SOCIAL BEHAVIOR AND PERSONALITY, 2007, 35 (9), 1163-1172

FUZZY PARTIAL CREDIT SCALING: A VALID APPROACH FOR SCORING THE BECK DEPRESSION INVENTORY

SEN-CHI YU
Huafan University, Taiwan
MIN-NING YU
National Chengchi University, Taiwan

In this study a new scaling method was proposed and validated, fuzzy partial credit scaling (FPCS), which combines fuzzy set theory (FST; Zadeh, 1965) with the partial credit model (PCM) for scoring the Beck Depression Inventory (BDI-II; Beck, Steer, & Brown, 1996). To achieve this, the Chinese version of the BDI-II (C-BDI-II) was administered to a clinical sample of outpatients suffering depression, and also to a nonclinical sample. Detailed FPCS procedures were illustrated and the raw score and FPCS were compared in terms of reliability and validity. The Cronbach alpha coefficient showed that the reliability of C-BDI-II was higher in FPCS than in raw score. Moreover, the analytical results showed that, via FPCS, the probability of correct classification of clinical and nonclinical was increased from 73.2% to 80.3%. That is, BDI scoring via FPCS achieves more accurate depression predictions than does raw score. Via FPCS, erroneous judgments regarding depression can be eliminated and medical costs associated with depression can be reduced. This study empirically showed that FST can be applied to psychological research as well as engineering. FST characterizes latent traits or human thinking more accurately than does crisp binary logic.

Keywords: fuzzy partial credit scaling, fuzzy set theory, Rasch model, depression, Beck Depression Inventory.

Sen-Chi Yu, Assistant Professor, Center for Teacher Education, Huafan University, Taipei County, Taiwan; and Min-Ning Yu, Professor, Department of Education, National Chengchi University, Taipei City, Taiwan.

The authors would like to thank the National Science Council of Taiwan for their support in the form of a research grant (project number: NSC-95-2413-H-211-001).

Appreciation is due to reviewers including: Steven H. Aggen, PhD, Department of Psychiatry, Virginia Commonwealth University, Virginia Institute for Psychiatric & Behavioral Genetics, 80 East Leigh St., Suite 1-120A, Richmond, VA 23219-1534, USA, Email: **ssaggen@vcu.edu**; Berlin Wu, PhD, Department of Mathematics Science, National Chengchi University, No. 64, Sec. 2, Zhinan Rd., Taipei City, 11605, Taiwan, Email: **berlin@math.nccu.edu.tw**

Please address correspondence and reprint requests to: Sen-Chi Yu, Center for Teacher Education, Huafan University, No.1, Huafan Rd., Shiding, Taipei County 223, Taiwan (R.O.C.). Phone: 886-2-26632102 ext. 3419; Fax: 886-2-29387737; Email: **rhine@cc.hfu.edu.tw**

Depression is among the most pervasive psychological problems in primary healthcare settings, accounting for 10.4% of all patients seen in such settings globally (Endler, Macrodimitris, & Kocovski, 2000). Self-reported measures of depression are most straightforward and important tools in various healthcare settings in the diagnosis and classification of different levels of depression. Therefore, a valid scoring schema is essential to accurately reflect the severity of depressive symptoms. The most popular scoring schema applied in psychological inventories is raw score, or "method of successive integral". In this scoring schema, alternatives listed in the scale are scored at equally spaced intervals. For example, a score of 4, 3, 2, or 1 is given if the alternatives *strongly agree*, *agree*, *disagree*, or *strongly disagree* respectively, are chosen. However, this approach has been criticized on the grounds that it is too simplistic (Nunnally & Bernstein, 1994; Yu, 2005).

First, raw score fails to achieve "meaningful measurement" for nonlinearity, and sample and test dependence (Wright, 1999). By contrast, item response theory (IRT) approach. Rasch models (Rasch, 1960) transform raw score into linear measures and, consequently, achieve a more objective and meaningful psychological measurement. Second, the options used in the rating scales, without clear and mutually exclusive distinctions, could be viewed as "linguistic variables". Linguistic variables, as defined in fuzzy set theory (FST; Zadeh, 1965), are variables of which the values are not numbers but words or sentences in a natural or artificial language (Klir & Yuan, 1995; Zimmermann, 1996). For instance, "sadness", a question adapted in the Beck Depression Inventory II (BDI-II; Beck, Steer, & Brown, 1996), is a linguistic variable if it takes a value such as "I felt sad all the time", or "I felt sad much of the time". Moreover, these terms are not clearly defined and no definite boundaries exist between, for example, "much of the time" and "all the time". Lacking clear definitions for the variables, the arithmetic performed on linguistic variables is beyond the capability of traditional binary crisp logic. Therefore, the newly developed fuzzy logic is the preferred solution for measurement.

Furthermore, the distinctions between two adjacent alternatives may be so polarized or extreme that none of the alternatives can reflect an individual's mental state exactly. Considering the example quoted above, the discrepancy between two adjacent alternatives such as "I did not feel sad," and "I felt sad much of the time" seems so strong that examinees who felt sad only occasionally would not be easily able to select an alternative. Under such circumstances, assuming someone entirely belongs to one particular alternative may be questionable and debatable. Such an assumption of a crisp set view originated in Aristotle's binary logic, where each individual can be dichotomized into a set member (those who certainly belong to the set) or nonmember (those who certainly do not). Carrying on such logic, test constructors ask examinees to choose one alternative (set) in

each item. However, people generally feel depressed or happy in the continuum within two opposing extremes rather than as a yes-or-no dichotomy. Therefore, human thinking is multivalued, transitional and analogue, instead of bivalued, clear-cut, and digital. Fuzzy set theory, by providing a systematic framework for dealing with the vagueness and imprecision in human thoughts, is a powerful tool to analyze and animate human thinking (Dubois, Ostasiewicz, & Prade, 2000). Nevertheless, few FST applications have been found in psychometric studies.

Considering the FST, the degree to which an element belongs to a given set, denoted by "membership", is a continuous value, gradually changing from zero to one (Kosko, 1993). Therefore, a fuzzy set can be defined mathematically by assigning a value representing its membership grades to each possible individual in the universal discourse. The membership function, or the character function of a fuzzy set, corresponds to the level of similarity, likelihood, or compatibility with the concept represented by the fuzzy set (Bilgic & Turksen, 2000; Dubois et al., 2000; Zimmermann, 1996).

FST has advanced in many disciplines: for example, in artificial intelligence, computer science, decision theory, logic and pattern recognition (Dubois et al., 2000). Since FST provides a systematic framework for dealing with the vagueness and imprecision inherent in the human thought process, it should be beneficial in psychometric investigations for several reasons. First, variables of interest in psychology are poorly defined, latent, and imprecise, corresponding to the vagueness in FST. Second, each item presented in psychological inventories could be regarded as a linguistic variable. However, in contrast with the many engineering studies discussing FST, only a few such works have been published in psychological measurement. These works include a series of studies conducted by Berlin Wu and his associates (Nguyen & Wu, 2006; Wu & Lin, 2002), which revealed that the FST is more efficient in predicting human behavior and public opinions than is traditional logic. Furthermore, another series of studies by Yuan-hong Lin (Wu & Lin, 2002) demonstrated that, based on computer simulations and real case study, the fuzzy set approach is more reliable and accurate than the traditional scoring of Likert scales. Up to now, the scaling and membership generating of these studies, using FST in psychological measurement has been based on classical test theory rather item response theory (IRT). However, as mentioned above, raw score suffices to accomplish a "meaningful measurement" (Wright, 1999).

The present study proposed a new scaling method, fuzzy partial credit scaling (FPCS), which utilizes partial credit model (PCM; Masters, 1982), an IRT approach one parameter logistic model, to construct fuzzy numbers and utilize these fuzzy numbers to score psychological measurements. To certify whether FPCS is a more valid scoring approach than raw score, the reliability and predictive validity of FPCS and that of raw score were compared.

METHOD

Traditional and Fuzzy Scoring

In FPCS, subjects are free to choose more than one alternative for each item and, in turn, assign percentages on the chosen alternatives. The assigned percentages represent the degree of membership that subjects feel to each category. Moreover, the sum of percentages of the chosen categories is restricted to 100%. Next, the triangular normal fuzzy numbers $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{D}$ were constructed to represent alternatives 1 to 4, respectively.

Table 1 shows the examples of fuzzy scoring (FS) and traditional scoring. As shown in this table, the category assigned the highest percentage is treated as the traditional scoring. If there are two most assigned categories, the lower score will be taken as traditional scoring. The sum of fuzzy numbers multiplied by its membership degree, constitutes the fuzzy scoring. Since the calculations of PCM require a crisp number, the results of traditional scoring were utilized as crisp data for PCM algorithms. The results of fuzzy scoring, still fuzzy numbers, will be utilized for sequent analysis.

TABLE 1
Examples of Fuzzy and Traditional Scoring: Two Alternatives Chosen

	Assigned Percentages (Degree of Membership)	Traditional Scoring (crisp value)	Fuzzy Scoring(FS) $FS = \sum \mu_{ijk} (\tilde{K})$ (interval value)
Alternative 1* ($\tilde{A}$)	80%	**1**	$0.8 \times \tilde{A} + 0.2 \times \tilde{B}$
Alternative 2 ($\tilde{B}$)	20%		
Alternative 3 ($\tilde{C}$)	0%		
Alternative 4 ($\tilde{D}$)	0%		

Note: * indicates the category assigned the highest percentage.

Fuzzy Partial Credit Scaling

Generating Fuzzy Numbers This study proposed triangular fuzzy number modified from Yu (2005) for scoring psychological measurements and the procedures were as follows:

Step 1: Subjects are asked to choose and assign percentages on alternatives of items. The sum of assigned percentages, representing the membership degrees, in each item must be constrained to 100%.

Step 2: Calculate the traditional scoring according to the procedures mentioned above.

Step 3: Calculate "step parameters" (δ_{ij}) defined in PCM

Step 4: Fuzzify crisp data into fuzzy data by constructing triangular fuzzy numbers using step parameters estimated in Step 3.

We tried to map linguistic variables, alternatives 1 to 4, into corresponding reasonable normal fuzzy numbers $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{D}$, with triangular membership functions $\mu_{\tilde{A}}$, $\mu_{\tilde{B}}$, $\mu_{\tilde{C}}$, and $\mu_{\tilde{D}}$. These membership functions are shown in Figure 1.

The x-axis represents ability, usually ranging from -3 to 3; while the y-axis represents degree of membership, ranging from 0 to 1.

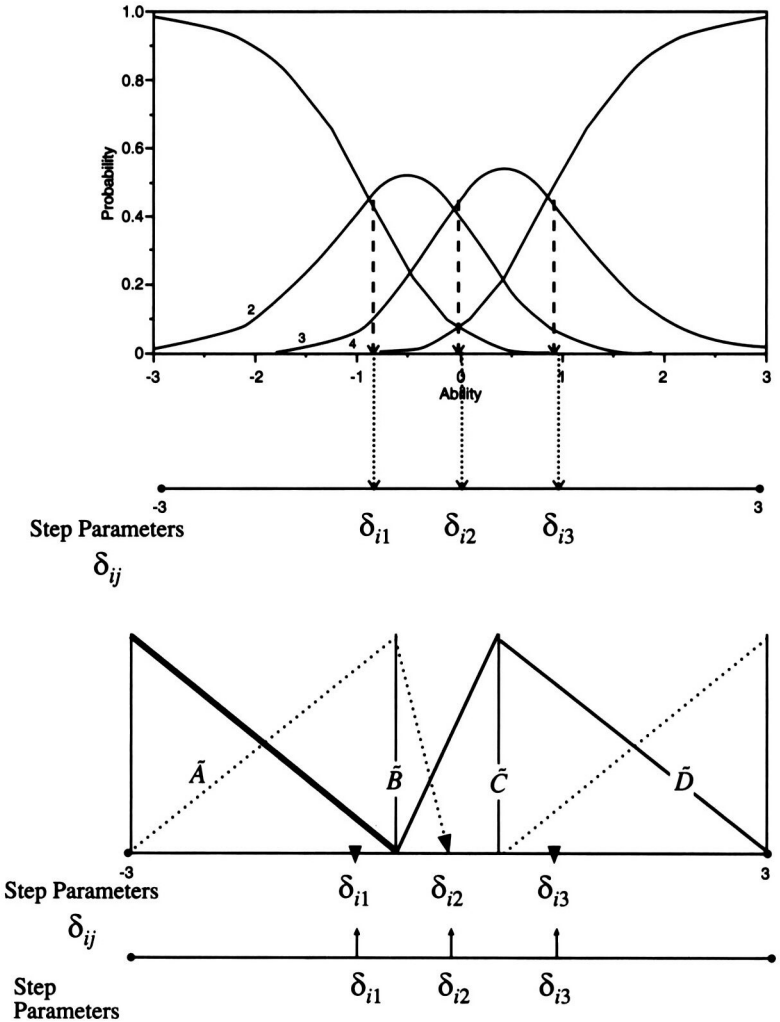

Figure 1: Generating Triangular Fuzzy Numbers via Step Parameters.

In Figure 1, we first found the "step parameters" (δ_{ik}) estimated by PCM. We proposed that a subject with ability located between -3 and "step parameter 1" (δ_{i1}) will choose alternative 1. For this reason, the triangular fuzzy number $\tilde{A}$ = (-3, -3), (δ_{i1} + δ_{i2})/2) with -3 and (δ_{i1} + δ_{i2})/2 being the lower and upper bounds, respectively, and -3 as the most likely value for $\tilde{A}$. In Figure 1, we drew a line segment from (-3 , 1) to (-3, 0) and ((δ_{i1} + δ_{i2})/2, 0) to characterize the membership of function of $\tilde{A}$.

Next, we proposed that a subject with ability located between "step parameter 1" (δ_{i1}) and "step parameter 2" (δ_{i2}) will choose alternative 2 and the middle point between these two step parameters should receive the maximum degree of membership. Therefore, the triangular fuzzy number $\tilde{B}$ = (3, (δ_{i1} + δ_{i2})/2, (δ_{i2} + δ_{i3})/2) with -3 and (δ_{i2} + δ_{i3})/2 being the lower and upper bounds, respectively, and (δ_{i1} + δ_{i2})/2 being the middle point which is the most likely value for $\tilde{B}$. In Figure 1, we drew a line segment from ((δ_{i1} + δ_{i2})/2, 1) to (-3, 0) to represent the left leg and another line segment from ((δ_{i1} + δ_{i2})/2, 1) to (δ_{i2} + δ_{i3})/2, 0) to represent the right leg of the triangular fuzzy number.

Likewise, we proposed $\tilde{C}$ = ((δ_{i1} + δ_{i2})/2, (δ_{i2} + δ_{i3})/2, 3) and $\tilde{D}$= ((δ_{i2} + δ_{i3})/2, 3, 3) to characterize the likelihood of alternatives 3 and 4, respectively.

Scoring of FPCS The addition of triangular fuzzy numbers $\tilde{M}$ (m, α, β) and $\tilde{N}$ = can be defined as $\tilde{M}$(+) $\tilde{N}$ = (m + n, α + γ, β + δ) and the subtraction is $\tilde{M}$(-) $\tilde{N}$ = (m - n, α + δ, β + γ) (Chen & Huang, 1992). Therefore, assuming subject j completed a three-item scale. The scoring of the three items were denoted as triangular fuzzy numbers i_1 = (0, 1, 2) i_2 = (1, 2, 3), and i_3 = (0, 1, 2) respectively. Consequently, the aggregate fuzzy score (AFS), still a fuzzy number, was: AFS = (0, 1, 2) + (1, 2, 3) + (0, 1, 2)= (1, 4, 7).

For sequent statistical operation, AFS was defuzzified into a crisp number using the center of gravity (COG) method. COG calculates the center of gravity of the support of the fuzzy number weighted by the membership grade. The center of gravity of fuzzy set $\tilde{X}$ with membership function $\mu_{\tilde{A}}$, GR(X) = $\frac{\int_{-\infty}^{\infty} x\mu_x(x)dx}{\int_{-\infty}^{\infty} \mu_x(x)dx}$

For a triangular fuzzy number $\tilde{X}$ (a, b, c), GR(X) = (a+b+c)/3 (Zimmermann, 1996). The defuzzified AFS, called total fuzzy score (TFS) were used for sequent statistical analysis.

Sample and Procedure

The total sample used in this study consisted of participants recruited from two separate populations: (a) outpatients of a psychiatric clinic who were diagnosed as suffering from depression as the clinical sample, and (b) undergraduates as the nonclinical sample.

Since depression symptoms may appear in many mental disorders, the diagnosis of the outpatients who took part in this study included the following disorders: Major Depression Disorder, Bipolar Disorders, Dysthymic Disorder, and Adjustment Disorder with Clinical Mood. Outpatients who were in partial or full remission of depression might have low total BDI-II scores, and this might have contaminated the classification results of clinical and nonclinical depression. Therefore, 36 outpatients whose depression was diagnosed as in remission and in partial remission were eliminated and the other 204 (123 female and 81 male) subjects were retained in the clinical sample. Participants ranged in age from 15 to 78 years (M = 35.56, SD = 14.09). The self-report instrument utilized in this study was administered by the researcher while the severity of depression was diagnosed by a psychiatrist. The diagnosis was the external criterion to compare the predictive validity of BDI-II via FPCS and raw score. Informed consent was obtained from patients, and participation was voluntary. As for the nonclinical sample, a total of 321 (265 female and 56 male) students in Taiwan were recruited. Participants ranged in age from 18 to 39 years (M = 25.30, SD = 5.36).

Instrument

The instrument in this study was the Chinese version of the Beck Depression Inventory II (C-BDI-II). BDI-II (Beck et al., 1996) is a self-report instrument for measuring the severity of depression in adolescents and adults through items showing varying degrees of the main cognitive, affective, and physiological aspects of clinical depression. The C-BDI-II was adapted from the original BDI-II by the Chinese Behavioral Sciences Society and made available in 2000.

A number of studies have generally found that the BDI–II has high internal consistency and moderate to strong convergent validities with other self-report measures (Krefetz, Steer, Gulab, & Beck, 2002).

Since this study applied IRT approach analysis, the dimensionality of the BDI must be examined. The dimensionality of the BDI was evaluated via principal component analysis (PCA) and fit indices. The analytical results of PCA on the BDI showed that the first eigenvalue was 10.44 while the second eigenvalue was only 1.09; therefore, a dominant factor exists. According to Stout's "essential unidimensionality", the dominant factor is so strong that the examinee's trait level is robust to the presence of smaller specific factors (Yu, 2005). Judging from these, we concluded that the unidimensionality of the BDI was tenable.

The INFIT mean square (MNSQ) fit indices were applied to evaluate the goodness of fit. An INFIT MNSQ value of (1+ x) indicates (100x)% more variation between the observed and the expected value. Reasonable MNSQ ranges for clinical observations are 0.5-1.7 (Bond & Fox, 2001). The analytical results showed that INFIT MNSQ of items of BDI ranged from 0.78-1.40,

indicating a reasonable fit. Given these findings, we concluded that the BDI is unidimensional.

RESULTS

To verify whether FPCS is a more valid scoring method than raw score, the reliability and validity of FPCS and that of raw score were compared and were listed as follows.

FPCS Reliability

Since the fuzzy data generated by FPCS are interval-valued fuzzy numbers rather than precisely valued crisp numbers, traditional IRT based reliability indices cannot be computed owing to the nature of the number. Therefore, the Cronbach alpha coefficient was utilized to measure reliability.

The analytical results demonstrated that the alpha coefficient for FPCS was .951, while that of raw score was only .939. The FPCS thus achieved higher reliability than raw score.

FPCS Validity

In this study, two different scoring schema, raw scores and FPCS, yielded two different predictors, while diagnosis of depression by a psychiatrist provided the external criterion. This study applied logistic regression to investigate the relation between scoring schemas and diagnosis of suffering from depression (binary outcome).

Concerning FPCS, the estimated regression function was $\hat{y}$ 2.281 + 0.01 x_1. Where x_1 denotes scoring via FPCS. The Wald statistics equal 131.957 ($p < .001$), showing that the regression coefficient is significant at $\alpha = .001$. The probability of correctly classifying clinical and nonclinical depression was 80.3%

Regarding raw score, the estimated regression function was $\hat{y}$ -2.097 + .134 x_1. Where x_1 denotes scoring via raw score. The Wald statistics equaled 131.414 ($p < .001$), showing that the regression coefficient is significant at $\alpha = .001$. The probability of correct classification of clinical and nonclinical depression was only 73.2%.

Clearly, the predictive validity of raw score is inferior to that of FPCS. These findings reveal that FPCS, compared with raw scores, yields better model fit and can more accurately predict depression.

DISCUSSION

Measurement errors comprise systematic and nonsystematic error components. In this study it was argued that certain systematic error components, such as

errors of leniency and severity, could be eradicated through multivalued fuzzy logic.

The error of leniency (severity) indicates that raters tend to rate higher (lower) than they should (Guilford, 1954). Taking as an example the item "sadness" presented in the BDI, individuals who only feel sad occasionally must choose between "I do not feel sad" (the first alternative) and "I feel sad much of the time" (the second alternative). However, the distinction between two adjacent alternatives is so polarized or extreme that neither alternative can reflect exactly an individual's mental state. When respondents are asked to choose just a single alternative from the rating scales to describe their mood state or attitude, force fitting and rounding off are inevitable. Such reductions may involve errors of leniency and severity, reducing instrument reliability. Comparatively, FST uses transitional membership degrees to characterize individual state and force fitting and rounding off are avoided.

Predictive validity was employed in this study to investigate the validity of FPCS. In this study, via FPCS, 7.1% of erroneous judgments regarding depression inferred from self-reported inventory were reduced. Regarding the costs associated with depression, the US spends $43.7 billion annually on medical expenses and lost productivity (Endler et al., 2000), and it causes inestimable human suffering. This study showed that the FPCS provides a more accurate scoring schema than does raw score. Through FPCS, erroneous judgments of depression can be eliminated and related medical costs can be reduced.

The analytical results also indicate that fuzzy logic conveys human thinking more accurately than does crisp logic. Theoretically, fuzzy logic offers researchers an interpretive algebra, a language that is half verbal conceptual and mathematical analytical (Ragin, 2000). The membership degree of FST, which conveys both qualitative and quantitative properties of the chosen alternatives, can characterize the measured latent construct more genuinely and honestly. Based on these findings, FPCS was empirically verified as a valid scoring schema for psychological measurement.

REFERENCES

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory–II*. San Antonio, TX: Psychological Corporation.

Bilgic, T., & Turksen, B. (2000). Measurement of membership function: Theoretical and empirical work. In D. Dubois & H. Prade (Eds.), *Fundamentals of fuzzy sets* (pp. 195-232). Boston: Kluwer.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwam, NJ: Lawrence Erlbaum Associates.

Chen, S. J., & Huang, C. L. (1992). *Fuzzy multiple attribute decision making: Methods and applications*. Berlin: Springer-Verlag.

Dubois, D., Ostasiewicz, W., & Prade, H. (2000). Fuzzy sets: History and basic notions. In D. Dubois & H. Prade (Eds.), *Fundamentals of fuzzy sets* (pp. 21-124). Boston: Kluwer.

Endler, N. S., Macrodimitris, S. D., & Kocovski, N. L. (2000). Controllability in cognitive and interpersonal tasks: Is control good for you? *Personality and Individual Differences*, **29**, 951-962.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Klir, G., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic: Theory and applications*. Englewood Cliffs, NJ: Prentice Hall.

Kosko, B. (1993). *Fuzzy thinking: The new science of fuzzy logic*. New York: Hyperion.

Krefetz, D. G., Steer, R. A., Gulab, N. A., & Beck, A. T. (2002). Convergent validity of the Beck Depression Inventory-II with the Reynolds Adolescent Depression Scale in psychiatric inpatients. *Journal of Personality Assessments*, **78** (3), 451-460.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, **47**, 149-174.

Nguyen, H. T., & Wu, B. (2006). *Fundamentals of statistics with fuzzy data*. New York: Springer.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Ragin, C. C. (2000). *Fuzzy-set social science*. Chicago: University of Chicago Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. [Danish Institute of Educational Research 1960, University of Chicago Press 1980, MESA Press 1993] Chicago: MESA Press.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know* (pp. 235-255). Hillsdale, NJ: Lawrence Erlbaum.

Wu, B. L., & Lin, Y. H. (2002). *Fuzzy mode and its applications in educational and psychological assessment analysis*. Paper presented at the 2nd International Conference on Information, Beijing, China.

Yu, S. (2005). *Fuzzy partial credit scaling: Applying fuzzy set theory to scoring rating scales*. Unpublished doctoral dissertation, National Chengchi University, Taipei, Taiwan.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, **8**, 338-352.

Zimmermann, H. J. (1996). *Fuzzy set theory and its applications*. Boston: Kluwer.