons. 2. Type I error rate: value > |2| at the .05 level. 3. K-S Test: Kolmogorov-Smirnov Test.

As shown in Table 1, the means of five standardized item-fit indices are very stable about the expected value of 0.0. In the next column, a little variations of standard deviations are found in five fit indices. The standard deviations of three item-fit indices (UWz, Wz , and Lz) have slightly lower values ranging from 0.7 ~ 0.8 than the expected value of 1.0. On the other hand, the ECI2z index has a standard deviation of 0.9 and the ECI4z index has a standard deviation of 1.0. To examine the hypothesis that these standardized item-fit indices are close to the normal distribution, the Kolmogorov-Smirnov test were conducted using SPSS computer program. The results in Table 1 indicate that all five standardized item-fit indices do not depart significantly from a normal distribution. The Type I error rates of the four indices except the ECI4z index, however, do not match the normal distribution. It is expected that stan-

dardized item-fit indices greater than 2.0 or less than -2.0 will occur approximately 5% of the time when data fits the model. In these simulations, 5 in a sample of 100 items are expected as the Type I error rate at the 0.05 level. For the UWz, Lz, and Wz index, values greater than |2| occurred at a rate of 1 percent (1 out of 100 items). For the ECI2z index, values greater than |2| occurred at a rate of 2 percent (2 out of 100 items). This means that if the critical value of 2.0 were to be used, for the ECI2z index, then the Type I error rate would approximate 0.02. Therefore, the critical value of 2.0 could not be interpreted as unit normal. However, for the ECI4z index, values greater than |2| occurred at a rate of 4% (4 out of 100 items), which are not far from the rates expected under the normal distribution. Based on the results shown in Table 1, it appeared that the mean, standard deviation, and Type I error suggest that the ECI4z index is close to the use of normal approximation for obtaining cut-off values when data fits the model. The intercorrelations among the indices along with the difficulty levels are reported in Table 2 and 3.

Table 2

The Intercorrelations among five Standardized Item-Fit Indices

Fit Index	UWz	Wz	Lz	ECI2z	ECI4z
UWz	1.00				
Wz	.71	1.00			
Lz	-.83	-.97	1.00		
ECI2z	.85	.93	-.98	1.00	
ECI4z	.73	.94	-.95	.97	1.00

Note: All indices are significant at the .01 level.

Table 3

Correlations between Difficulty Levels and five Standardized Item-Fit Indices

Fit Index	Difficulty Level
UWz	.04
Wz	.01
Lz	.00
ECI2z	-.02
ECI4z	-.02

As shown in Table 2 five standardized item-fit indices are highly intercorrelated ranging from -0.83 ~ 0.97, which are significant at the 0.01 level. Table 3 shows all indices yield extremely low correlations with the difficulty level ranging from -0.02 to 0.04.

The effect of Sample Size on five standardized item-fit indices

In the next simulation, the number of persons was varied from 50 to 800 for test lengths, 50 items. The responses of 50 items were generated to fit the model. The item difficulties had uniform distribution ranging from -3 to +3 logits. The person abilities were normally distributed with a mean of zero and a standard deviation of one. The result is shown in Table 4.

Table 4

The Means and Standard Deviations of Item Fit Indices: Number of Persons varied

	Number of persons				
Fit Index	50	100	200	400	800
UWz	-0.1 (0.8)	-0.1 (0.8)	0.0 (0.9)	-0.1 (0.7)	0.0 (0.7)
Wz	-0.1 (0.7)	-0.1 (0.7)	0.0 (0.7)	0.0 (0.7)	0.0 (0.8)
Lz	0.0 (0.7)	0.0 (0.8)	-0.0 (0.7)	0.0 (0.7)	0.0 (0.8)
ECI2z	0.0 (0.9)	0.0 (1.0)	-0.0 (0.9)	0.0 (0.9)	-0.0 (0.9)
ECI4z	-0.0 (1.0)	-0.0 (1.0)	-0.0 (0.9)	0.0 (1.0)	0.0 (1.0)

Note: Each parenthesis indicates its own standard deviation.

Table 4 shows the result of varying the number of persons in the sample from 50 to 800 on a 50-item test. All indices reveal that there is little variation in the means over increased number of persons. However, it is interesting to note that the standard deviations of the UWz, Wz, and Lz index have slightly lower values than the expected value of 1.0. On the other hand, the standard deviation of the ECI4z index is very close to the expected value of 1.0 compared to other four indices.

Power of detecting differential item functioning (DIF)

The final analysis is involved with the power of five standardized item-fit indices to detect DIF. The procedures to conduct simulation of DIF were as follows. First, two groups (reference vs. focal) of 500 persons were generated using SIMTEST computer program, each with a mean ability of zero and a standard deviation of one logit. The number of items was set at 50. The item difficulties for the reference and focal group had uniform distribution with a standard deviation of one logit. Second, in order to simulate DIF, responses for the focal group were generated with item difficulties 0.5 and 1.0 logits more difficult than the item difficulties used with the reference group. The number of items that contain the bias was varied (10, 15, and 20 out of 50). The two groups were then combined for the analysis. Smith (1992, p. 177; 1994a, p. 46) originally dem-

onstrated the usefulness of this strategy specifically for item bias or DIF. Third, to compute the probabilities of observed scores, the ability and difficulty parameters were estimated by the BIGSTEPS computer program for each examinee and item. And then, the IRT parameters were used to calculate unweighted and weighted standardized fit index using BIGSTEPS computer program, standardized likelihood index, ECI2z, and ECI4z using IFIT computer program for the item-fit indices. The results are shown in Table 5, 6, 7, 8, and 9.

Table 5

Power of UWz Index for detecting DIF

The Amount of Bias	No. of Biased Items	Misclassified Items	Detected Items	Detected Percent
0.5	10	4	1	10.0
0.5	15	0	1	6.6
0.5	20	1	0	0.0
1.0	10	1	0	0.0
1.0	15	2	1	6.6
1.0	20	2	1	5.0

Note: Each line has 50 items and 1000 persons.

Table 6

Power of Wz Index for detecting DIF

The Amount of Bias	No. of Biased Items	Misclassified Items	Detected Items	Detected Percent
0.5	10	7	1	10.0
0.5	15	0	1	6.6
0.5	20	2	1	5.0
1.0	10	0	0	0.0
1.0	15	3	0	0.0
1.0	20	2	2	10.0

Note: Each line has 50 items and 1000 persons.

Table 7

Power of Lz Index for detecting DIF

The Amount of Bias	No. of Biased Items	Misclassified Items	Detected Items	Detected Percent
0.5	10	7	1	10.0
0.5	15	0	1	6.6
0.5	20	2	0	0.0
1.0	10	0	0	0.0
1.0	15	2	0	0.0
1.0	20	2	1	5.0

Note: Each line has 50 items and 1000 persons.

Table 8

Power of ECI2z Index for detecting DIF

The Amount of Bias	No. of Biased Items	Misclassified Items	Detected Items	Detected Percent
0.5	10	7	1	10.0
0.5	15	0	1	6.6
0.5	20	2	2	10.0
1.0	10	2	0	0.0
1.0	15	3	0	0.0
1.0	20	3	1	5.0

Note: Each line has 50 items and 1000 persons.

Table 9

Power of ECI4z Index for detecting DIF

The Amount of Bias	No. of Biased Items	Misclassified Items	Detected Items	Detected Percent
0.5	10	8	1	10.0
0.5	15	0	1	6.6
0.5	20	3	2	10.0
1.0	10	1	0	0.0
1.0	15	3	0	0.0
1.0	20	3	2	10.0

Note: Each line has 50 items and 1000 persons.

As shown in Table 5, 6, 7, 8, and 9 the results indicate that the power of detecting DIF when the reference and focal groups contain 500 persons remains unchanged across the six conditions. In most cases, the percentage of biased items detected is less than 10%. Further, the Type I error rate for the unbiased items is not reasonable ranged from 0% to 20% for five standardized item-fit indices although 5% rejection rate would be expected under the null hypothesis. The only conclusion that can be drawn from Table 5, 6, 7, 8, and 9 is that these indices are quite ineffective index to detect DIF.

Conclusions

With regard to five standardized item-fit indices, three conclusions can be drawn. First, although these five standardized item-fit indices did not depart significantly from a normal distribution, it appeared that the Type I error rates were not reasonable. This result implies that the assumption associated with the |2| cut-off value is not warranted and that adjustments should be made in the cut-off values. This finding regarding

cut-off values,however, may capitalize on chance because this study is based on the single series of simulated data. Second, all five indices are highly intercorrelated and yield low correlation with difficulty level. Further, for the effect of sample size, the ECI4z index is appeared more stable than other indices over increased number of persons. Third, for the power of five standardized item-fit indices to detect DIF, the results show that all indices did perform poorly across various conditions. Therefore, it is possible to conclude that all indices used in this study are inadequate fit measures for detecting DIF. With respect to the UWz or Wz index, these findings are expected since Smith (1991a) already reported that total fit index was ineffective index for detecting biased items. In this context, it is interesting to note another fit index called "between fit" index. The "between fit" index is only available in the IPARM computer program (Smith, 1991b). The most distinctive feature of "between fit" index, claimed by Smith (1994a), is that because this fit index is based on relevent criterion subgroups, this makes it possible to divide total groups into subgroups, such as sex or race, to detect subgroup differences usually defined as bias. Smith (1994a) also demonstrated that between fit index is an useful tool to detect systematic aberrant response patterns. The proposed study did not attempt to examine the distributional properties and power of between fit index because Smith (1991a, 1994a, 1994b) already conducted a series of study concerning between fit index. In conclusion, another fit measures for item bias or differential item functioning analysis (e.g., between fit index) should be applied or, if necessary, developed under the Rasch model.

References

Blixt, S. L., and Dinero, T. E. (1985). An initial look at the validity of diagnoses based on Sato's caution index. *Educational and Psychological Measurement, 45*, 293-299.

Donlon, T. F., and Rindler, S. E. (1979). *Consistency of item difficulty for individuals and groups in the Graduate Record Examination*. Paper presented at the annual meeting of the American Educational Research Association. San Francisco.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement, 6*, 297-308.

Drasgow, F., and Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10*, 59-67.

Drasgow, F., Levine. M. V., and McLaughlin, M. E. (1987). Detecting inappro-

priate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59-79.

Drasgow, F., Levine, M. V., and Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Glaser, R. (1952). The reliability of inconsistency. *Educational and Psychological Measurement*, 60-64.

Glaser, R. (1949). A methodological analysis of the inconsistency of responses to test items. *Educational and Psychological Measurement, 9*, 721-739.

Harnisch, D. L., and Linn, R. L. (1981a). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*, 133-146.

Harnisch, D. L., and Linn, R. L. (1981b). *Identification of aberrant response patterns*. (Final Report on Grant No. G-80-0003) Washington. D. C. : National Institute of Education.

Harnisch, D. L., and Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. K. Hambleton (ED.), *Applications of item response theory*, Vancouver, B.C.: Educational Research Institute of British Columbia.

Levine, M. V., and Drasgow, F. (1982). Appropriateness measurement: review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35*, 42-56.

Levine, M. V., and Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics. 4*, 269-290.

Linacre, J. M., and Wright, B. D. (1993). BIGSTEPS: *Rasch-Model Computer Program.* Chicago:MESA Press.

Luppescu, S. (1993). *SIMTEST: ver 2.3.* Chicago: MESA Press.

Masters, G. N. (1988). Item discrimination: when more is worse. *Journal of Educational Measurement, 24*, 15-29.

Meijer, R. R. (1996). Person-fit research: an introduction. *Applied Measurement in Education, 9*, 3-8.

Mosier, C. I. (1940). Psychophysics and mental test theory: fundamental postulates and elementary theorems. *Psychological Review, 47*, 355-366.

Reise, S. P. (1990). A comparison of item-and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*, 127-137.

Rost, J., and Daver, M. (1994). A conditional item-fit index for Rasch models. *Applied Psychological Measurement, 18*, 171-182.

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement. 18*, 171-182.

Scheuneman, J. (1975). *A new method of assessing bias in test items.* Paper presented at American Educational Research Association, Washington, D.C.

Seol, H. (1997). IFIT:*A Microsoft Fortran program for item-fit statistics in the Rasch model.* Unpublished computer program.

Smith, R. M. (1982). *Detecting measurement disturbances with the Rasch model.* Unpublished doctoral dissertation, University of Chicago.

Smith, R. M. (1985). *Validation of individual test response patterns.* International Encyclopedia of Education. Oxford: Pergamon Press, 5410-5413.

Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement, 46*, p. 359-372.

Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological Measurement, 48*, 657-667.

Smith, R. M. (1991a). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement, 51*, 541-565.

Smith, R. M. (1991b). *IPARM: item and person analysis with the Rasch model.* Chicago: MESA Press.

Smith, R. M. (1992). *Applications of Rasch measurement.* Chicago: MESA Press.

Smith, R. M. (1994a). A comparison of the power of Rasch total and between-item fit statistics to detect measurement disturbances. *Educational and Psychological Measurement, 54*(1), 42-55.

Smith, R. M. (1994b). Detecting item bias in the Rasch rating scale model. *Educational and Psychological Measurement, 54*, 886-898.

Smith, R. M., Schumaker, R. E., and Busch, M. J. (1994c). *Examing replication effects in Rasch fit statistics.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Smith, R. M., Schumaker, R. E., and Busch, M. J. (1995). *Using item mean squares to evaluate fit to the Rasch model.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*(1), 95-110.

Tatsuoka, K. K., and Linn, R. L. (1983). Indices for detecting unusual patterns: links between two general approaches and potential applications. *Applied Psychological Measurement, 7*(1), 81-96.

Tatsuoka, K. K., and Tatsuoka, M. M. (1982a). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7*, 215-231.

Tatsuoka, K. K., and Tatsuoka, M. M. (1982b). *Standardized extend caution indices and comparisons of their rule detection rates.* Urbana, Illinois: Computer-based Education Research Lab.

Thurstone, L. L., and Chave, E. J. (1929). *Measurement of attitude*. Chicago: University of Chicago Press.

Van der Flier, H. (1977). Environmental factors and deviant response patterns. In Y.H. Poortinga (Ed.), *Basic problems in cross-cultural psychology*. Amsterdam: Swets and Seitlinger, B.V.

Van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*, 267-298.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement. 14*, 97-116.

Wright, B. D., and Panchapakesan, N. A. (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement*, *29*, 23-48.

Wright, B. D. and Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Appendix

Smith (1991a) reported the unweighted and weighted total fit statistic. The formula for the unweighted total item-fit mean square, $MS(UT)_i$, is given as

$$MS(UT)_i = \frac{1}{N} \sum_{n=1}^{N} [(X_{ni} - P_{ni})^2 / P_{ni}(1 - P_{ni})] \quad (1)$$

The weighted total item-fit mean square, $MS(WT)_i$, is expressed as

$$MS(WT)_i = \sum_{n=1}^{N} (X_{ni} - P_{ni})^2 / \sum_{n=1}^{N} W_{ni} \quad (2)$$

where X_{ni} is the observed response for person n to item i, P_{ni} is the expected response for person n to item i, and $W_{ni} = P_{ni}(1 - P_{ni})$. These fit indices can be standardized to an approximate unit normal (0,1) by a cube root transformation, which is given by the following formula:

$$t = (MS^{\frac{1}{3}} - 1)(3/S) + (S/3)$$

Where S is the standard deviation of MS given above.

The second method is based on a likelihood function. Levine and Rubin (1979) proposed "appropriateness measurement" and this index was standardized by Drasgow et al.(1985). The standardized likelihood index can be expressed as

$$L_z = \{L(\theta) - E[L(\theta)]\} / \{VAR[L(\theta)]\}^{\frac{1}{2}} \quad (3)$$

Where

$$L(\theta) = \sum_{n=1}^{N} \{U_n[\ln P_n(\theta)] + (1 - U_n)[\ln Q_n(\theta)]\}$$

$$E[L(\theta)] = \sum_{n=1}^{N} \{ P_n(\theta)[\ln P_n(\theta)] + Q_n(\theta)[\ln Q_n(\theta)] \}$$

$$VAR\,[\,L(\theta)\,] = \sum_{n=1}^{N} P_n(\theta)\, Q_n(\theta) \{ \ln [\, P_n(\theta) / Q_n(\theta)] \}^2$$

U_n is the dichotomous person response, $P_n(\theta)$ is the probability of a correct response given θ,

$Q_n(\theta) = 1 - P_n(\theta)$, and N is the number of persons.

The third approach involves Extended Caution Indices (ECI). The standardized caution indices, ECI2z and ECI4z mathematically proven by Tatsuoka (1984), are given by the following formula (Drasgow et al., 1987, p. 65).

$$ECI2z = \sum_{i=1}^{n} [\, P_{ij}(\theta) - U_{ij}\,][\, G_i - \bar{G}\,] / [\sum_{i=1}^{n} P_{ij}(\theta)Q_{ij}(\theta)(G_i - \bar{G})^2\,]^{\frac{1}{2}} \qquad (4)$$

$$ECI4z = \sum_{i=1}^{n} [\, P_{ij}(\theta) - U_{ij}\,][\, P_{ij}(\theta) - \bar{P}_j\,] / [\sum_{i=1}^{n} P_{ij}(\theta)Q_{ij}(\theta)(P_{ij} - \bar{P}_j)^2\,]^{\frac{1}{2}} \qquad (5)$$

Where

i = Person (1 ...n), j = Item (1 ... N), U_{ij} is the observed response, and P_{ij} is the probability of a correct response, $Q_i(\theta) = 1 - P_i(\theta)$.

$$G_i = \frac{1}{N} \sum_{j=1}^{N} P_{ij}(\theta)$$

$$\bar{G} = \frac{1}{n} \sum_{i=1}^{n} G_i$$

$$\bar{P}_j = \frac{1}{n} \sum_{i=1}^{n} P_{ij}(\theta)$$

JOURNAL OF OUTCOME MEASUREMENT®, 3(3), 248–265

Developing a Unidimensional Instrument to Measure the Effectiveness of School-based Partnerships

Deborah L. Bainer
The Ohio State University, Mansfield

Richard M. Smith
Rehabilitation Foundation, Inc.

The purpose of this study was to refine an instrument designed to measure a single construct, the effectiveness of school-based partnerships. The instrument was designed to measure the "health" of the partnership teams and to identify specific problems for which intervention might be appropriate. The items were based on four theoretical models of partnering efforts. The partnerships studied were created to enhance the teaching of elementary school science and involved elementary teachers and resource professionals in school-based programs over a six-year period. The results show how Rasch analysis, using the item and person fit statistics, bias analysis using separate calibration groups for contrasts of interest, and principal component analysis can be used to evaluate the unidimensionality of a scale.

Requests for reprints should be sent to Deborah L. Bainer, The Ohio State University of Mansfield, 1680 University Drive, Mansfield, OH 44906, e-mail: bainer1@osu.edu.

This study was funded in part by grants from the Ohio State University, Mansfield, Professional Development Program and the U.S. Department of Education, OERI, Dwight D. Eisenhower Professional Development Program for Mathematics and Science Education.

Partnerships between schools and agencies have been advocated as a vehicle for professional development and education reform for the past two decades. Four models of group efforts, such as partnerships, appear in the literature which may inform the development of partnerships and help differentiate effective from less effective partnership efforts (Bainer, 1996; Cobb and Quaglia, 1994; Rigden, 1991; Wichienwong, 1988; Wright, 1994). A problem arises, however, in that the characteristics of effective partnerships remain theoretical and untested, especially in the elementary school context (Bainer, 1998). Further, because individual partnerships display unique organizational and personal interactions, it is not known why some partnerships succeed and persist while others soon disband (Miron and Wimpelberg, 1989). Cobb and Quaglia (1994) point out that we need to know more about school-based partnerships in order to ensure successful education reform.

This problem is aggravated in that a scale to measure the construct of school-based partnerships does not exist. An invariant metric is needed to measure this construct across different groups of people, including elementary school teachers and resource professionals from a variety of businesses, industries, agencies and health care organizations who volunteer to join with teachers to form partnership teams. Application of Rasch analysis holds promise to assist in the construction of this scale. Therefore, the purpose of this study was to develop and refine an instrument with one or more unidimensional scales to measure the effectiveness of school-based partnerships.

For the purposes of this paper, the term "partnership" refers to a long-term commitment (at least one year) among one or more elementary teachers and one or more volunteers from a business, industry, health care organization, or government agency to work closely together to improve science instruction in one or more elementary school classrooms. In the partnership teams thus formed, the teachers provide expertise in pedagogy and learning while the volunteers bring expertise in science content and access to resources and a network within the scientific community. Research has shown that these reform-based partnerships are effective at changing classroom instruction in the direction of what is considered "good" science teaching (Bainer, 1997; Bainer, Barron, and Cantrell, 1995; Bainer, Barron, and Cantrell, 1996/97) and at improving student achievement and influencing students' curricular and career choices (Science and Mathematics Network, 1996). Further, volunteers who engage in part-

nerships self-report enhanced job-related skills and competencies as a result of working with teachers and students (Bainer, Halon, and Williams, 1995; Bainer, Cantrell, and Barron, 1997).

Method

The purpose of this study was to refine an instrument to measure a single construct, the effectiveness of school-based partnerships. This instrument would measure the "health" of partnership teams and identify specific problems for which intervention might be appropriate.

Participants

Elementary school (K-6) teachers and volunteers who participated in partnering programs in one Midwestern state over the past five years were participants in the study. A total of 301 participants from 16 counties returned survey questionnaires, of which 297 were useable for the analysis. Of these, 220 (74.6%) identified themselves as females and 75 (25.4%) identified themselves as males. Further, 181 (60.9%) identified themselves as teachers and 62 (20.8%) reported working as volunteer resource professionals. Teachers' experience ranged from 1-35 years, with a mean of 14.2 years and a median and mode of 13 and 12 years respectively. Resource volunteers' experience ranged from 1-50 years, with a mean of 12.0 years, and a median and mode of 10 years.

Partnership teams were formed under three distinct yet parallel funded programs. Fifty-eight participants (19.3%) were involved in partnership teams formed as part of Sciencing with Watersheds, Environmental Education, and Partnerships (SWEEP), a federally-funded partnership program conducted during 1995-1998. One hundred thirty-three participants (44.2%) were involved with the state-funded Partnering for Elementary Environmental Science (IKE) program established during 1992-1995. The remainder of the partnerships, involving 110 participants (36.5%), were formed during the same time period by the Science and Mathematics Network of Central Ohio (Network).

Most participants (279, or 93.0%) had no partnering experience prior to the formation of their team. About half (146, or 48.7% of respondents) had experience team teaching, and many had worked on collaborative projects (212 participants, or 70.7%) or committees (229 respondents, or 76.3%) on their job. Many had worked collaboratively on community projects (207, or 69.0%) or community committees (150, or 50.0%). Few,

however, (85 participants, or 28.3%) had completed university course work in partnering, mentoring or supervising others although nearly half (132 participants or 44.0%) had taken job-related workshops or seminars on partnering, team building, or mentoring.

The partnership teams in which the participants were engaged ranged in size from 1-9 members, with a median team size of 4.1. The majority (80%) of the teams, however, had between two and five members. Nearly half of the teams (48.8%) had been in existence less than one year at the time of this study. Another quarter (22.1%) were in their second year of involvement. Other teams ranged up to six years of partnering. Most participants described the current level of activity of their team as "active or on schedule" (140 participants, or 47.1%), while others said their teams currently had "limited activity" (69 participants, or 23.2%) or were "very active and ahead of schedule" (48 participants, or 16.2%). Reportedly, 40 participants in the study (13.5%) were members of teams which had disbanded or were not currently active. When reflecting on their past partnership experience, 238 respondents (82.6%) reported that their team was "active" or "very active" during its first year, while 44 persons (15.3%) reported "limited activity" during their first year of involvement with the team. Only six individuals (2.1%) reported that their teams disbanded during their first year.

About one third (33.2%) of the participants reported that their entire team met formally on a monthly basis to plan, evaluate, or work on classroom-related activities. Fewer (64 participants, or 21.7%) reported formal meetings held regularly on alternate months while others (73 participants, or 24.7%) reported that their partnerships rarely met formally. Other active teams reported meeting twice monthly (10.8%) or weekly (4.7%). Informal interactions about the partnership team's activities, such as spontaneous conversations, needs-driven interactions such as phone calls or exchanges in the hallway or over lunch which usually did not involve the entire team, took place more frequently. These occurred monthly (64 respondents, or 21.7%), twice monthly (50 participants or 16.9%), weekly (73 participants, or 24.7%), or even daily (21 participants, or 7.1%). For other participants, informal exchanges occurred rarely (33 participants, or 11.2%) or regularly (45 participants, or 15.3%).

Overall, most participants reportedly were satisfied (159, or 51.1%) or highly satisfied (116, or 37.3%) with partnering. Further, most (206, or 65.8%) described themselves as "very confident" with partnering and the

partnering process; and while others were "somewhat confident," none of the respondents reported being "not confident at all." Finally, most participants (201, or 67.1%) said they would definitely recommend involvement in a partnership to their professional colleagues. Others (100, or 32.9%) said that they would recommend partnering "under certain conditions or situations." None said that they would not recommend participating in a partnership to others.

Instrument

The instrument used in this study, the *Models of Partnership Survey (MOPS),* was developed to measure the construct of effective partnering. Each question represented a characteristic of effective group efforts operationalized by one of the four theoretical models of group efforts identified in a literature search (Bainer, 1998). Wichienwong's model (1988) discusses group efforts as either cooperative or collaborative, based primarily on the level of involvement of the members in a partnership. Cobb and Quaglia (1994) provide a model of group efforts based on the level of interaction among participants in a partnership. They differentiate between group efforts which are partnership-driven vs. relationship-driven. Rigden (1991, 1992) discusses a continuum based on the level of impact which the effort has on schools, instruction, and student learning. Pivotal points along the continuum are "helping hands" partnerships which are minimally involved, "project driven" partnerships which are more focused, and "reform-based" partnerships which strive for systemic change. Finally, Wright (1994) explores the situational nature of group endeavors and develops a model based on their level of organization. Wright develops a mathematical model of group productivity based on Rasch measurement, describing "teams" as unions working together in perfect agreement, "packs" which work as collections of perfect disagreements, and "chains" which work as connections of imperfect agreements.

The first part of the instrument contained 26 demographic questions eliciting information about the structure of the partnership team (size, activity, location, duration, etc.), the relationships among its members (formal and social interactions, methods of interaction, etc.), and the satisfaction and confidence of team members. The purpose of the demographic questions was to distinguish among effective and less effective partnerships teams. The second part of the instrument consisted of 78 Likert-type questions with four response categories (strongly disagree, disagree, agree, strongly agree). This part of the instrument was to measure the

theoretical construct of effective partnering operationalized in the four models in the context of elementary schools. Of the 78 items, 56 were positively worded so that "strongly agree" resulted in the highest score and 22 were negatively worded so that "strongly disagree" resulted in the highest score, indicating the most effective partnering characteristic.

The instrument was piloted on a group of 26 experienced teachers involved in partnerships. The instrument was subsequently revised to reflect changes in wording and format recommended by the teachers. Two forms of the questionnaire were developed so that the items could be presented in different order to control for order effects. The instrument was administered via mail, using three rounds of mailings as recommended in the Total Design Method (Dillman, 1978). Of the 477 questionnaires mailed, 24 proved to be duplicate mailings, 12 were sent to individuals who said they never were engaged in partnership programs, 4 were sent to participants who had relocated and could not be located, and 2 of the program participants were deceased. Of the adjusted sample of 435, 134 did not respond after three rounds of mailings. A total of 301 participants (69.2%) returned survey instruments of which 297 were entered in the analysis.

Data Analysis

Data were analyzed using the BIGSTEPS program (Wright and Linacre, 1996) and SPSS. The total data set was subjected to Rasch calibration and principal component analysis. Rasch fit statistics were used to detect the presence of multi-dimensionality. On the basis of these findings, an analysis of the reduced data set was performed. Reanalysis was based on deleting misfitting persons, then on deleting misfitting items. Results from the analysis deleting misfitting persons and items indicated two possible factors, which were analyzed separately.

To ensure that the factors identified were the result of differences in items and not a result of differences in the person responses to the questionnaire, a variety of subpopulation calibrations were performed. These calibrations were based on: 1) form of the instrument (Form 1 vs. Form 2), 2) gender (male vs. female), 3) profession (teachers vs. resource professionals), and 4) training program group (SWEEP vs. IKE vs. Network). A t-test procedure (Wright and Stone, 1979) was used to determine differences in subpopulation calibration difficulties for each item.

The unidimensionality of the instrument was investigated by applying principal components factor analysis as described by Banerji, Smith,

and Dedrick (1997). An unrotated principal component analysis was used since the varimax rotation has been shown to make the factor structure difficult to interpret (Smith, 1996).

Results

The results of the calibration of the full set of items and persons with BIGSTEPS yielded a real person separation index of 0.96 and a real item separation index of 0.97. The magnitude of these values suggests that the persons are adequately separating the items along the difficulty continuum and that the items are adequately separating the persons along the same difficulty continue. These results suggest that there is a variable that predominates the analysis and that further investigation is warranted to determine if that variable is unidimensional. The distribution and the magnitude of the item and person fit statistics can often give an indication of the dimensionality. The mean person unweighted total fit (OUTFIT) was -0.40 with a standard deviation of 3.1. This statistic is the cube root transformation of the person unweighted mean square and has an approximate normal 0, 1 distribution (Smith, 1991b). The magnitude of the standard deviation is a sign of an unusually large number of large values. In fact, 66 of the 297 persons had an unweighted total fit statistic greater than 2.00. This indicates that the responses for about 22% of the respondents were inconsistent with the overall definition of difficulty from the calibration. The mean item unweighted total fit was -0.30 with a standard deviation of 3.8. This statistic is the cube root transformation of the item unweighted mean square and has an approximate normal 0, 1 distribution (Smith, 1991b). Again the magnitude of the standard deviation is an indication that there are an unusually high number of items with large fit values. In fact, 18 of the 78 items had unweighted total fit statistics greater than 2.00. This indicates that the responses to about 23% of the items were inconsistent with the overall definition of difficulty from the calibration. The Type I error rate for these statistics is in the 3-4% range (Smith, 1991a), thus indicating that there is a considerable amount of misfit in these data. The unexpectedly large amount of person and item misfit may be the result of multidimensionality in the instrument.

To further investigate the possibility of multidimensionality, a principal component analysis of the instrument was performed using SPSS. The results of the principal component analysis indicated that there was a single strong first factor with seven weaker factors also present. The eigen-

value for the first factor was 29.26 and it accounted for 37.5% of the total variance. The eigenvalue for the second factor was 4.59. The addition of this factor raised the explained variance to 43.4 %. The eigenvalue for the third factor was 2.56, and the addition of this factor raised the explained variance to 46.7 %. By the eighth factor the eigenvalue was down to 1.47 (a magnitude that could occur by chance, according to Smith and Miao, 1994), with an explained variance of 61%.

A total of 53 items (68%) had a factor loading greater than 0.50 on factor one. Of the remaining items, 7 had factor loadings less than 0.30 and 18 items had factor loadings between 0.30 and 0.50. There was a high correlation between the items with a low factor loading on factor one and the misfitting items identified in the BIGSTEPS analysis. All 8 of the items with factor loadings less than 0.30 had unweighted total fit values greater than +2.0. Of the 11 items with factor loadings between 0.30 and 0.50, seven had unweighted total fit values greater than +2.0. Overall 14 of the 18 items with factor loadings less than 0.50 on factor one misfit (t>2.0) in the Rasch calibration and one item with a factor loading less than 0.50 had a fit standardized value of +1.90. The marked similarity of the results from the Rasch calibration and principal component analysis should be expected (Smith, 1996; Wright, 1996).

The second factor in the principal component analysis identified two clusters of items. There were 10 items with factor loadings greater than 0.30 and 11 items with factor loadings less than -0.30. The content of these items is listed in Table 1. Initial inspection of the item numbers might suggest a boredom/fatigue factor with the earlier items loading positively on the second factor and the later items loading negatively on that factor. However, this is not a plausible suggestion since the questionnaire was administered in two scrambled forms. A more likely explanation is suggested by the wording of the items, or the characteristic of effective partnering which they operationalize. The 10 items with positive loadings on the second factor tend to describe the role of the various members within the partnership. These items were derived from the Wichenwong (1988) model of group efforts. The 11 items negatively loading on the second factor talk about roles or characteristics of the partnership itself. These are derived from characterisitcs of reform-based partnerships in the Rigden (1992) model. Because these roles were carried out within the school setting, partnership members who were not teachers may not have understood the education jargon in these questions. Further, they may not

Table 1

Text of Items Loading on Second Factor

Cluster 1 - Positive loading items (>.30)

2 In my partnership, some participants contributed relatively little. (.48)
3 In my partnership, all members of the group worked together from the initial planning stage, throughout the implementation of our program, and through final evaluation and reflection activities. (.42)
5 In my partnership, all members helped develop the goals and expectations for the partnership. (.35)
9 In my partnership, all members seemed equally committed to bringing about an effective partnership. (.31)
10 In my partnership, all members contributed equally during the initiation stage to identify topics and resources, establish a working structure, and get to know each other. (.42)
11 In my partnership, all members contributed equally during early planning to design our action plan and partnering program, identify roles for each member, identify and secure resources, equipment, and facilities, and complete general plans for the classroom. (.39)
12 In my partnership, all members contributed equally to implement our action plan by developing detailed lessons, presenting activities, modifying the plan as needed, and communicating with each other and others in the building and district. (.36)
13 In my partnership, all members contributed equally to evaluating and reflecting on the effectiveness of our lessons, action plan, and partnership. (.33)
14 In my partnership, all members contributed equally in preparing interim and final reports about our activities and effectiveness when required. (.32)
19 In my partnership, I feel I provided more leadership or direction than others did. (.31)

Cluster 2 - Negatively loading items (<-.30)

59 My partnership could best be described as innovative and pioneering. (-.32)
61 My partnership could best be described as exhibiting a classroom- or student-focused program. (-.37)
62 My partnership could best be described as providing hands-on activities for students. (-.32)
63 My partnership could best be described as providing age appropriate information and activities for students. (-.31)
64 My partnership could best be described as reflecting national education goals. (-.37)
65 My partnership could best be described as reflecting state education and curricular goals. (-.36)
67 My partnership could best be described as bringing about change in the classrooms involved. (-.37)
68 My partnership could best be described as bringing about change in the school(s) involved. (-.39)
73 My partnership could best be described as providing resources for school programs. (-.33)
77 My partnership could best be described as changing attitudes in the classroom and school. (-.40)
78 My partnership could best be described as changing classroom practices in the

have been able to observe if, indeed, the day-to-day operation of the classroom had been affected by the partnership in these ways. Since the factor analysis suggested a dimensionality problem, closer investigation was required to identify what contributed to this multidimensionality.

Two aspects of the design of the original questionnaire were identified that might contribute to multidimensionality. The first was that the questionnaire was administered as two forms with scrambled items. The second was that there were 56 positively worded items and 22 negatively worded items on the questionnaire. Further, there was concern that differences in the persons responding (ie., teachers vs. volunteers) might cause the misfit. There were three factors investigated here: gender of the respondent, the respondents role in the partnership (teacher vs. volunteer resource person), and in which of the similar partnership training programs the respondents participated (SWEEP, IKE, or Network training).

To investigate the potential differences in item difficulty due to form design differences, separate calibrations for persons responding to Form One and persons responding to Form Two were conducted. The item difficulties resulting from the two calibrations were then compared using the t-test approach (Wright and Stone, 1979). This analysis indicated that 15 of the 78 items had t-values greater than |2.00|. This represents 19% of the items, about four times the expected Type I error rate. To address the differences between male and female responses, separate calibrations for those two groups were preformed and the resulting item difficulties compared with the t-test approach. This analysis indicated that 14 of the 78 items (18%) had t-values greater than |2.00|. Again this is about 4 times the expected Type I error rate. For the role in partnership analysis, 13 of the 78 items (17%) had t-values greater than |2.00|. For the training program analysis, 3 separate t-tests were performed since there were 3 different training programs and the t-test procedure only allows pair-wise comparisons. The results of the three t-tests indicated that 12 (15%), 8 (10%), and 8 (10%) items had t-values greater than |2.00|. Although there were unexpectedly large numbers of items with t-values greater than|2.00|, there was little disordering in the definition of the variable in any of the comparisons. In addition, the same items did not have t-values greater than |2.00| across analyses. This was taken as an indication that the misfit in the original calibration was not due to the use of different forms of the questionnaire, or to differences in the definition of the variable for males and females, nor to different partnership training programs.

To investigate the possibility that the positive and negative wording of the items might be the cause of the misfit, the item fit values in the original calibration were reviewed. Of the 22 negatively worded items, 15 had a unweighted total fit value greater than +2.0. Comparing the negatively worded items with the items that had factor loadings less than 0.50 on the first factor showed 13 of the 18 items that did not load on this factor were negatively worded items. These two findings are a strong indication that the dimensionality problem was due to the presence of the negatively worded items.

There is always the possibility that the misfitting persons identified in the original analysis were responding in a random fashion and that their responses caused the item misfit and altered the definition of the variable. To ensure that this was not the cause of the misfit, a separate calibration was performed omitting the 112 persons who had large misfit values on the first calibration. The results of this calibration showed no real difference in the person separation reliability, which remained at 0.96, or in the item misfit. The mean item unweighted total fit was -0.60 and the standard deviation was 3.3. Both of these values were close to those in the original calibration (-0.30 and 3.8). There were 16 items with unweighted t-values greater than +2.0, down two from the original analysis. All 16 of the misfitting items were identified in the original analysis. Finally, the item difficulties from this analysis were compared to the original analysis using the t-test approach. In this comparison, only 1 of the 78 items had a t-value greater than |2.0|. The combination of these results suggests strongly that the item misfit in the original calibration was not due to the presence of the misfitting persons.

To check the possibility that the misfit was unrelated to the negative wording, the 19 worst fitting items from the original total person calibration were delated from the calibration that included all of the sample. These results showed little change in the person and item separation reliability from the first calibration (PSR: .96 vs. .97; ISR: .97 vs. .96). The person fit improved slightly. The means were the same, -0.4, but the standard deviation of the reduced item set was slightly better at 2.9 vs. the 3.1 in the original calibration. As expected, the mean of the item fit moved slightly from -0.3 to -0.2, but more importantly the standard deviation was reduced from 3.8 in the original calibration to 2.1 in the reduced item set calibration. Despite removing the 19 worst fitting, items there were now 9 new items of the 59 in the calibration with unweighted total fit greater than +2.0. Of

these, five of the nine were negatively worded items. When added to the 15 negatively worded items detected as misfitting in the first calibration, this brings the total to 20 of the 22 negatively worded items misfit on the two analyses. This adds strength to the hypothesis that the negatively worded items are causing the multidimensionality in the form that is picked up in the item fit statistics and the principal component analysis.

To further test this hypothesis, a calibration of the total sample was again performed, this time with 26 items deleted, the 22 negatively worded items and 4 positively worded items that were identified as misfitting in the first analysis. Table 2 shows a list of the item numbers, reverse coding status, and factor loadings on the first factor in the principal component analysis. Table 3 shows the wording for these 26 items. The results

Table 2

Items Deleted from the Revised Instrument

Item Number	Misfit<2.0	Reverse Coded	Fact.Load.<.50
2	*	*	*
8	*	*	0.56
15	-2.2	*	0.56
16	*	*	0.52
17	*	*	*
18	*	*	*
21	1.2	*	*
22	*	*	*
23	*	*	*
24	*	*	0.51
31	-0.6	*	*
34	-0.1	*	*
35	*	*	*
38	1.9	*	*
40	1.3	*	0.61
45	*	*	*
46	*		*
47	-0.3	*	0.51
56	*		*
57	*	*	*
60	-2.7	*	0.67
66	*	*	*
70	*		*
71	*	*	0.58
72	0.1	*	0.64
75	*		*

* = Yes, exceptions are shown with actual values.

Table 3

Text of Items Deleted from the Revised Instrument

2	In my partnership, some participants contributed relatively little.
8	In my partnership, I spent more time working by myself than working with others in the group or in group meetings.
15	In my partnership, I felt excluded from much of the planning and other activities.
16	In my partnership, I largely planned and developed the lessons and activities I presented to the students on my own.
17	In my partnership, the lessons and activities I presented to the students were assigned or given to me by others in the group.
18	In my partnership, I felt I provided more leadership or direction than did others.
21	In my partnership, I felt somewhat left out of the communication, such as being informed about changes in schedules or programs.
22	In my partnership, the action plan and related activities were conceptualized and implemented mainly by the teachers.
23	In my partnership, the resource people serve primarily as consultants, guest speakers, or aides for activities which the teachers lead.
24	In my partnership, some members have more influence or power than do others.
31	In my partnership, I should spend more time involved in the joint efforts of our partnership.
34	In my partnership, time is not used productively.
35	In my partnership, insufficient time is allocated for planning and the class-related program.
38	My partnership could best be described as overly structured and inflexible.
40	My partnership could best be described as one-dimensional, or dominated by one group or person's agenda.
45	My partnership could best be described as significantly altering my work schedule.
46	My partnership could best be described as significantly changing the way I teach or work with groups of people.
47	My partnership could best be described as not very concerned about feedback and evaluation.
56	My partnership could best be described as task-oriented.
57	My partnership could best be described as concerned about image and publicity.
60	My partnership could best be described as lacking in mutual, clearly defined goals.
66	My partnership could best be described as utilizing traditional assessment strategies.
70	My partnership could best be described as supported from multiple levels within multiple agencies.
71	My partnership could best be described as short-term.
72	My partnership could best be described as ineffective.
75	My partnership could best be described as addressing a specific academic problem within the school.

of this calibration again show little change in the item or person separation reliability despite excluding approximately 33% of the items (person separation reliability = 0.96, item separation reliability = 0.96). Again the person fit improved slightly, with the new mean unchanged at -0.4 again, but with a reduced standard deviation of 2.9. The fact that all of

the negatively worded items could be removed without any loss in person separation reliability again suggests multidimensionality based on negative item wording.

To further test this hypothesis, the negatively worded items were calibrated separately for the total sample. The results of this calibration yielded a person separation reliability of 0.84 and item separation reliability of 0.97. The mean person unweighted total fit was -0.2 with a standard deviation of 1.8. The mean item unweighted total fit was 0.0 with a standard deviation of 2.8. The large standard deviation for the item fit suggests that there are misfitting items remaining in this subset. Four items from this subset had item fit values greater than 2.0. This suggested that these items were not working effectively with the remaining negatively worded items.

To further refine the negatively worded item set, the four misfitting items were deleted and the calibration rerun with the 18 best fitting items. The separation reliability statistics remained virtually unchanged when the four items were deleted. The item separation reliability was 0.97, while the person separation reliability increased slightly to 0.86 from 0.84. The person fit remained unchanged (mean = -0.3, S.D. = 1.8). However the item fit improved (mean = 0.3, S.D. = 1.3).

To further investigate separately the dimensionality of the positively worded and negatively worded items, two additional principal component analyses were performed. The analysis of the 18 best fitting negatively worded items showed one major factor with an eigenvalue of 6.37 accounting for 35.4% of the variance. Three additional factors were extracted with eigenvalues of 1.58, 1.22, and 1.05. These three values are low enough to have been created by chance. All of the 18 items loaded on the first factor, with the lowest loading being .48. Only three of the items had factor loadings lower than 0.50. The principal component analysis of the 52 best fitting positively worded items showed one main factor with an eigenvalue of 24.29 and accounting for 46.7 % of the variance. A total of seven additional factors had eigenvalues greater than 1.0, however only three of these had values greater than 1.5: factor two (3.44), factor three (1.84), and factor four (1.67). All of the items loaded on the first factor with the smallest loading being 0.49, the only item with a factor loading less than 0.50. These results seemed to support the hypothesis that there are two primary factors consisting of the 52 best fitting positively worded items and the 18 best fitting negatively worded items.

To investigate the relationship between these two factors, a person measure file was created during the last calibration run for the positive worded and negatively worded best fitting sets. The two person files were read into SPSS to determine the correlation between the two measures. The unattenuated correlation between the measures from the best fitting positively worded items and the best fitting negatively worded items is 0.80 (not much lower than the person separation index value of .86 for the 18 negatively worded items). A plot of these measures is shown in Figure 1. It is clear that these two measures are strongly related, enough so that a sum of these scores might be warranted. (It can't be any worse than adding the Mathematics, English, Social Studies, and Science scores together on the ACT Assessment to arrive at a total score).

After all of these analyses it is possible to conclude that there is a 70 item scale, that can be reported either as a single score or two subscores that adequately reflects the effectiveness of the school-based partnerships. The decision as to one total score or two subscores seems more philosophical

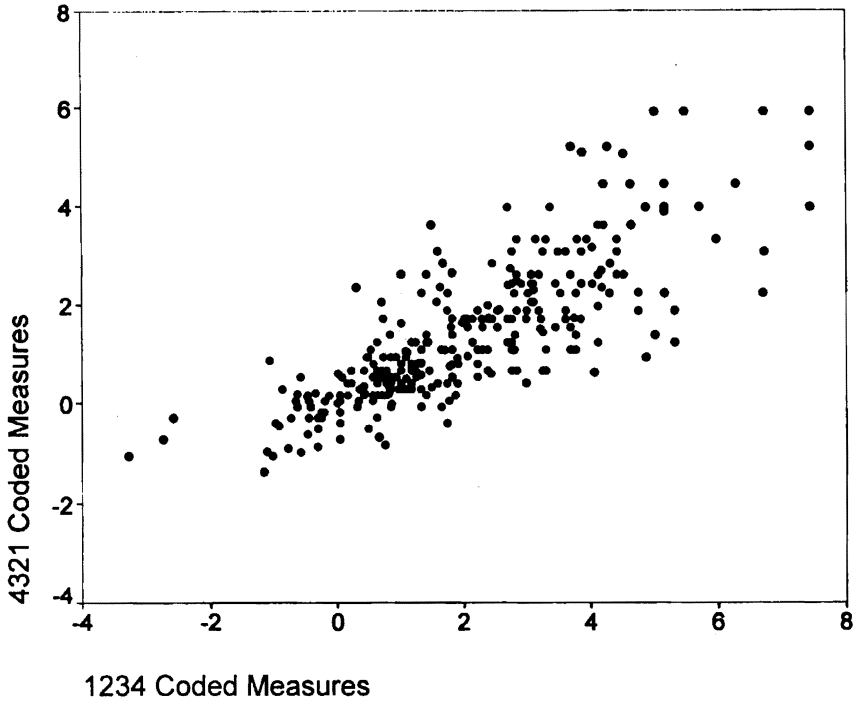

Figure 1. Scatter plot of positive and negative worded measures.

than practical, with a single score being more in keeping with the purpose of the instrument. However, the results illustrate the sensitivity of the Rasch unweighted item total fit statistic to mild multidimensionality.

Conclusions

There are a number of conclusions that can be drawn from this analysis. These conclusions serve as recommendations for researchers and educators developing instruments. First, it is important to start with a theoretical base for the design of an instrument. In this case, the use of the theoretical base reduced the number of misfitting items to 4 positively worded and 4 negatively worded items out of a total of 78 items—far fewer than is normally expected. Second, be careful when introducing reverse coded or negatively worded items into the instrument. Although this practice has been recommended as a means of offsetting response set biases, there are clear indications in a variety of settings that the responses to the negative worded items do not measure the same underlying construct as the positively worded items. There may be a substantial correlation between the two variables, as there was in this case, but the combination of the positively and negatively worded items in the same calibration often causes the item fit statistics to have an unexpectedly high proportion of misfitting items. Finally, in cases where there is an unexpected amount of misfit it is often prudent to combine Rasch analysis with principal component analysis to be able to detect mulitdimensionality in as wide a possible range of cases. See Wright (1996) and Smith (1996) for a more detailed discussion of this point.

References

Bainer, D. L. (1998). A comparison of four models of group efforts and their implications for establishing educational partnerships. *Journal of Research in Rural Education, 13*(3), 1-10.

Bainer, D. L. (1997). *With a new lens: How partnering impacts teachers' views of and approaches to teaching science.* Columbus, OH: ERIC Center for Science, Mathematics, and Environmental Education. (ERIC Document Reproduction Service No. 406 222).

Bainer, D. L., Barron, P., and Cantrell, D. (1995, April). *The impact of reform-based partnerships on attitudes toward environmental science and partnering and on classroom instruction.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Bainer, D. L., Barron, P., and Cantrell, P. (Winter 1996/97). Enhancing science instruction in rural elementary schools through partnering. *The Rural Educator, 18*(2), 16-21.

Bainer, D. L., Cantrell, D., and Barron, P. (1997, November). *Professional development of resource professionals through school-based partnerships.* Paper presented at the annual meeting of the North American Association for Environmental Education, Vancouver, B.C., Canada.

Bainer, D. L., Halon, C. S., and Williams, D. (1996, October). *Educating the professions through school-based partnerships.* Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago.

Banerji, M., Smith, R. M., and Dedrick. R. F. (1997). Dimensionality of an early childhood scale using Rasch analysis and confirmatory factor analysis. *Journal of Outcome Measurement, 1*(1), 56-85.

Cobb, C. and Quaglia, R. J. (1994, April). *Moving beyond school-business partnerships and creating relationships.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans LA.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method.* New York: Wiley and Sons.

Miron, L. F. and Wimpelberg, R. K. (1989). School/business partnerships and the reform of education. *Administrator's Notebook, XXXIII*(9), 1-4.

Rigden, D. W. (1991). *Business/school partnerships: A path to effective restructuring.* New York: Council for Aid to Education.

Rigden, D. W. (1992). *Business and the schools: A guide to effective programs.* New York: Council for Aid to Education.

Science and Mathematics Network of Central Ohio. (1996). *Excite the Mind: Evaluating the Impact of Network Partnerships.* Columbus, OH: Science and Mathematics Network of Central Ohio.

Smith, R.M. (1991a). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement, 51*, 541-565.

Smith, R.M. (1991b). *IPARM: Item and person analysis with the Rasch model.* Chicago: MESA Press.

Smith, R.M. (1992). *Applications of Rasch measurement.* Chicago: MESA Press.

Smith, R.M. and Miao, C.Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (ed.) *Objective measurement: Theory into practice II.* Norwood, N.J.: Ablex Publishing Corp.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling, 3*(1), 25-40.

Wichienwong, W. (1988). *The relationship of administrators' involvement in the evaluation process and evaluation attitudes.* Unpublished doctoral dissertation, The Ohio State University, Columbus, OH.

Wright, B. D. (1994). Composition Analysis. *Mid-Western Educational Researcher*, *7*(2), 29-36, 38.

Wright, B.D. (1996). Comparing Rasch measurement and factor anaylsis. *Structural Equation Modeling*, *3*(1), 3-24.

Wright, B. D. and Linacre, J. M. (1996). *BIGSTEPS: Rasch analysis for all two-facet models.* Chicago: MESA Press.

JOURNAL OF OUTCOME MEASUREMENT®, *3*(3), 266–277

Application of Rasch Measurement to a Measure of Musical Performance

Kathleen A. Haley

Boston College

This purpose of this paper is to describe the Rasch calibration of a portion of the Watkins-Farnum Performance Scale (WFPS), using a sample of 218 sixth graders from a middle school in Rhode Island. The WFPS is a test of instrumental music performance and consists of fourteen exercises of increasing difficulty, of which students play as many as possible until they fail two consecutive exercises. The WFPS has demonstrated reliability and validity. However, classical test theory did not allow its authors to calculate a measure of difficulty for each bar (because some students did not play all bars), or to allow judges the flexibility to shorten the scale (because of low reliability). Using Rasch scaling, item difficulties can be estimated, the test can be administered more efficiently, and perhaps most importantly, diagnostic information can be easily obtained.

Requests for reprints should be sent to Kathleen Haley, Boston College, Campion Hall 323, Chestnut Hill, Massachusetts 02467, e-mail: haleykc@bc.edu.

Introduction

In 1942, John G. Watkins completed his landmark dissertation, *Objective Measurement of Instrumental Performance*, in which he developed a highly valid and reliable scale for the measurement of cornet performance. It consists of fourteen short exercises of progressive difficulty. Each exercise consists of 16–36 bars. For purposes of this paper, each bar is considered an item. Each item is scored correct if there are no errors, and incorrect if one or more errors are made. Stephen E. Farnum later modified the scale to be useful for any instrument (Watkins and Farnum, 1954). The result was the Watkins-Farnum Performance Scale (WFPS), a scale which has been in frequent use by many instrumental music teachers ever since. However, there were two things that Watkins wanted to do that classical test theory did not provide the technology to do.

1) Since the scale is composed of exercises that become progressively more difficult, Watkins stopped his subjects once "a sheer chaos of sound in no manner resembling music was coming forth from the horn" (Watkins, 1942). This practice prevented him from being able to calculate difficulty values for each bar, since each item was not taken by the same number of students. Instead, he calculated the probability of playing a given exercise with a specified number of errors. Using item response theory, difficulties can be estimated even if all subjects in a sample do not take the item.

2) In an effort to save time, Watkins wrote a preliminary exercise which began quite simply, but became difficult quickly. The intent was to give test administrators an idea of the playing level of the subject, so that advanced players would not have to play the easiest exercises. The preliminary exercise was dropped because of low reliability. However, given that the shortest exercise consists of sixteen bars, which are to be treated as individual items, one or a few exercises will be sufficient to estimate ability under item response theory, as long as the student does not play them perfectly or completely fail them. Thus the scale will have the built-in time saving feature Watkins had hoped to create.

Overview of the Literature

Most of the research emphasis on measurement in music has been in the cognitive and affective domains. A great deal of research is available regarding tests of musical aptitude, including the Seashore (1919) and the Gordon (1965) tests. A sampling of some of the most com-

monly used tests of music achievement (in this sense meaning a test of a student's music literacy and appreciation) includes the Kwalwasser-Ruch Test of Musical Accomplishment (1952), the Iowa Test of Musical Literacy (Gordon, 1970), and the Silver Burdett Music Competency Tests (Colwell, 1979). However, Colwell (1982, personal communication 1998) points out that the Watkins-Farnum Performance Test remains the only published test of instrumental musical performance. Madsen and Madsen (1970, p. 40) state that

> "It would seem that the music researcher should also be concerned about the lack of valid and reliable live performance tests. For some time it has been evident that paper and pencil studies provide only partial answers to performing variables...Perhaps after many years of basic experimentation in all areas of music, adequate performance tests can be constructed."

The emphases on the cognitive and affective domains are not surprising, however, given the emphasis on these domains in the schools (Boyle, 1992) and their relative ease of measurement when compared to the psychomotor domain. However, as Watkins (1942) points out, this is putting the proverbial cart before the horse, considering that validation studies of aptitude tests generally involve a performance measure as the dependent measure. This was Watkins' professed motivation for the construction of his measure of cornet performance.

A common type of performance measure consists of a band director listening to each student play a passage, and subjectively giving the student a score or ranking, with or without the use of predetermined categories and scoring criteria. It should not be surprising that such scoring is quite unreliable (Fiske, 1977a). Much of the research in musical performance assessment has focused on finding factors that affect the reliability of this type of assessment. However, the research has generally shown that many factors commonly assumed to improve reliability actually have little or no effect. Among the many factors which have been considered in relation to reliability are judge specialization on the instrument in question (Fiske, 1977b), judges' performing ability (Wapnick and Rosenquist, 1991, Fiske, 1977a) and scores on music aptitude tests (Mullin, 1979), the use of analytic scoring (Wapnick, et al., 1993), and use of the musical score by judges (Wapnick, et al.). Music educators are always searching for more consistent methods of evaluating their students, and Watkins was among the first to provide such a method. When Watkins published,

there were few standardized performance tests of any kind. A few measures of sight-singing were available, as well as one measure of organ performance (Watkins, 1942). Watkins was the first to widely publish an attempt at objective instrumental performance.

In the 1970s, rating scales, consisting of a series of items descriptive of a performance, began to be considered seriously in the music literature. The groundbreaking rating scale work was by Abeles (1973), who used a qualitative analysis followed by a factor analysis to create a Clarinet Performance Rating Scale. This is a thirty-item Likert scale in which judges are presented with statements such as "The attacks and releases were clean.", and asked to "highly agree," "agree," "neither agree nor disagree," "disagree," or "highly disagree." Unfortunately, despite highly desirable psychometric properties, a true rating scales are not commonly used in the classroom, perhaps because the study has not been replicated for all instruments.

Research Methods

Sample

The sample consists of 218 students from a Rhode Island middle school. About 62% of the students are girls. All the participants are in the sixth grade, and each is a student of some wind band instrument. About three-quarters of the students played some woodwind instrument, primarily flute (30%), clarinet (22%), and saxophone (18%). The remaining students were brass players, primarily trumpet (14%). The students were administered the test at the end of their second year of instrumental instruction.

Instrument

The Watkins-Farnum Performance Scale is a set of 14 exercises of increasing difficulty, ranging from very simple to quite difficult. Each exercise is sixteen to 36 bars long. Examinees play the exercises in order, and each bar is scored either correct or incorrect. There are a number of types of errors and very specific criteria for what constitutes an error. Watkins was attempting to create an instrument that could be scored completely objectively. Therefore, wrong notes, rhythms and articulations are considered errors, as is failure to observe a dynamic marking. Poor tone quality or intonation is not considered an error. The test is appropri-

ate for a wide range of ability, but because of the ability level of these students, only the first eight exercise are calibrated in this paper. This results in a total of 132 items.

Under Watkins' and Farnum's scoring method, only one error is scored in each bar. Therefore, the possible scores for each bar are one and zero only. The maximum possible score on each exercise is a given standard (ten for most exercises), and the total points scored equals the standard for the exercise minus the number of bars containing an error. A student is to be allowed to continue until he/she scores zero on two consecutive exercises. The total score for the test is the sum of the individual exercise scores.

Watkins (1942) reports a parallel forms reliability coefficient of .95 for the Cornet Performance Scale. Farnum (1954) reports parallel forms reliabilities ranging from .87 to .94. Farnum reports concurrent validity for his scale as the rank order correlation between the instructor's ranking of the students and the students' WFPS score. The coefficients range from .77 to .87 for different instrument groups. These coefficients show quite acceptable levels of reliability and validity.

Procedure

Students were asked to sight read the exercises at the end of their 2nd year of instrumental study. The students took the test as a part of the normal requirements of the music class. The scoring was completed by their band director.

The Rasch dichotomous model was used to analyze the data. Quest Interactive Test Analysis System (Adams and Khoo, 1993) was used to obtain a difficulty estimate for each item, and an ability estimate for each student. An item fit analysis was also performed. Those items with fit statistics greater than t=2.0 were closely examined, and possible explanations for the misfit were discussed.

Results

The resulting variable map, shown in Figure 1 generally conformed to Watkin's theory about the relative difficulty of the exercises. Higher numbered items belong to more difficult exercises and are generally located higher in the scale than the lower numbered items. Items are numbered on the right side, with the more difficult items toward the top, while students are designated as X's on the left side, with the more able students

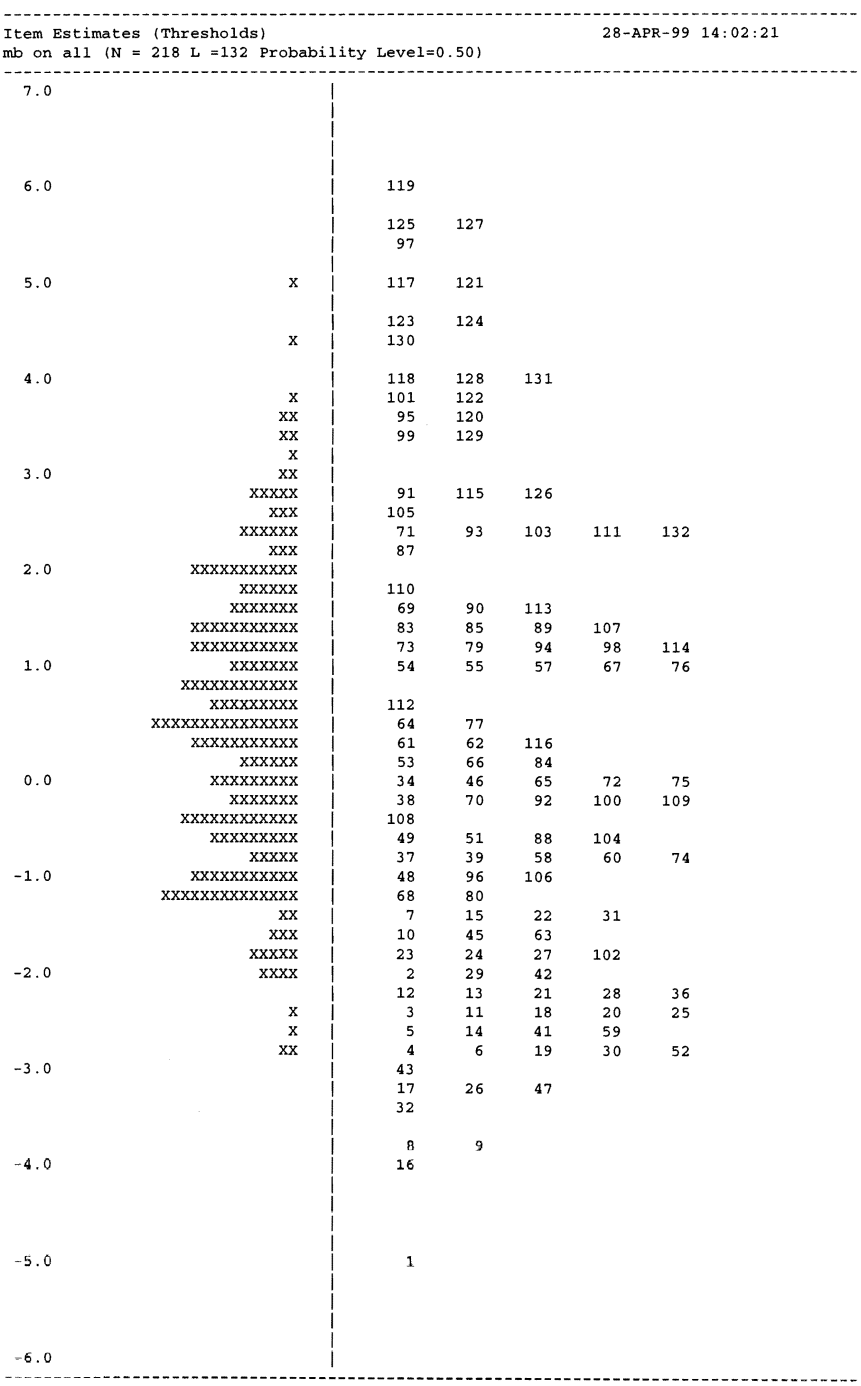

Figure 1. Variable Map of WFPS Items

toward the top. Only the first eight exercises were calibrated because this set of items is particularly well matched to the ability of the sample. There are items appropriate for the entire range of student ability, with no gaps in the difficulty continuum.

It was expected that items within an exercise would cluster together, but that there should be no particular order within an exercise. That is, the highest items on the scale should generally be from exercise 8, in no particular order. Those from exercise 7 should generally be below these. There are many exceptions to the general trend of items belonging to difficult exercises being difficult themselves. This is not surprising since within a difficult exercise one would not be surprised to find isolated bars which low-performing students are able to play correctly. Figure 2 shows the difficulty of items by exercise.

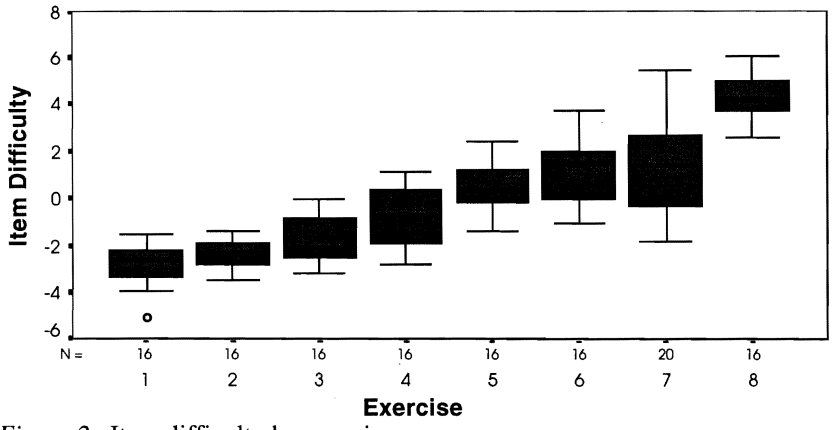

Figure 2. Item difficulty by exercise

Note that the items from exercise 1 and exercise 2 are perhaps more similar in difficulty than any other pair of adjacent exercises. This is due to the fact that almost all students were able to play most of these bars, and therefore there is little information to distinguish the difficulty of exercise 2 from that of exercise 1. If more very low performing students were added, some would be expected to correctly play bars from exercise 1 but not those from exercise 2, thus separating the items.

The data in general were quite a good fit to the model. Two items showed a substantial amount of misfit, meaning that higher-scoring students are not necessarily playing the bar correctly at a higher rate than low-

scoring students. One of the items, item 64 (t=4.7), was an easy bar with a repeat sign in it. If the student missed the repeat, the bar was marked wrong. The misfit says that a high-scoring student is no more likely than a low-scoring student to correctly play the repeat. The other item, item 54 (t=3.9), was the first bar that contained a note altered by the key signature (B-flat). Since subsequent bars containing a B-flat did not show misfit, a reasonable interpretation might be that high-scoring students are no more likely than low-scoring students to correctly play the key signature at the first opportunity. However, they do seem to be more likely to correct themselves and play it correctly on subsequent opportunities.

Discussion

When Watkins created the scale, he intentionally wrote exercises of increasing difficulty. He verified this judgment by asking experts to rate the difficulty of the exercises. In pilot testing the instrument, he was further able to test this assumption. However, he was only able to test this at the exercise level, because different numbers of students took each item. With Rasch scaling, difficulty values can now be estimated for each item. This is important because, as mentioned, the difficulties of the exercises are rough; many easier bars are found within difficult exercises, and vice versa. However, the two most important contributions of the Rasch scaling are described in the following paragraphs.

The ordinary method of administering the WFPS is to have the student begin at exercise one, and perform each exercise in turn. The scorer listens, marks errors, subtracts the number of errors in each exercise from a maximum score for the exercise, and totals the exercise scores. While a teacher may know that a student plays at a level far above the first exercises, s/he must still listen to the early exercises to reach a score. Given the large number of students in many band programs, this is probably not the best use of a teacher's time. It would be possible, using the Rasch calibration of the WFPS, to create a quasi-adaptive version. The test must still be scored by a human judge. It would still be possible to create a fully adaptive computer version of the WFPS, but it would require the response strings to be entered by the test administrator after each exercise. Each exercise would be considered a "testlet," a small series of items considered as a unit for the purposes of choosing the next item or testlet. This seems somewhat too cumbersome for the purposes of a classroom music teacher. Instead, the selection of exercises would follow an algorithm that depended in part on human judgment. The computer would be used only for scoring after completion of the test.

The second major advantage of the Rasch scaling is the quality of information gained. Under classical test theory, which underlies the WFPS, the information gained from the administration of a test is a single number reflecting the number of correct responses. The Rasch model provides such information as well. However, since the items are given difficulty scores on the same scale as the student ability scores, a teacher can determine what types of bars a student is and is not likely to perform correctly.

The variable map shown earlier is repeated as Figure 3 with example items included.

This illustrates what is perhaps the greatest advantage provided by the Rasch model. Suppose a new student entered a band program and scored a -2 on the WFPS. The director would know that she would have a 50% chance of correctly playing the third example item and would be more likely than not to play examples one and two correctly and the others incorrectly. Similarly, if a new student scored a 3, she would be likely to play six and seven incorrectly, but all the others correctly. A chart could be created from which students could easily read what they need to practice most to achieve the next level.

One issue that may reasonably be troubling the reader is the idea that this scale is objective, measuring only the presence or absence of errors, while the quality of a musical performance is a subjective decision. Indeed, this is a well-founded concern and one which receives little treatment in Watkins' dissertation. However, the subjective qualities of musical performance, often specified as "expression," "musicality," or similar labels, are highly correlated with global ratings and with the more objective qualities measured in the WFPS (Fiske, 1975; Burnsed, Hinkle, and King, 1985). Nevertheless, it is not recommended that a music teacher use the WFPS, either with or without the Rasch item response scaling that is applied in this study, as the sole basis for decisions that affect the future of an individual student. That is, this test should not, except in part, determine a student's grade or be used to make a decision whether or not to admit a student to a performing ensemble. Some appropriate uses would be:

1) A first hearing of new students, to allow a director to become better acquainted with a student and his or her ability.

2) An initial seating[1] of students within a section, given that there will be future opportunity for students to move up within the section.

3) A monitor of progress in specific areas, such as rhythm or articulation, or in sight reading.

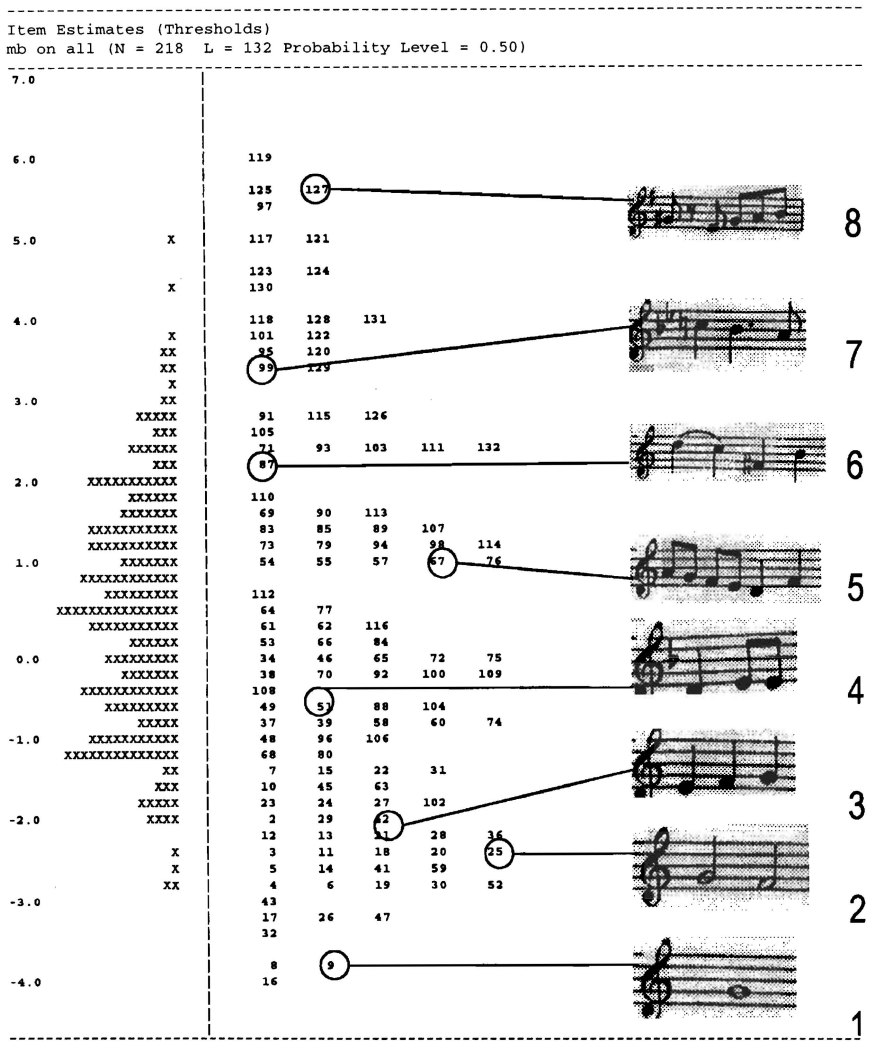

Figure 3. Variable Map of WFPS Items with Example Items Added

4) A valid measure of musical performance to be used to report group scores in future research studies.

Again, there are many difficulties in assessing musical performance validly and reliably. The WFPS does not solve, or even address, them all. But the instrument has been shown to have sound psychometric properties and is widely respected. The addition of Rasch scaling improves the efficiency of the administration and makes diagnostic information simple

to obtain. More research is of course necessary to calibrate the entire scale. But it appears from this initial foray into Rasch scaling that the Rasch model may prove to be a promising new measurement tool for both music teachers and researchers.

Footnote

[1] Many band directors audition students not only for admission into an ensemble but also for placement. For example, the best clarinetist will be "first-chair clarinet" and be seated accordingly.

Acknowledgment

The editorial comments of Dr. Larry Ludlow are gratefully acknowledged.

References

Abeles, H. F. (1973). A facet-factorial approach to the construction of rating scales to measure complex behaviors. *Journal of Educational Measurement, 10*(2), 145-51.

Adams, R., and Khoo, S. (1993). *Quest: The Interactive Test Analysis System.* Victoria: Australian Council for Educational Research.

Burnsed, V., Hinkle, D., and King, S. (1985). Performance evaluation reliability at selected concert festivals. *Journal of Band Research, 21*(1), 22-29.

Boyle, J. D. (1992). Evaluation of music ability. In R. Colwell (Ed.). *Handbook of research on music teaching and learning* (pp. 247-65). New York: Schirmer Books.

Colwell, R. (1979). *Silver Burdett Music Competency Tests.* Morristown, NJ: Silver Burdett.

Colwell, R. (1982). Evaluation in Music Education: Perspicacious or Peregrine. In Colwell, R. (Ed.), *Symposium in Music Education.* Urbana: University of Illinois.

Fiske, H. E. (1975). Judge-group differences in the rating of high school trumpet performances. *Journal of Research in Music Education, 23*(3), 186-96.

Fiske, H. E. (1977a). The relationship of selected factors in adjudication reliability, *Journal of Research in Music Education, 25*(4), 256-63.

Fiske, H. E. (1977b). Who's to judge: New insights into performance evaluation. *Music Educators Journal, 64*(4), 23-25.

Gordon, E. (1965). *The musical aptitude profile.* Boston: Houghton Mifflin.

Gordon, E. (1970). *Iowa Test of Musical Literacy.* Iowa City: The University of Iowa, Bureau of Educational Research and Service.

Kwalwasser, J. and Ruch, G. M. (1952). Kwalwasser-Ruch Test of Musical Accomplishment for Grades Four through Twelve, Bureau of Educational Research and Service, State University of Iowa.

Madsen, C. K. and Madsen, C. H., Jr. (1970). *Experimental Research in Music*, from the Contemporary Perspectives in Music Education Series, Charles Leonhard (ed.). Englewood Cliffs, New Jersey: Prentice-Hall.

Mullin, A. (1979). Melodic/rhythmic decision-making by senior secondary music students. Unpublished master's thesis, University of Western Ontario.

Seashore, C. E. (1919). *Seashore Measures of Musical Talent*. New York Columbia Phonograph Co.

Wapnick, J., and Rosenquist, M. (1991). Preferences of undergraduate music majors for sequenced versus performed piano music. *Journal of Research in Music Education, 39*, 152-160.

Wapnick, J., Flowers, P., Alegant, M., and Jasinskas, L. (1993). Consistency in Piano Performance Evaluation, *Journal of Research in Music Education, 41*(4), 282-92.

Watkins, J. G. (1942). *Objective measurement of instrumental performance*, New York: Teachers' College Bureau of Publications, Columbia University, 1942.

Watkins, J.G. and Farnum, S.E. (1954). *The Watkins-Farnum Performance Scale: Form A*. Winona, Minn.: Hal Leonard, 1954.

JOURNAL OF OUTCOME MEASUREMENT®, *3*(3), 278–294

The Flow Experience: A Rasch Analysis of Jackson's Flow State Scale

Gershon Tenenbaum
Gerard J. Fogarty
University of Southern Queensland

Susan A. Jackson
Queensland University of Technology

Jackson and Marsh (1996) reported the development of a Flow State Scale (FSS) for use in sport and physical activity. The FSS contains 36 items measuring the nine dimensions of flow described by Csikszentmihalyi (1990, 1993). Jackson and Marsh reported high internal consistency estimates for the subscales and evidence for nine first-order factors and one second-order factor when confirmatory factor analytic techniques were used. The present study extended this validation work by subjecting the data from the original sample (N = 394) of elite younger athletes and a subsequent sample (N = 398) of older athletes to Rasch analysis. These Rasch analyses showed quite clearly that the flow dimensions may be conceptualised as a continuum with "autotelic experience" being experienced more readily than dimensions such as "transformation of time", a state that may only be encountered at the height of a deep flow experience. The Rasch analyses provide useful additional information about the areas of the flow continuum tapped by the items and scales of the FSS and, in so doing, help to confirm the construct validity and generalisability of the scale itself.

Requests for reprints should be sent to Gershon Tenenbaum, Department of Psychology, University of Southern Queensland, Toowoomba, Queensland, 4350, Australia, e-mail: tenenbau@usq.edu.au.

Stages of the Flow Experience: a Rasch Analysis of Jackson's Flow State Scale

Flow is an optimal psychological state that has been described at length by Csikszentmihalyi (1990, 1993) and adapted to sport and physical activity settings by sport and exercise psychology researchers interested in identifying and understanding the nature of the experience in these environments. Understanding the experience of the state of flow in sport settings has been the focus of Jackson's (e.g., Jackson, 1995, 1996; Jackson and Marsh, 1996) research in this area. The present study is an attempt to further describe and explain the process of flow as it may occur in physical activity settings.

Flow has been described by Csikszentmihalyi (1990, 1993) as comprising the following nine dimensions:

1. *Challenge-Skill balance.* In flow, there is a feeling of balance between the demands of the situation and personal skills.
2. *Action-Awareness Merging.* Involvement is so deep that there is a feeling of automaticity about one's actions.
3. *Clear Goals.* A feeling of certainty about what one is going to do.
4. *Unambiguous Feedback.* Immediate and clear feedback is received, confirming feelings that everything is going according to plan.
5. *Concentration on Task at Hand.* A feeling of being really focussed.
6. *Sense of Control.* The distinguishing charactersitic of this feeling in the flow state is that it happens without conscious effort.
7. *Loss of Self-Consciousness.* Concern for the self disappears as the person becomes one with the activity.
8. *Transformation of Time.* Time can be seen as passing more quickly, more slowly, or there may be a complete lack of awareness of the passing of time.
9. *Autotelic Experience.* Csikszentmihalyi (1990) describes this as the end result of being in flow, a feeling of doing something for its own sake, with no expectation of future reward or benefit.

Jackson (1996) found support for these dimensions in a qualitative analysis of elite athletes' flow descriptions. Jackson and Marsh (1996) further argued that a multimethod approach is needed to understand flow, incorporating both qualitative and quantitative research. In particular, they

urged the importance of establishing the validity of the various constructs said to underlie the flow experience and relating these dimensions to other psychological states. Toward this end, they developed the 36-item Flow State Scale (FSS) and used conventional item analysis techniques along with confirmatory factor analysis to establish the construct validity of the scale. The main findings of importance to the present study were that nine first-order factors were needed to give a good account of the variance and that a single second-order factor could also be extracted to help explain the correlations among the nine factors. Thus, the FSS provided support for Csikszentmihalyi's nine correlated factors model of the flow experience as well as a second order "global flow" factor that can be measured by these nine first-order factors.

Although the structure of the FSS was as expected, Jackson and Marsh (1996) reported one or two interesting features of the factor pattern. The first feature was the low communalities of the Transformation of Time and Loss of Self-Consciousness factors, suggesting that these factors are of less overall importance than the other factors. They noted that there is support for this point of view elsewhere (Jackson, 1996). The second feature was unexpected and had to do with the moderate loading of the Autotelic Experience factor on the second-order factor. This factor is described by Csikszentmihalyi (1990) as central to the flow experience. Jackson and Marsh (1996) concluded that its low loading on the second-order factor ruled out the possibility of its having a central role and that perhaps enjoyment of sport is taken for granted among athletes. Other possible explanations were also forwarded. Interestingly, in a qualitative study of the flow dimensions in athletes (Jackson, 1996), autotelic experience was found to be the most salient, or frequently experienced, dimension.

The question of the relative importance of factors in the global flow continuum is a difficult one to answer. However, there are techniques that are ideally suited to such questions. One such technique is known as Rasch analysis. Originally developed on measures of cognitive ability (Rasch, 1960), Rasch analysis allows test administrators to gauge the extent to which each item in a test taps an underlying latent trait. It does so by modelling response patterns for the full set of items and testing the goodness of fit for the item pattern as a whole and also the fit for each individual item. The details of this mathematical process will not be described here, instead the reader is referred to introductions by some of the main

adherents of Rasch analysis (e.g., Wright and Stone, 1979). What follows is a very basic introduction to some of the main features of Rasch analysis, intended mainly to show that it can be applied to the FSS.

When the Rasch Model is applied to response data, the hierarchy of the items (i.e., item calibrations on the linear continuum) is tested for consistency across the sample of the respondents. Furthermore, the ordering of item values can be compared for defined groups of persons, abilities, attitudes and other characteristics, to ultimately define the variable. The assumption underlying the model is that the items share a single dimension or trait and that the people are relatively homogeneous with respect to the measured variable. The presence of a strong second order factor in the FSS suggests that it is amenable to Rasch analysis.

When applied to attitudinal data, the model assumes that response X_{vi}, which occurs when person v takes item i, is governed by the person's attitude β_v and the item's value δ_i and nothing else. Since both β_v and δ_i share a common linear continuum, X_{vi} is a function of their difference ($\beta_v - \delta_i$). X_{vi} is probabilistic in nature and so is the mathematical procedure associated with the model. The logistic function provides a model which makes both linearity of scale and generality of measure possible. In addition, "fit statistics" indicate how much a person deviates in responding to single items with respect to his/her expected ratings (based on his/her total raw score), and how each single item is rated in relation to its affective/attitudinal value. Further details of this method can be read in Wright and Stone (1979) and Wright and Masters (1982).

Method

Participants

The first sample was that used by Jackson and Marsh (1996) and is described fully in that paper. Briefly, the sample consisted of 394 athletes from a variety (41) of sports. The sample comprised 264 males and 130 females. A range of participation levels were represented, from recreational to national representatives. The second sample consisted of 398 participants in the 1994 World Masters Games competition. The athletes, 243 males and 155 females, were primarily from Australia (84%), although 13 nationalities were represented in the sample. Four sports made up the sample: track and field, triathlon, swimming, and cycling. The mean age of the participants was 46 (SD = 10.9), and as a group they had participated in their sport from 14-19 years.

Instrument

The Flow State Scale (FSS: Jackson and Marsh, 1996) measures flow in sport and physical activity settings. It consists of 36 items divided into 9 subscales, each representing a different dimension. The 36 items can also contribute to a broad second-order general flow factor. The items were derived from research on the flow state within and outside sport settings and qualitative analysis of interviews with elite athletes. Participants are asked to recall one specific experience that occurred whilst participating in sport or physical activity that constituted an optimal experience. They then respond to the flow items using a 5-point Likert response format where 1 indicates "Strongly disagree" and 5 indicates "Strongly agree". The alpha internal consistency estimates of all subscales ranges from 0.79 to 0.86. Confirmatory factor analytic research has supported a model with nine first-order and one higher-order factor (Jackson and Marsh, 1996). Abbreviated descriptions of the items can be found in Table 1.

Statistical Analysis

The Ascore package (Andrich, Sheridan, and Lyne, 1991) was used to estimate the locations and fit statistics for the 36 FSS items. The package can handle various response types, including rating scales, and returns item locations with standard errors, threshold estimates for item categories, and fit statistics for each item and person in accordance with Wright and Masters (1982).

Results

The Rasch analysis was applied to the "flow" scale ratings of 328 participants from the first sample who had "valid" responses. Sixty-two participants were excluded from the final analysis because there was evidence they had responded in an invalid manner. These invalid responses were mostly cases where extreme responses were given to all items (all 1s or 5s), producing fit values > |3.00|. The item values (calibrations) in logits, standard errors (SE), fit statistics and item-trait interaction test of fit values are presented in table 1.

The "easiest" item to score high (i.e., the majority of respondents rated it highly), with a location estimate of -1.16 logits, was item 36 ("I found the experience extremely rewarding"). This feeling may thus be experienced very early in the flow experience. The "hardest" item to rate highly (i.e., the majority of respondents rated it low), with a location estimate of 1.24 logits,

Table 1

The FSS dimensions, abbreviated description, items' calibration and fit

Dimension	Item	Abbreviated description	Calib. (Logits)	SE	Fit Value
Autotelic Experience	36	experience was extremely rewarding	-1.16	.09	-2.3
(ENJY)	8	enjoyed experience	-0.96	.10	-0.5
	15	wanted to recapture the feeling	-0.83	.08	-1.2
	18	experience left me feeling great	-0.71	.07	0.3
Clear Goals	2	knew what I wanted to do	-0.64	.08	1.3
(GOAL)	21	strong sense of what I wanted to do	-0.54	.08	-1.6
	28	knew what I wanted to achieve	-0.37	.08	-0.3
	32	clearly defined goals	-0.04	.07	-2.6
Challenge-Skill Balance	16	competent to meet demands	-0.45	.08	-1.2
(CHAL)	9	abilities matched challenge	-0.16	.07	-0.2
	19	challenge and skills equally high	0.06	.07	-0.3
	1	skills met challenge	0.02	.07	1.0
Concentration on Task	17	total concentration	-0.44	.07	-1.9
(CONC)	23	completely focused on task	-0.27	.07	-1.3
	4	attention focussed	-0.09	.07	-0.2
	11	kept my mind on what was happening	0.28	.06	1.8
Paradox of Control	24	in total control of body	-0.32	.07	0.9
(CONT)	12	I could control what I was doing	-0.16	.07	-0.3
	30	feeling of total control	0.06	.07	-2.8
	5	in total control	0.18	.07	-1.5
Unambiguous Feedback	33	knew how well I was doing by the way I was performing	-0.36	.07	-0.0
(FDBK)	3	clearly doing well	-0.09	.07	0.1
	29	knew how well I was doing while performing	0.16	.07	0.6
	22	aware of how well I was performing	0.20	.07	1.1
Action-Awareness Merging	20	things happened automatically	0.10	.07	-0.4
(ACT)	31	spontaneous and automatic	0.14	.07	-0.1
	27	performed automatically	0.19	.07	0.2
	10	correct movements without thinking	0.23	.07	1.6
Transformation of Time	14	time different from normal	0.30	.06	3.5
(TRAN)	7	altered time	0.49	.06	3.0
	35	slow motion	0.97	.05	3.3
	26	time stopped	1.13	.05	5.2
Loss of Self-Consciousness	6	not concerned with others	0.46	.06	2.8
(LOSS)	34	not worried about others	0.50	.06	0.7
	25	not concerned with presentation	0.87	.05	4.1
	13	not worried about performance	1.24	.05	6.6

was item 13 ("I was not worried about my performance during the event"). This feeling comes only when the individual is very much into the flow experience. The rest of the items (i.e., experiences) are spread along the linear "flow" continuum between these two items.

Table 1 also shows that five of the items were possible "misfits". They were items 13 ("I was not worried about my performance during the event"), item 14 ("The way time passed seemed to be different from normal"), item 25 ("I was not concerned with how I was presenting myself"), item 26 ("I felt like time stopped while I was performing"), and item 35 ("At times, it almost seemed like things were happening in slow motion"). From a psychometric point of view, misfit items might be deleted from the final version of the measurement scale. However, re-examination of the five misfit items revealed that four of the five items (25, 35, 26, and 13) were located at the upper end of the linear continuum, i.e., they were experienced only when persons were very much in the "flow" experience. Also the fifth item (14) was located in this region. It is believed that such experiences may elicit similar responses from persons in different flow modes and, as a consequence, result in exaggerated standardised residuals and large misfit values. Also, deleting these items would result in shortening the length of the linear scale from 2.4 logits (-1.16 to +1.24) to 1.66 logits (-1.16 to +0.50). This would drastically limit the discrimination of persons who differ substantially in the "flow" experience. Finally, the internal consistency (Cronbach's Alpha) of the scale was 0.93 and the biserial correlations of these items with the total "flow" score were in the satisfactory range of 0.30 - 0.40. It was decided, therefore, that there were no firm grounds for altering the scale in any way.

Rasch analysis was also applied separately to data generated by elite athletes (international and national calibre) and adult recreational athletes (non-elite) athletes. The item values (calibrations) with respect to gender and athletic calibre were contrasted to each other through an X-Y plot. The two contrasts are presented in Figures 1 and 2.

The analyses and plots clearly indicate that item values (calibrations) remain relatively consistent across both gender and athletic calibre dimensions. However, there were some items that did not fall near to the line in both figures and these are worth mentioning. Items that fall appreciably above the identity line drawn through Figures 1 and 2 are "easier" for female and non-elite athletes respectively. Items that fall appreciably below the line are "easier" for male and elite athletes respectively. From Figure 1 it can be seen that items 36, 8, 18, 20, 31, and 28 all fell above the identity line. The first three of these items help to define the Autotelic Experience factor, the next two help to define Action-Awareness Merging factor. It appears that females find it easier to experience these states than

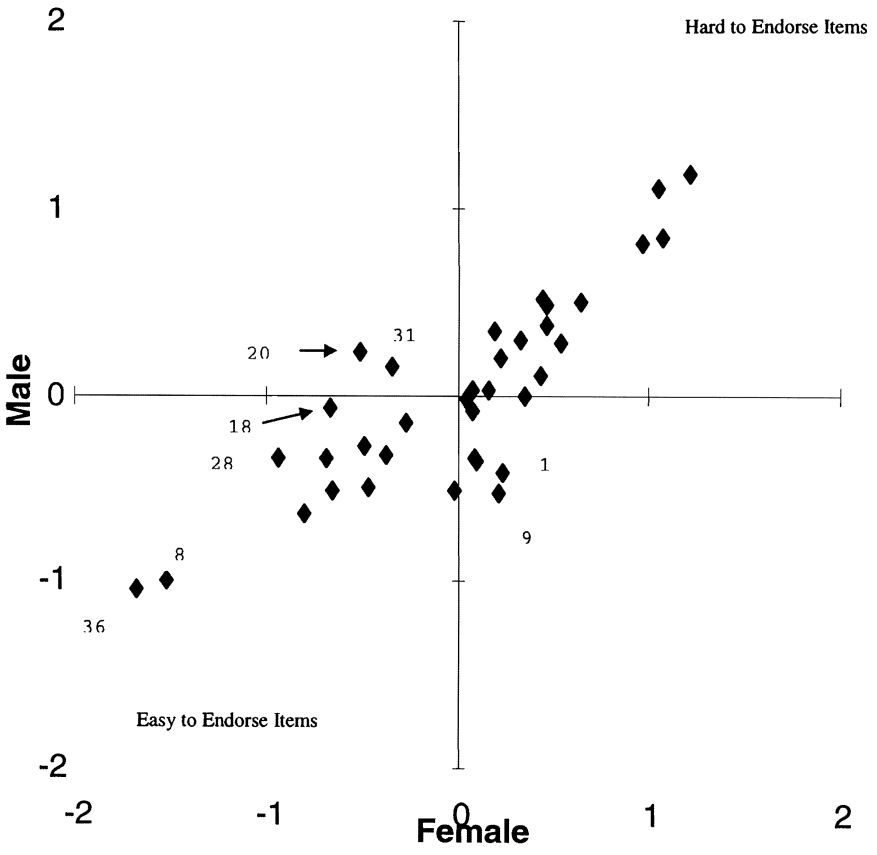

Figure 1. Comparison of Item Locations for Males and Females

do males. On the bottom side of the identity line, the two most distant items were 1 and 9, both measuring the Challenge-Skill balance factor, suggesting that males find it easier to endorse competency type items than do females. Independent t-tests revealed significant ($p < .01$) differences between males and females on these items. From Figure 2, which compares elite and non-elite athletes, a small number of items were some distance from the identity line (they are marked in the figure) but they did not as a group represent any of the FSS factors. Significant differences between the two samples of athletes were evidenced only for item 29 (Knowing how well one does while performing) and item 10 (Correct movement without thinking).

By and large, most items from Figures 1 and 2 fell close to the identity line indicating that males and females, as well as high level athletes and recreational athletes, perceive the experiences (i.e., items) of flow in

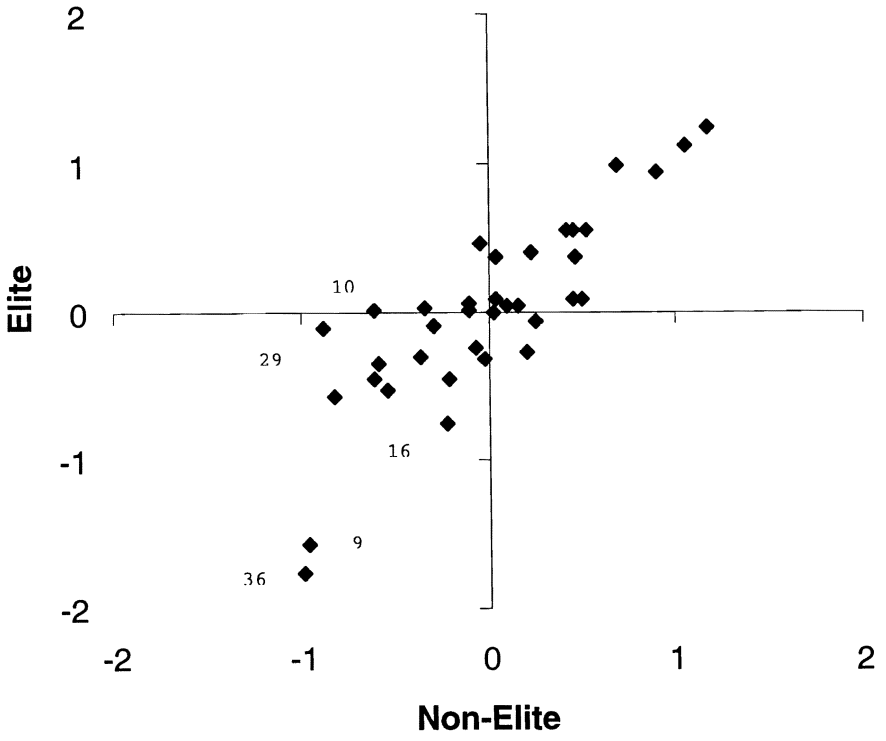

Figure 2. Comparison of Item Locations for Elite and Non-Elite Athletes

the same linear fashion. Jackson and Marsh's (1996) finding of differential factor loadings among the flow dimensions can be interpreted by the Rasch model as indicating that some dimensions of flow are more extensively experienced than others, or even that they occur at different points in time in the flow experience. Grouping the items under their factors shows what sections of the linear continuum are covered by each of the factors. To facilitate the interpretation of such a display, items are shown with their respective factors in Table 1.

If factors do represent underlying aspects of flow, then it is reasonable to expect that some of these aspects will be experienced before others. That is, athletes will find it easier to rate them more highly. If that is the case, there will be systematic differences in the location values for items marking different factors; the items for some factors will be at the "easy" end of the continuum, the items for others will be at the opposite end, others will be in the middle. Andrich (1975) gives an illustration of

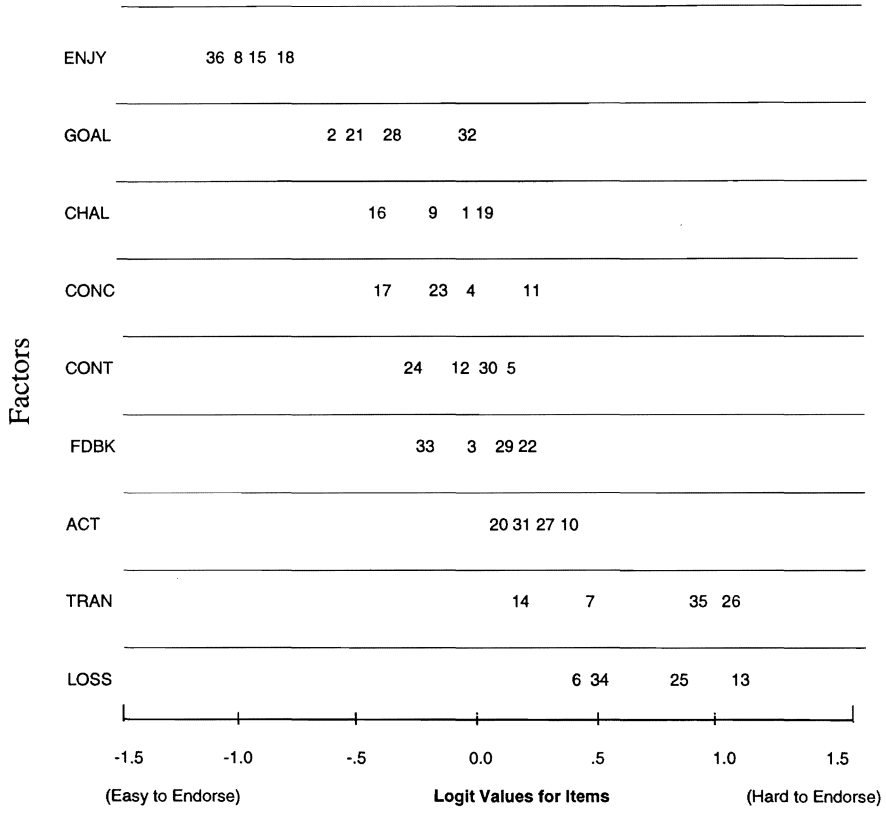

Figure 3. Location Estimates for Nine Dimensions of Flow Scale—First Sample

this technique. Figure 3 shows the location of the items grouped under their factorial headings for the full sample.

When grouped in this way, it is apparent that there is an ordering among the factors of the FSS (see Figure 3). The ninth factor, "autotelic experience", would be felt by most people, an indication that it is very easy to capture the feelings of enjoyment. This would be followed by the third dimension, "clear goals". Similarly, some other experiences would be felt by most people, such as competency in meeting demands (item 16), total concentration (item 17), total control (item 24), knowledge of what the person is doing (item 33), and focus on task (item 23). At the deeper levels of flow, the seventh dimension ("Loss of self-consciousness") and the eighth dimension ("Transformation of time") would be experienced. In other words, one should be in a "deep flow" to experience these feelings.

Comparison of Sample 1 with Sample 2

A somewhat abbreviated analysis can be presented here, the main purpose being to cross-validate the findings reported with the first sample. If the items are good representatives of the flow experience, their values are expected to remain stable because, as pointed out above, one of the properties of the Rasch analysis is that item estimates are sample free, unless the item has properties of instability. Demonstration of this property in the second data set would argue strongly for the stability of the FSS and provide further confirmation of the sections of the global flow continuum covered by this scale.

The item locations were compared across the two samples used in this study. This comparison is important because the athletes were quite different in many important respects. The first sample (N = 394) comprised mostly younger (mean age = 22) athletes who were either at the peak of their sporting prowess or rapidly approaching this point. The second group (N = 398) comprised mostly older athletes (mean age = 46), not necessarily elite, who were more likely to be engaged in sport simply because they liked to do so. If the item locations are the same across these diverse samples, it provides a substantial basis for interpretation of aspects of the flow experience.

When the data for the second sample were analysed, it was apparent that most of the estimates shared similar sign though somewhat different magnitudes. These similarlities and differences are shown in the scatterplot depicted in Figure 4. Dealing first with the points that did not fall on the identity line, it can be seen that items 8, 15, 18, and 36 were all endorsed more readily ($p < .01$) by the younger than the older athletes. These items all define the Autotelic Experience factor, suggesting that when they are in the flow experience younger athletes are enjoying their sport more. Of the six items that fell appreciably below the identity line, four of them (2, 21, 28, and 32) define the Clear Goals factor, suggesting that older athletes were more conscious of knowing what they wanted to do and what they wanted to achieve. The differences between the groups on these items were also significant ($p < .01$), though young/competitive and older/masters athletes rated the Clear Goals factor higher than most of the other flow dimensions.

Additional flow items where group differences were found included items 13, 25, 20, 31, 14, 22, and 11. Items 13 and 25 were from the Loss of Self-Consciousness dimension, where younger athletes endorsed these

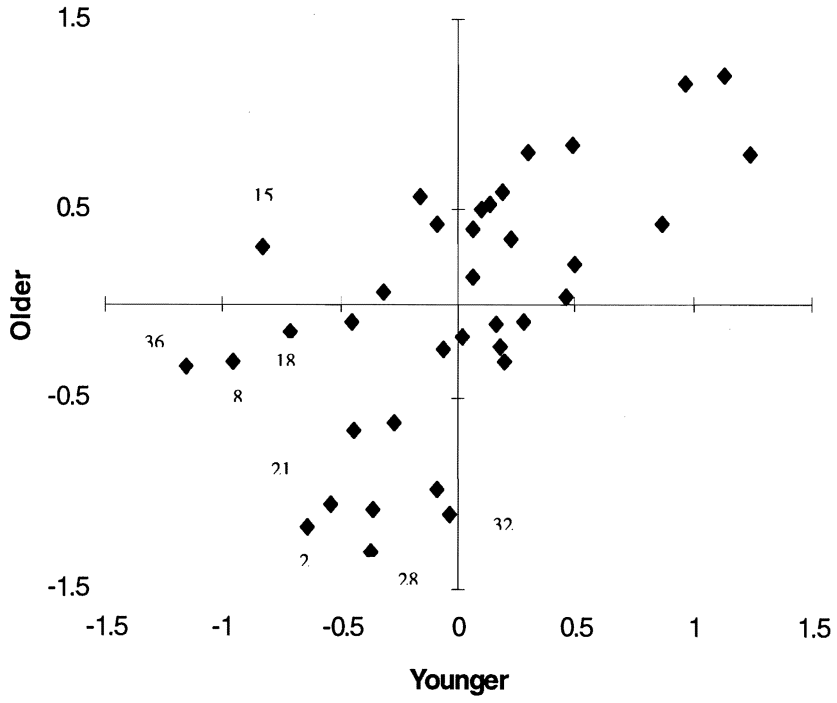

Figure 4. Comparison of Logit Values for Younger and Older Athletes

items less readily than the older group (1.24 and 0.87 versus 0.78 and 0.40 respectively). The opposite was found for items 20 and 31, which represent the Action-Awareness Merging dimension (0.10 and 0.14 versus 0.50 and 0.52 respectively). Additional significant differences in item locations were obtained for item 14 ("time different from normal": young, 0.30; older, 0.79), item 22 ("aware of how well I was performing": young, 0.20; older, -0.32), and item 11 ("kept my mind on what was happening": young, 0.28; older, -0.09). The magnitudes of these differences, however, were not striking.

These differences aside, it is apparent from this plot that, although not exactly the same, the location estimates did not vary greatly across these two quite different samples. The flow symptoms that pertain to the Autotelic Experience (ENJY) were "easy" to experience, as were the symptoms of Clear Goals (GOAL) and Unambiguous Feedback (FDBK). These were followed by Challenge-Skill Balance (CHAL), Concentration on Task at Hand (CONC), and Paradox of Control (CONT). Then

Action Awareness (ACT) and Loss of Self-Consciousness (LOSS) were experienced. Only athletes with high flow experience felt the Transformation of Time (TRAN) symptoms. This result indicates that item locations remain relatively consistent across different samples of athletes as well as within each sample for elite versus non-elite athletes and males versus females. This is an indication of the generalizability of this scale across sample-types.

Assessment of Location of Experiences Along the Linear Continuum

In the final stages of these analyses, raw scores from the FSS were converted to logits, the units of measurement that result from a Rasch analysis. The conversion is useful because it provides an estimate of where an athlete is located on the flow continuum that has been converted into a true linear scale with a fixed zero point and equal units of measurement. If the "flow" scale items are scored from 0 - 4, rather than from 1 - 5 as they currently are with the FSS, the total score when all items are summed ranges between 0 - 144. However, since raw scores derived in this manner are not placed on a linear continuum, a score that is transformed into logit units by Rasch analysis is preferable if one wants to know the stage of flow achieved. The transformation from raw scores to logits is presented in Appendix A. Note that none of the subjects in the sample scored under a raw score of 63 which is equivalent to -0.35 logits. The mean logit score of the sample was 1.45 logits, which indicates that most of the items were scored highly by most of the persons.

Discussion

The Rasch analyses conducted here highlighted some aspects of the flow experience that have not been clear in previous writings. There is no doubt that Autotelic Experience, as defined by the items of the FSS, is one of the easiest dimensions of flow to be experienced whereas Transformation of Time, Loss of Self-Consciousness and, to a lesser extent, Action Awareness Merging, are among the least experienced. In other words, autotelic symptoms such as "enjoyed experience", "wanted to capture the feeling", "feeling great", and "extremely rewarding" tended to be experienced extensively by athletes. Jackson (1996) drew a similar conclusion after her qualitative analysis of athletes' experiences.

The Rasch analysis of the FSS subscales demonstrated a high degree of consistency across and within samples, testifying to the reliability

and generalisability of the FSS scale. The Rasch analysis has shown that the various dimensions of flow share a relatively consistent pattern that can be viewed in quantitative stages. It allows us to describe the flow state in terms of the unique experiences that are located on the linear continuum.

It would be interesting to use the FSS with athletes who have completed outstanding performances. Some elite athletes were included in the present samples but they were not necessarily questioned after top level performances. Higher levels of endorsement would be expected for the difficult dimensions, such as Transformation of Time, under these circumstances. Such an investigation would further validate the FSS and give a better indication of the range of the flow experience covered by the scale.

In summary, the present study adds another link to the understanding of the flow experience in athletes. As argued by Jackson and Marsh (1996), any attempts to investigate flow are "fraught with difficulties and limitations" (p. 32). By examining the placement of the Flow Scale (Jackson and Marsh, 1996) items along a linear continuum, it has been possible to show that the state of flow involves a variety of experiential characteristics that may be experienced to different degrees by different athletes. By contrasting the location of items for elite and non-elite, male and female, and older and younger athletes it has also been possible to show that although overall these groups tend to rate the items similarly, Rasch analysis highlights differences that may well be worth exploring further. For example, we would hypothesise on the bases of the analyses conducted here that when in the flow state, females are more conscious of enjoyment, males of competency. Older athletes are more conscious of a sense of achieving goals and younger athletes of a sense of pleasure and enjoyment. Intuitively these findings make sense but they require further elaboration and testing.

References

Andrich, D. (1975). The Rasch multiplicative binomial model: Applications to attitude data. *Research Report Number 1*. Measurement and Statistics Laboratory, Dept. of Education, University of Westen Australia.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47*, 105-113.

Andrich, D, Sheridan, B., and Lyne, A. (1991). *ASCORE: Manual of procedures*. Faculty of Education, University of Western Australia.

Csiksentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper and Row.

Csiksentmihalyi, M. (1993). *The evolving self*. New York: Harper and Row.

Jackson, S. A. (1996). Towards a conceptual understanding of the flow state in elite athletes. *Research Quarterly for Exercise and Sport, 67* (1), 76-90.

Jackson, S. A., and Marsh, H. W. (1996). Development and validation of a scale to measure optimal experience: The Flow State Scale. *Journal of Sport and Exercise Psychology, 18*, 17-35.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Education Research. Reprinted, 1992, Chicago: MESA Press

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., and Stone, M. (1979). *Best test design*. Chicago: MESA Press.

Appendix A

Transformation of Total Raw Scores to Logit Scores for Items in FSS

Raw Score	No. of Respondents	Logit Score	SE
63	1	-.32	.17
64	2	-.29	.17
65	1	-.26	.17
66	0	-.23	.17
67	0	-.20	.17
68	1	-.17	.17
69	0	-.14	.17
70	0	-.11	.17
71	1	-.08	.17
72	2	-.05	.17
73	2	-.02	.17
74	1	.01	.17
75	2	.04	.17
76	2	.07	.17
77	2	.10	.18
78	4	.13	.18
79	1	.16	.18
80	0	.19	.18
81	1	.22	.18
82	2	.25	.18
83	1	.29	.18
84	2	.32	.18
85	6	.35	.18
86	2	.38	.18
87	9	.42	.18
88	5	.45	.18
89	6	.48	.18
90	3	.52	.18
91	7	.55	.19
92	3	.58	.19
93	4	.62	.19
94	5	.66	.19
95	5	.69	.19
96	0	.73	.19
97	4	.76	.19
98	4	.80	.19
99	11	.84	.19
100	5	.88	.20
101	6	.92	.20
102	10	.95	.20
103	5	.99	.20

(Continued on next page)

Appendix A (cont.)

Raw Score	No. of Respondents	Logit Score	SE
104	12	1.04	.20
105	11	1.08	.20
106	15	1.12	.21
107	4	1.16	.21
108	14	1.20	.21
109	10	1.25	.21
110	9	1.29	.21
111	6	1.34	.21
112	7	1.39	.22
113	6	1.43	.22
114	7	1.48	.22
115	10	1.53	.23
116	5	1.58	.23
117	13	1.64	.23
118	4	1.69	.23
119	9	1.75	.24
120	5	1.80	.24
121	6	1.86	.25
122	7	1.92	.25
123	8	1.99	.25
124	8	2.05	.26
125	10	2.12	.26
126	9	2.19	.27
127	3	2.27	.28
128	3	2.34	.28
129	7	2.43	.29
130	7	2.51	.30
131	4	2.60	.31
132	1	2.70	.32
133	6	2.81	.33
134	5	2.92	.34
135	5	3.04	.36
136	4	3.18	.38
137	4	3.33	.40
138	4	3.51	.43
139	2	3.71	.47
140	1	3.95	.52
141	4	4.26	.59
142	3	4.68	.72
143	2	5.40	1.01
Mean		1.45	
SD		.97	

CONTRIBUTOR INFORMATION

Content: *Journal of Outcome Measurement* publishes refereed scholarly work from all academic disciplines relative to outcome measurement. Outcome measurement being defined as the measurement of the result of any intervention designed to alter the physical or mental state of an individual. The *Journal of Outcome Measurement* will consider both theoretical and applied articles that relate to measurement models, scale development, applications, and demonstrations. Given the multi-disciplinary nature of the journal, two broad-based editorial boards have been developed to consider articles falling into the general fields of Health Sciences and Social Sciences.

Book and Software Reviews: The *Journal of Outcome Measurement* publishes only solicited reviews of current books and software. These reviews permit objective assessment of current books and software. Suggestions for reviews are accepted. Original authors will be given the opportunity to respond to all reviews.

Peer Review of Manuscripts: Manuscripts are anonymously peer-reviewed by two experts appropriate for the topic and content. The editor is responsible for guaranteeing anonymity of the author(s) and reviewers during the review process. The review normally takes three (3) months.

Manuscript Preparation: Manuscripts should be prepared according to the *Publication Manual of the American Psychological Association* (4th ed., 1994). Limit manuscripts to 25 pages of text, exclusive of tables and figures. Manuscripts must be double spaced including the title page, abstract, text, quotes, acknowledgments, references, and appendices. On the cover page list author name(s), affiliation(s), address(es), telephone number(s), and electronic mail address(es). On the second page include a 100 to 150 word abstract. Place tables on separate pages. Include photocopies of all figures. Number all pages consecutively.

Authors are responsible for all statements made in their work and for obtaining permission from copyright owners to reprint or adapt a table or figure or to reprint a quotation of 500 words or more. Copies of all permissions and credit lines must be submitted.

Manuscript Submission: Submit four (4) manuscript copies to Richard M. Smith, Editor, *Journal of Outcome Measurement*, Rehabilitation Foundation Inc., P.O. Box 675, Wheaton, IL 60189 (e-mail: jomea@rfi.org). Prepare three copies of the manuscript for peer review by removing references to author(s) and institution(s). In a cover letter, authors should indicate that the manuscript includes only original material that has not been previously published and is not under review elsewhere. After manuscripts are accepted authors are asked to submit a final copy of the manuscript, original graphic files and camera-ready figures, a copy of the final manuscript in WordPerfect format on a 3 ½ in. disk for IBM-compatible personal computers, and sign and return a copyright-transfer agreement.

Production Notes: Manuscripts are copy-edited and composed into page proofs. Authors review proofs before publication.

SUBSCRIBER INFORMATION

Journal of Outcome Measurement is published four times a year and is available on a calendar basis. Individual volume rates are $35.00 per year. Institutional subscriptions are available for $100 per year. There is an additional $24.00 charge for postage outside of the United States and Canada. Funds are payable in U.S. currency. Send subscription orders, information requests, and address changes to the Subscription Services, Rehabilitation Foundation, Inc. P.O. Box 675, Wheaton, IL 60189. Claims for missing issues cannot be honored beyond 6 months after mailing date. Duplicate copies cannot be sent to replace issues not delivered due to failure to notify publisher of change of address. Back issues are available at a cost of $12.00 per issue postpaid. Please address inquiries to the address listed above.

 ISSN 1090-655X.