DEVELOPING A MEASURE OF AUTHENTIC ASSESSMENT STANDARD FOR CHILDREN'S DEVELOPMENT AND LEARNING USING MANY-FACET RASCH MODEL

Nor Mashitah M.R.[1], Mariani M.N.[2], Jain Chee [3], Che' Mah Binti Yusof[4]
Universiti Malaya

Mas_mradzi@yahoo.com.my

ABSTRACT

The Authentic Assessment Standards (AAS) for young children was developed to provide standard assessment instrument that covers standard learning curriculum and domains of child development based on children's performance. The purpose of this study is to investigate AAS score reliability and validity on the performance including children's ability, judges effect and domain difficulty using Many-Facet Rasch Model (MFRM). In this study, data were collected from children aged three and four years by two independent judges and the total measure using data for 60 children. MFRM was applied to analyze the data. The results indicated that all children can perform the task given at different levels depending on the domain and item difficulty. It has been found that judges' effect in scoring and when item statistics were examined, they fulfilled the purpose of evaluation for young children. Therefore, the present study illustrates a procedure for evaluating and improving such measures using a highly flexible and sophisticated psychometric approach among children by authentic assessment. Standardized measurement tools including documentation with systematic assessment and measurement models were recommended to increase quality of evaluation in early childhood education

Key Words: authentic assessment, children's performance, rasch measurement, MFRM

INTRODUCTION

During most of the early childhood years, it is difficult to measure and evaluate bits of knowledge and skills that are isolated from other types of knowledge and skills. Young children are not reliable test takers due to the many different confining personal, developmental, and environmental factors, which affect their behaviours. However, many challenges associated with assessing the learning and development of young children have been well documented (Shonkoff & Phillips, 2000; Snow & Van Hemel, 2008). Program for young children and their families are oneof the ways to increase the quality of life for children. However it was not supported by Evans (2001) as he claimed that in order to ensure children's rights. Efforts have to be made to provide for the basic needs of young children. He indicates four main facets in Early Childhood Care and Development (ECCD) programmes. First, children have the right to develop their potentials, and should provide children with the best possible intellectual, physical and social tools that would enable them to reach their highest intellectual, physical, and social capabilities. Secondly, early attention to the child's needs is critical and leads to long-term benefits for the child. Evidence from the related area showed that the early years of one’s life are the most critical. Next, is the derived benefit for parents, family and communities at large.

A daunting task to any administrator, researcher and practitioner is to identify an instrument that will be reliably and valid to assess the tremendous amount of development and learning that occurs during the first 5 years of a child’s life. The development of infants, toddlers, and preschool children is influenced by a variety of factors such as genetics, environment, and culture. Young children acquire new behaviours and skills at varying rates and learn in differing ways. The emerging developmental competencies of infants and toddlers give way to more differentiated and refined skills and behaviours of preschool children. Development and learning occur within the context of family as well as during time spent in out-of-home care. The behavioural and attention spans of young children differ significantly, and there is a variation in how the developmental process that unfolds within a young child as well as between children during the first 5 years of life.

The developmental process adds to the complexity of factors to consider in selecting appropriate instruments (Shonkoff & Phillips, 2000; Snow & Van Hemel, 2008). In an effort to ensure that information obtained from assessments represents authentic and accurate portrayals of young children’s development and learning, professional organizations have formulated guidelines that address what is considered “best practice” for assessment in early childhood (Kim & Smith, 2010).

Traditional or conventional practices of using standardized or norm-referenced measures to assess the early acquisition of skills in young children are being replaced by developmental approaches to assessment. Bagnato (2005) contends that “early childhood measurement is significant into authentic assessment, the optimal alternative to conventional testing” (p. 17). An authentic approach to assessment uses materials and activities that are familiar to the child to examine skills and behaviors that occur throughout his or her daily activities and routines (Neisworth & Bagnato, 2005). This developmental and authentic approach to assessment is intended to focus on the identification of young children’s developmental strengths as well as areas of concern. Information obtained from authentic assessments can then be used to provide functional information for professionals for the purposes of planning, implementing, and evaluating developmentally appropriate experiences and interventions (Bagnato, Neisworth, & Munson, 1997). Neisworth and Bagnato (2005) recommend that assessment information should (a) be useful; (b) be socially valid; (c) be authentic; (d) be collaborative across families and professionals; (e) provide “functional, reliable, and valid information” (p. 49); (f) “accommodate individual differences” (p. 49); (g) be sensitive enough to detect small changes in the progression of developmental skills; and (h) be designed for and field validated with the populations of children who will be assessed using the measure.

The topic of assessing young children’s development and learning is also being discussed within the context of educational outcomes and school readiness. Early care and educational programs have become increasingly accountable for “promoting standards of learning and monitoring children’s progress in meeting those standards” (Snow & Van

Hemel, 2008, p. 1) As assessment outcomes generate newly found significance with respect to educational implications, instruments that assess young children's development and learning should have a "one-to-one correspondence to measures taken later" (Snow &Van Hemel, 2008, p. 73) and must be evaluated for evidence of psychometric properties (e.g., reliability, validity), appropriateness for diverse populations, and the domains being assessed. Evidence usefulness and quality of the information collected through the use of the measure need to demonstrate validity and reliability.

Authentic Assessment Standard (AAS) well- known observational measures based on children's performance for assessing young children's progress represent a standard on their performance. Authentic approach is an appropriate systematic method and classifies children's learning and development in six domain developments such as cognitive, language, physical, creativity, socio-emotional and spiritual morality. This standard assessment includes the standard learning curriculum and domains of child development based on children's performance. Besides that, measurement models to measure many facets of children was implemented. Raters are asked to observe and evaluate 183 items representing developmental domains by using a 5-point rating scale. The Vygotsky's area of Zone of Proximal Development looks deeply at the ability and level of development of children as individual persons. *Scaffolding* emphasizes on children's level of performance in problem solving skills. AAS was also constructed on the view that individual variability of children's learning styles can affect the achievement in every activity according the responses. The focus of this authentic assessment is monitoring and assessment of each individual child's development. An important component of this assessment is the systematic approach, provided to teacher, a framework for systematically collecting and processing children's work and performances. In addition, AAS describes how to observe, how to collect, and how to analyze the observation.

Many Facet-Rasch Model (MFRM) was applied to analyze the collected data (Rasch, 1980). MFRM has been developed in recent years to overcome some of the problems and assumptions associated with Classical Test Theory and to provide information for decision-making that is not available through Classical Test Theory (Linacre, 1993). MFRM has several distinct advantages over classical data analysis (Smith, Schumaker & Bush, 1998; Elhan & Atakurt, 2005). First, Rasch measurement places each facet of the measurement context on a common underlying linear scale. This result in a measure can be subjected to traditional statistical analysis, while allowing for unambiguous interpretation of age group performance as it relates to judging severity and domain difficulty. Second, the Rasch-based calibration examines, domain difficulty and judging is sample-free. In other words, Rasch techniques remove the influence of sampling variability from its measures so that valid generalizations can be made beyond the current sample of groups, collections of domains and judges. Third, Rasch 'fit' procedures can be used to derive unexpected response patterns that are useful for evaluating the extent to which individual age groups, domains or judges are behaving in ways that are inconsistent with the measurement model (Engelhard, 1992; Wright & Linacre, 1994). FACETS software program developed by Linacre (1993) was used to apply MFRM. This software was able to give detailed information about the calibration of the three aspects of the study (group age performance, domain difficulty and judge severity or lenient.

According to this extension of Rasch, MFRM have twofold: (a) there is no restriction to the analysis of only two facets (children and items), and (b) the data being analyzed need not be dichotomous. In an analysis of performance assessments, the MFRM model allows one to take account of additional facets of that setting that may be of particular interest, such as raters, tasks and criteria. Moreover, raters award scores to children using ordered scale categories (rating scales).Therefore, the data in most instances comprise multiple responses. For each facet, the model represents each element such as each individual child, rater, item, domain by a separate parameter value. The parameters denotes distinct attributes of the facets involved in, such as, proficiency of children, severity of rater, and difficulty of items, domains or scoring criteria. Following the principle of measurement invariance, when the data fits the model, these measure compensates for rater severity or leniency differences, that is, the examinee proficiency measures are independent of the particular sample of raters involved. The purpose of this study was to evaluate AAS score reliability and validity of scores from children's measure and performance including children's ability, judges effect and domain difficulty using Many-Facet Rasch Model (MFRM).

Method

Design and Participants

Sixty children aged 3 years (n: 30) and 4 years (n: 30) in private nursery participated in the study. This study focuses on all domain of child development in natural settings and each childis responses to the activities. The children's background is not taken into account and all children are given the same activities guided bya teachers.

Raters

To obtain reliable scores for children's performance, raters should have mastered the rating of items and should have practical experiences as well as theoretical knowledge of the constructs to be measured (Wolfe, 2004). Thus three, adapted early into childhood expertise and satisfied the following criteria (a) majored in early childhood education at the doctoral level, (b) have enrolled in graduate courses related to domain of child development and learning standard curriculum, (c) have training with the Authentic Assessment Standard (AAS) in terms of administration and rating at least 4 times, and (d) have taught children for at least 5 years in the education settings, and were trained as raters. All raters were informed about the purpose of this study, and independently completed their ratings.

Instrument

The AAS is a qualitative measure of children's learning and development in holistic skills performance aged 3 and 4 years. The instrument covers six domain development such as cognitive, language, physical, creativity, socioemotional and spiritual. Each sub-test is performed twice and includes a different number of items ranging from a minimum of one to a maximum of three. Every rater scores each item with 1 (less ability) to 5 (most ability), depending on whether the participant's ability to perform each item with or without guidance.

Procedures

Two trained raters administered all thetest sessions. The first rater took part in the main instruction as well as demonstration, while the second rater assisted to control measurement condition and videotaping. All test sessions were conducted in a nursey setting during routine and learning process. The tests were administered following sequences according to the AAS Manual.

Data Analysis

Many-Facet Rasch Model (MFRM)

The standard Rasch model comprises two facets, item difficulty and person ability dichotomous responses (Rasch, 1996). This model shows the probability of a respondent endorsing a given item. This is the net result of the interaction between the ability of the respondent and the difficulty of the item (Wright & Mok, 2000). Based on Linacre (1994), the MFRM is an extension of the basic Rasch model. For MFRM, the probability of a given AAS response is the result of the children's level of performance, the difficulty of the domain, and the leniency of the rater or judges (the tendency of the rater to be a harsh or lenient judge by providing ratings that are systematically low or systematically high, Linacre & Wright, 2002). The MFRM was used to evaluate the level of ability of children (facet 1) adjusting for the effects of the rater (facet 2), difficulty of domain (facet 3). The MFRM was calibrated using FACETS computer software (Linacre, 2004).

Steps of MFRM Analyses and Interpretive Guidelines

The MFRM includes several critical steps.

.

Rating scale functioning. A FACET (Linacre, 2004a) provides severalindicators of adequacy of rating scale performance. Six indicators wereevaluated in the present analyses, as detailed by Linacre (2002) andBond and Fox (2001). First, "category uses statistics," or the frequency ofresponses in each category, were assessed for a consistent distributionacross rating categories. A recommended guideline is at least 10 observed responses per category. Second, the average Rasch respondent leniency estimates for those who endorsed a given response category and were examined to assess the degree to which higher category utilization was associated with increasing respondent leniency. Third, threshold estimates, indicating the point at which the probability of endorsement of two adjacent categories is equal, should be spaced by at least 1.0 logit, indicating that each rating scale transition is a distinct point on the rating scale continuum. Fourth, step calibrations were evaluated to determine difficulty of selecting one response category over another and should increase as the response category increases. Fifth, category fit statistics were examined as an indication of the degree to which categories performed as predicted. Standardized OUTFIT values exceeding 2.0 indicate that the category contributed more "noise" than precision to the data. Sixth, the response probability curves provided an illustration of the statistics described above.

Model fit. FACETS provides several statistics for both items and respondents that quantify the degree to which the observed data fit the expected model. Fit statistics are based on the differences between the observed and expected response for each person on each item (Bond & Fox, 2001). Thus, small residual values indicate that the observed response was close to the expectation, and large residual values indicate that the response was unexpected. The ZSTD OUTFIT value has been found to perform most effectively for identifying misfitting items (Smith, 2000; Smith, Schumaker, & Bush, 1998).

Separation reliability. FACETS also provide separation reliability estimates for each facet in the model. Separation reliability refers to the number of levels of a given facet reliably differentiated by the other facets in the model (Smith, 2001). In the present context, the item separation reliability indicates the number of levels of therapist adherence defined by the items (Bond & Fox, 2001).

Results and Discussions

Rating Scale Functioning

Category Use Statistics

The category uses statistics for the 5-point rating scale for AAS revealed that the percentage endorsement for response options 1-5 were 0, 4, 17, 32, and 47%, respectively. In determining the functional categories of the scale used for each construct in AAS, three things are seen. First, the shape of the distribution by looking at the number of times of each category; second, the frequency of each category must be at least ten; and third, the mean size is increasing.

In this study, the five-category scale is used for all constructs in the AAS, which is shown in Table 1 and Table 2 Children who can do without the guidance of activities considered to have very high levels of measurement and are indicated by the scale of measurement 5, while children who need guidance in doing activities considered to have very low levels of ability and is characterized by the scale of measurement 1.

Table 1: Levels of measurement, measurement scales and explanations

Levels of measurement	Scales	Explanations
Very Capable	5	Children do not need any guidance
Ability	4	Children can make their own with supervision
Moderate Ability	3	Children do themselves with a little guidance
Ability Low	2	Children do with guidance
Ability Very Low	1	Children need guidance in full

Table 2: Frequency of categories and the mean size of AAS

Label Category	Number of observations	(frequency) Min Size
1	47	-3.62
2	805	-1.55
3	3680	0.18
4	7034	1.60
5	10388	3.26

The results of the analysis carried out showed and supported that the five-category scale of measurement used is appropriate. Selection of the five response categories used by researchers is consistent with the standard criteria discussed by experts because there is no proof of the optimal number of measurement scales to measure a particular construct (Lopez, 1996). Five measurement scales used have been proven empirically to show its appropriateness.

This finding is consistent with results of the analysis performed on all five domains measured in assessment development, the study showed that the five categories used in AAS works as it should be based on the Rasch measurement model. This is because of the shape of the distribution of the five categories of measurement scales for each domain development shows that it is normally distributed with a slight negative skew. Each category also shows up in uniform and everything works as it should.

Threshold Estimates

The threshold value (threshold) or calibration step is another important criterion that needs to be seen in determining the appropriateness of the category for which the measurement scale is used. The calibration step is the estimation of difficulty of selecting a response category compared to other categories. The magnitude of the distance between the threshold value is also important. Distance threshold is to show each step of defining the different positions of the category. This indicates that the estimate is too close or too far away for logit scale used. According to Linacre (1999), the threshold value must exceed at least 1.0 logit (for the five categories of measurement scale) to show the difference in distance between the categories used and not more than 5.0 logit in order to avoid a very large gap between the categories used.

Table 3 is the five category scale for the measurement of AAS. The threshold for the 5 categories showed a uniform decrease. Although it does not increase, the two values still meet the Rasch measurement model, the distance or the range of the threshold between category more than 1.0 logit (1.95, 1.53 and 0.93). This shows the range of categories that can be used for measurement and any category that does not fulfill theaspects of measurement based on the Rasch measurement model.

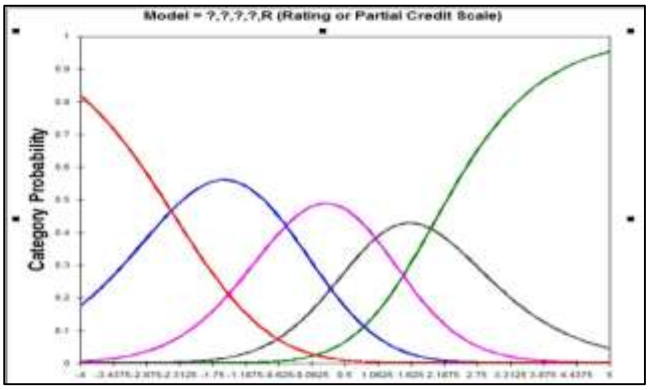
Table 3: The threshold for the five categories of measurement scales AAS

Category level	Threshold value	Range of categories
1	N/A	
2	-2.46	
3	-.51	1.95
4	1.02	1.53
5	1.95	0.93

The analysis of the threshold is supporting the findings by showing that the threshold value for each domain in the development of AAS showed a nearly constant for values less than 1.0 logit. But it still shows the range of categories that are suitable for not more than 5 logit. This shows that the range of categories that can be used for measurement based on Rasch measurement model. When examined the probability curve diagram for each domain it shows each category has a clear peak, and this shows the scale of measurement was functioning as it should for each developmental domain in AAS.

Step Calibrations

Disordering of step calibrations for CB and MON items was present, as Illustrated for the CB dimension in Figure 1 The curves for each response option formed peaks which were the predicted probability of a given rating at each point along the x-axis. Each curve should form a "peak" above the other curves at some point along this continuum. If the curve does not form a peak, it indicates that the response option was not readily or accurately distinguished from the Adjacent options. As illustrated for categories, the curves for ratings of "2," "3," and "4" did not emerge as peaks.



Category Fit Statistics

The OUTFIT mean-square value for each response category was within the accepted bounds, indicating that the response options were not used in a “noisy” or unpredictable manner.

Table 4 Category Statistic

Category	Usage	Average Measure	Expected Measure	Outfit	Threshold
1	0%	0.20	0.08	1.0	
2	4%	0.83	0.69	1.3	-2.46
3	17%	1.41	1.34	1.1	-0.51
4	32%	1.87	1.99	0.7	1.02
5	47%	2.75	2.70	1.0	1.95

The statistical analysis of the consistency of the correspondence also shows that less than 2.0 of the outfit for all AAS development domain. This shows that all these categories, fit or are suitable statistically and meet Many-facet measurement model.

Summary

The results of the Rasch rating scale analysis provided evidence that the 5-point rating scale did not function as intended for AAS. The middle response options were under-utilized and there was limited differentiation between adjacent response categories. Thus, prior to evaluation of item fit, the rating scale categories were adjusted according to the methods described by Linacre (2002). This is described in the next section.

Model Fit

Apart from using the frequency category and increase uniformly in mean size to determine the appropriateness of the scale categories, the compatibility (fit) statistics can also be used (Lopez, 1996, Wright & Masters, 1982). The congruity (fit) statistics provide another criterion for assessing the quality of the measurement scales. According to Linacre (1999), the outfit MNSQs exceeding 2.00 indicates that there is "more misinformation than information", which means that for a specific domain there are many situations that do not fit in the measurement process. Such domains require further investigation and may be empirically appropriate aspects to be combined with adjacent domains. Table 5 shows the outfit MNSQs for each developmental domain.Table 5 Value outfit MNSQ for six developmental domains AAS

Table 5: Value outfit MNSQ for six development domains AAS

	Nilai
Spiritual	1.64
Creativity	1.16
Physical	0.99
Socioemotion	1.03
Cognitive	0.88
Language	0.85

MNSQ the outfit shown in Table 5 above shows the value of around 2:00 for all domains contained in AAS. This shows all the domains or the corresponding statistical fit and meet MFRM.

Separation Reliability

Table 6 gives the overall summary fit statistics for the person (children), rater and domain facets of the model from the Facets computer program. There are several things to note in this summary table. First of all, the person facet is not centered (M= 2.19, SD= .67). The raters facet is centered with a mean of zero (SD= .25), and the domain facets is also centered at zero (SD. 59). The convention in the Many Facet model is to center all of the facets except the one that represents the object of measurement. Table 6 provides reference for interpreting the locations of the persons, raters and domains on the variable map.The summay statistics (Infit and Outfit), are close to the expected values of 1.00 with a standard deviation of .20 indicating fairly good model-d the rating scale structure for these data was examined.

Table 6 overall summary fit statistics

	Persons	Raters	Domains
Measure			
Mean	2.19	0.00	0.00
SD	0.67	0.25	0.59
N	60	2	6
Outfit			
Mean	1.01	1.01	1.09
SD	0.30	0.09	0.27
Infit		s	
Mean	1.02	1.03	1.12
SD	0.28	0.09	0.25
Separation statistic			
Reliability of separation	0.99	1.00	1.00
Chi-square	*4878.5	*655.7	*2737
(df)	59	1	5

*p< .05

Facets

MFRM collects measurement data in parallel and it is implemented using AAS which can isolate the child to a different ability, decisiveness rater, and the difficulty of the domain shown in Table 7.

Children

Table 7 shows the measurement report of children who are assessed as a whole for the six domains of development in a range of sizes AAS children, the RMSE, the index of segregation and the chi-square. Range of children's sizes are between 12:38 (SE = 0.06 children 2) a measure of the ability of children of the lowest and highest, 3:44 (SE = 0.11, 56 children). RMSE values for the measurement of children is 0:08. The separation index was 8.69 which shows these children can be separated or segregated into different capabilities. This is further confirmed by the significant chi-square where the measurement of children is the value of x ^ 2 = 4878.5, p <0.05, df = 59.

Based on analysis,that shown that AAS are used to evaluate performance of children by evaluators were able to separate the children into some extent. The analysis also shows that the difference in the ability of children is statistically significant. The highest separation index is 12:38 and the lowest value is 3:44. This shows that the results of the use of AAS, children can distinguish at least 12 different levels. This is appropriate, such as those suggested by Linacre (2002), that a good instrument will be able to separate students with at least two different levels.

Rater

Table 7 shows the overall rater assessment measurement reports appearing in a range of AAS measure of firmness rater, the RMSE, the index of segregation and the chi-square. Rater firmness measurement range is between -0.25 (SE = 0.01, Rater 1) a measure of the lowest and highest rater severity of -1.19 (SE = 0.01, Rater 2). RMSE values were 0.03 for the measurement domain. The separation index is 20:32 which show domains can be separated or segregated into different difficulty levels. This is further confirmed by the significant chi-square where the measurement domain value is x ^ 2 = 2737, p <0.05, df = 5.

Based on the results of the analysis done on the firmness of the evaluators indicated that each evaluator has a different emphasis and different firmness and this is evidenced by the chi-square test significance which reaffirms that each evaluator has evaluated the different firmness. Assessing varying emphasis is a potential or possibility of the existence of problems in the measurement if using raw scores. This was proven by Banno (2008), which showed that a significant variation between the scores given by different evaluators for the same performance has been reported by several studies. To reduce stress differences between assessors, training sessions can be conducted. However, according to Lunz and Stahl (1990), evaluators can not be trained to achieve the same firmness. This statement was supported by Bonk and Ockey (2003), Lunz et. al. (1990) and Weigle (1994), which shows that the stress evaluator is still different even though they have gone through training courses. The other alternative other than training that can be used is to use MFRM. MFRM estimates the capacity of children for measurements made for candidates is independent of item difficulty and rigor of assessment or in other words, this model provides the measurement of the "independent valuer", "free item" and "independent student" which means that the resulting measurement is not depend on the sample or item or assessors as long as they have the appropriate fit with the Rasch measurement model (Linacre, 1994)

Domain

Table 7 shows the overall rater assessment measurement reports appearing in a range of AAS measure of firmness rater, the RMSE, the index of segregation and the chi-square. Rater firmness measurement range is between -0.25 (SE = 0.01, Rater 1) a measure 0.03 for the measurement domain. The separation index 20:32 show domains can be separated or segregated into different difficulty levels. This is further confirmed by the significant chi- square where the measurement domain value x ^ 2 = 2737, p <0.05, df = 5.

Table 7 Measuremement Facet Summary

Facets	Measure	SE	RMSE	Separation Index	Chi-square P<0.05 Df
Person (children)	0.38-3.44	0.08	0.08	8.69	4878.5
Raters	-0.25-0.25	0.01	0.01	18.08	655.7
Domain	-1.19-0.61	0.03	0.03	20.23	2737

Based on analysis, shown that AAS are used to evaluate performance of children by evaluators. They were able to separate the domain to some extent. The analysis also shows that the difference in the level of difficulty is statistically significant. This shows that the results of the use of AAS, domain can distinguish at least 20 different levels.

CONCLUSION

This study evaluated the psychometric properties of a recently developed,brief measure of Authentic Assessment Standard. Complexities of the data, including repeatedmeasurements and multiple levels of nesting, limited the viability of traditionalpsychometric methods based in Classical Test Theory. Instead,a unique application of the Many-Facet Rasch Model was utilized, providinga flexible approach with a number of strengths.Because the ultimate value of a measure is its ability to predict outcomes, future investigations should evaluate the associations between the authentic performance and documentation system and key outcomes. In the context of the present study, this can be accomplished by utilizing the validated AAS components to predict children achievement. To further facilitate the performance of children, evidence-based practices were changed into community-based settings, future work should also continue to develop and evaluate competency of teacher in nursery and their readiness to have good documentation using measurement tools and model.

The present study illustrates a procedure for evaluating and improving such measures using a highly flexible and sophisticated psychometric approach among children by authentic assessment.

REFERENCES

Banno, E. (2008). Investigating an oral placement test for learners of Japanese as a second language. Thesis Ph.D. Temple University.

Bagnato, S., (2005). The authentic alternative for assessment in early intervention: An emerging evidence-based practice. *Journal of Early Interventipon, 28,* 17-22.

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

Bonk, WJ. & Ockey, GJ. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing 20:*89-110.

Do-Hong Kim & JanDiane Smith (2010). Evaluation of two observational assessment
Systems for Children's Development and Learning, NHSA Dialog: A Research-to- Practice *(Journal for the Early Childhood Field, 13)* (4), 253-267.

Lopez, W. (1996). Communication validity and rating scales. (*Rasch Measurement Transactions*), 10, 482.

Linacre, J.M. (1994). *Many-facet rasch measurement.* Chicago, IL: MESA Press. Linacre, J.M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement 2*(3):266-283.

Linacre, J.M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *(Rasch Measurement)Transactions 16*(2):878

Lunz, M.E.,& Stahl, J.A. (1990). Judge consistency and severity across grading period. *Evaluation and the Health Professions* 13:425-444.

Lunz, M.E., Wright B.D., & Linacre, J.M., (1990). Measuring the impact of judge severity on examination scores. Applied Measurement in Education 3:331-345.

Shonkoff, JP. & Phillips, DA (Eds) (2000). *From neurons to neighbourhoods: The science of early childhood development* (Report of Committee on Integrating the Science of early Childhood Development, Board on Children, Youth, and Families, Commission on Behavioural and Social Science and Education, National Research Council and Institute of Medicine). Washington, DC: National Academies Press.

Snow, C.E., & Van Hemel, S.B. (Eds) (2008). *Early childhood assessment: Why, what, and how* (Report of the Committee on Development Outcomes and Assessments of Young Children, Board on Children and Youth and Families, Board on Testing as Assessment, Division of Behavioral and Social Science and Education, National Research Council) Washington, DC: National Academies Press.

Bagnato, S.J., Neisworth, J.T, & Munson, S.M (1997). *LINKing assessment and early intervention: An authentic curriculum-based approach (*3rd ed). Baltimore:

Brookes.Engelhard, J.G., (1992). The measurement of writing ability with a many-facet Rasch Model. *Applied Measurement in Education,* 5, 171-191.

Linacre, J.M., (1993) *Generalizability yheory and many-facet Rasch measurement.* Paper presented at the annual meeting of the American educational research association.

Bagnato, S.J., (2005). The authentic alternative for assessment in early intervention: An emerging evidence-based practice. *Journal of Early Intervention, 28,* 17-22.

(Neisworth, J.T., & Bagnato, S.J.,(2005). DEC recommended practices: Assessment. In S. Sandall, M.L. Hemmeter, B. J. Smith, & M.E. McLean (Eds.) *DEC recommended practices: A comprehensive guide for practical application in early intervention/*

early childhood special education (pp. 45-69). Longmont, CO: Sopris West. Rasch, (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Smith, E. V., Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *(Journal of Applied Measurement 2,)* 281-311.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *(Journal of Applied Measurement, 1,)* 199-218.

Smith, R. M., Schumacker, R. E., & Bush, J. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *(Journal of Outcome Measurement, 2,)* 66-78.

Elhan, AH. & Atakurt, Y. (2005). Olceklerin degerlendirilmesinde nicin Rasch analizi kullanimali, Ankara Universitesi Tip Fakultesi Mecmuasi, 58, 47-50

Wolfe, E. W. (2000). Equating and item banking with the Rasch model. *(Journal of Applied Measurement,)* 1,409-434.

Wright, BD., & Mok, M. (2000). Rasch models overview. Journal of Applied Measurement, 1, 83-106.

Wright BD. & Linacre,J.M. (1994). Combining and splitting of categories. *Rasch Measurement Transaction 6*(3): 233.

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

Weigle, SC. (1994). Using FACETS to model rater training effects. *Language Testing 15(*2): 263-287.