 Anthropologist, 23(3): 612-619 (2016)

A Research on the Classification Validity of the Decisions Made According to Norm and Criterion-referenced Assessment Approaches[1]

Duygu Gizem Ertoprak[1] and Nuri Dogan[2]

[1]Amasya University, Faculty of Education, Department of Educational Sciences, Amasya, Turkey
[2]Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkey
E-mail: [1]<duygugizemertoprak@gmail.com>, [2]<nurid@hacettepe.edu.tr>

KEYWORDS Classification. Criterion. Decision Validity. Discriminant Analysis. Norm

ABSTRACT In this study, the criterion-referenced assessment and norm-referenced assessment applications were examined comparatively to see whether the decisions made about the students led to any differences in their classification validity, and the results obtained were analyzed. For this purpose, the evaluation results belonging to the students were transferred into the passed/failed decisions in accordance with the rules in criterion and norm-referenced assessment systems, and the obtained findings were analyzed with the discriminant analysis. The study was conducted with 1007 students from six universities who attended courses in the 2011-2012 academic year, and who had previously taken education (pedagogical formation) courses. At the end of the study, the classification validity of the decisions made about the students with the criterion-referenced assessment approach was seen to be higher. Thus, it can be said that making classification decisions about students using the criterion-referenced assessment approach can give more accurate results than the decisions made according to norm-referenced assessment approach.

INTRODUCTION

Validity is defined as the extent to which the tool or method of measurement can measure the variable intended to be measured (Rankin and Vadum 2001). The accuracy of decisions to be made after administering a test is associated with the validity of that test scores. When seen from the perspective of decision-makers, it may be contended that the decisions to be made following an invalid test scores will be ineffective, and when seen from the perspective of individuals, it may be claimed that the decisions to be made will be unfair (Murphy and Davidshofer 2005).

Cronbach (1970), pointing out that a great deal of evidence is needed so as to make decisions as to whether or not a tool of measurement can measure the things that it intends to measure, states that validity should be considered as the process of collecting evidence demonstrating that the test serves to the purpose for which it is used rather than as a single definition or a coefficient. In this case, more than one type of validity is in question. In the APA (1954) system, validity is defined under four headings: content validity, construct validity, predictive validity, and concordance (synchro) validity. Murphy and Davidshofer (2005) suggest that inferences to be made on the basis of test scores could be "inferences associated with the measured property" and "inferences likely to influence the decisions to be made with regard to the individual tested", and that validity should be classified as "measurement validity" and "decision validity".

According to Turgut and Baykul (1992), no matter what psychological construct is measured, mostly decisions are made at the level of classification based on the measurement results. Decisions are made as pass/fail if it is a measure used in education, as accept/refuse if it is a measure used in selecting employees for a workplace,

[*] Designed on the basis of the M.A. Thesis entitled "A Study on the Classification Validity of the Decisions made According to Norm-references and Criterion-referenced Assessment Approaches", which was completed in Hacettepe University Department of Measurement and Evaluation in Education in 2012.

Address for correspondence:
Duygu Gizem Ertoprak
Lecturer
Amasya University, Faculty of Education,
Department of Educational Sciences,
05100 Amasya, Turkey
Telephone: (+90) 358 252 62 30

as positive/negative attitudes if it is a measure used in collecting data on an attitude, as ill/healthy if it is a measure used in diagnosis; and the following orientation of the individuals is made in accordance with those decisions. According to Erkus (2003), such decisions as pass/fail, acceptable/unacceptable, low depression/high depression, and positive attitude/negative attitude about individuals are also examples for decisions at the level of classification. Erkus (2004) believes that "statistically" classification or ordering or both types of decisions can be made based on the measurement results. In deciding about individuals, one should necessarily depend on a criterion.

The criteria employed in evaluating students in education, and the evaluations made based on these may be divided into two main categories. In other words, the form of evaluation differs according to the criterion to be used (Erkus 2006): (1) If the measure is criterion-referenced, the assessment is called a criterion-referenced (that is to say, criterion-based) assessment; (2) if the measure to be used is norm-referenced, the assessment is called a norm-referenced (that is to say, norm-based) assessment.

Erkus (2006) defines a criterion-referenced assessment as the evaluation made according to a minimum standard (cut-off score) specified through various ways so as to be able to consider competency in a field. According to Cohen and Swerdlik (2002), the measures in such evaluations consist of the standards and values that institutions require individuals to have or a series of behaviors that are expected of the candidates. According to Salvia et al. (2012), when one is interested in a student's knowledge about a single fact, one compares a student's performance against an objective and absolute standard (criterion) of performance. Thus, to be considered criterion-referenced, there must be a clear, objective criterion for each of the correct responses to each question or to each portion of the question, if partial credit is to be awarded.

Norm-referenced assessment is a type of evaluation in which student achievement should be examined in comparison with the other members of the group in which the student is included. The grades of each member in the group are determined by making comparisons in terms of areas below the normal distribution curve, percentages, the standard scores such as z and T, the average achievement of the class, and the standard deviation of the class (Atilgan et al. 2009). Sometimes testers are interested in knowing how a student's performance compares to the performances of other students - usually students of similar demographic characteristics (age, gender, grade in school, and so forth). In order to make this type of comparison, a student's score is transformed into a derived score. This type of assessment is called norm-referenced (Salvia et al. 2012).

On checking the regulations for education, teaching and examinations adopted by the Universities in Turkey, it is found that some of them prefer the criterion-referenced assessment while others prefer a norm-referenced assessment. It is also remarkable that criterion-referenced measures are used in primary as well as in secondary education in classifying the students, whereas universities choose the measures to be used in accordance with their decisions. Moreover, it is also noticed occasionally that the same university can use differing measures in different faculties or colleges. It is an expected phenomenon for schools to use different measures in examinations intended for different purposes. Although the same or similar curricula are pursued in institutions of higher education, it is interesting that differing measures are used. In this case, issues such as, which system of assessment should be preferred and most importantly "which of the assessment systems used would be more accurate in the pass/fail decisions to be made for students" come into mind.

When assessment tasks are set for students in universities and colleges, a common practice is to advise them of the criteria that will be used for grading their responses (Sadler 2009). Grading students and owing to this, assessment criteria comparisons were examined in the different countries' education systems. For instance, between 1962 and 1994 a norm-referenced centralized grading system, where individual grades were determined on the basis of a normal distribution, was employed by the Swedish education authorities for the primary purpose of ranking students and managing their entry into higher education. The system was criticized, however, because the grades did not provide much information about students' actual level of knowledge, but focused inordinately on their relative performances within a specified population. Norm-referenced grading was also questioned on the grounds of social responsibility since it was thought to promote

competition rather than collaboration among students. As part of an extensive educational reform during the mid-1990s, the norm-referenced grading system was replaced by the criterion-referenced system that is still in effect today (Redelius and Hay 2012).

Another article seeks to illuminate the gap between UK policy and practice in relation to the use of criteria for allocating grades. It critiques criterion-referenced grading from three perspectives. Twelve lecturers from two universities were asked to 'think aloud' as they graded two written assignments. The study found that assessors made holistic rather than analytical judgments. A high proportion of the tutors did not make use of written criteria in their marking and, where they were used, it was largely a post hoc process in refining, checking or justifying a holistic decision. Norm referencing was also found to be an important part of the grading process despite published criteria. The authors develop the notion of tutors' standards frameworks, influenced by the students' work, and providing the interpretive lens used to decide grades (Bloxam et al. 2011).

When the process is seen as a whole, following an explicit assessment model and taking the steps correctly influence the students' judgment. Furthermore, this is dependent on attaining the conditions of validity. Within this context, this study examines whether or not the criterion-referenced and norm-referenced approaches of assessment produce any differences in the pass/fail decisions regarding the students, and it determines in which type of assessment the decisions yield results more concordant with discriminant analysis.

METHODOLOGY

Type of Research

This study, which is among fundamental research studies, is a piece of comparative research.

Study Group

The study group was composed of 1007 students from six universities in the 2011-2012 academic year, and who had previously taken educational courses (pedagogical formation courses). The universities using the norm-referenced assessment included Dokuz Eylül University and Sakarya University (which meant 404 students), whereas those using the criterion-referenced assessment included Adiyaman University, Ege University, Hacettepe University, and Trakya University (which meant 603 students).

Data Collection

Two different kinds of data were collected in the study. The first was the data coming from students' recorded grades of the psychology course (transcripts), and they were obtained from the Student Affairs Office. The second type of data was the test scores obtained by giving an achievement test to the students.

The data collection tool utilized in the research and taken into consideration in the analyses was the 20-item achievement test, which included sub-tests on developmental psychology and learning psychology. The questions were determined by choosing from 360 questions from the three trial tests of Educational Sciences, which were administered across Turkey and each of which had been applied to 1000 students. The item discrimination index of each question chosen was above 0.70, while item difficulty index was in the 0.60-0.85 range. Having done the preselection for item selection, expert opinion was consulted for the test content. Thus, the data collection tool was established by considering both the expert opinion and the statistical features.

Data Analysis

Firstly, the dependent and the independent variables were determined in the analysis of the data. The averages for scores that students received from the developmental psychology and the learning psychology sub-tests were transformed into the hundred-pointed grade system, and the test score obtained was labeled as the variable score. Students whose scores were above the standard score according to the criterion-referenced and the norm-referenced approaches were regarded as successful/pass, and those whose scores were below the standard score were regarded as unsuccessful/fail, and a new variable, 'group' was added. The *dependent variable* of the model was the group variable showing the pass/fail status of the students. The successful/unsuccessful or the pass/fail classification for the students here was made

by regarding criterion-referenced value as 60 and by calculating the norm-referenced value by considering the arithmetic means and the standard deviations of each group. The universities that the students attended, the type of study, their departments, gender, passing grades (transcripts) for the educational psychology course, test scores for the developmental psychology test and for the learning psychology test were the *independent variables* of the model. Discriminant analysis was employed so as to research the relations between the dependent and the independent variables. The reason for using discriminant analysis was to determine the fit between the pass/fail decisions made for students in accordance with the test scores and the classification decisions made by considering the education course grades. In other words, the percentages for the decisions made for students according to criterion-referenced and norm referenced measures were determined in this way.

Prior to the analyses performed, it was checked whether or not the data met the presuppositions of the discriminant analysis, it was concluded that the data was suitable for analyzing via this method of analysis.

RESULTS

The classification decisions, which were obtained by assessing the students through their own system of assessment and through criterion-referenced assessment system for 404 students assessed through norm-referenced measure, and the results for classification obtained through discriminant analysis are shown in Table 1.

On examining the results for classification shown in Table 1, it becomes clear that 348 (90.4%) of the 385 students who were to pass and 19 (100%) out of 19 students who were to fail, and all of whom were assessed through norm-referenced assessment were classified "accurately" when assessed through their own system of assessment. 37 out of 385 students who had been decided to pass were classified incorrectly. According to discriminant analysis, although fail decision should be made for those students who had been classified incorrectly, pass decision was made for them in the norm-referenced measure. The total percentage of accurate classification for the discriminant function was 90.8 percent.

When the students who had been assessed through norm-referenced assessment were assessed through the criterion-referenced assessment, 213 (98.6%) of the 216 students for whom pass decision was made and 188 (100%) of the 188 students for whom fail decision was made were classified "accurately". 3 out of 213 students for whom pass decision was made were classified incorrectly. According to discriminant analysis, although fail decision should be made for those students who had been classified incorrectly, a pass decision was made for them in the criterion-referenced measure. The total percentage of accurate classification for the discriminant function was 90.3 percent.

According to these results, it might be said that making decisions for the 404 students on the basis of criterion-referenced assessment seems to produce more accurate results.

The classification decisions which were obtained by assessing the students through their own system of assessment and through norm-referenced assessment system for 603 students assessed through criterion-referenced measure, and the results for classification obtained through discriminant analysis are shown in Table 2.

Table 1: The results for classification obtained through norm-referenced and criterion-referenced assessment systems for students assessed through norm-referenced assessment

		Pass		*Fail*		*Total*	
	Actual Status	*f*	*%*	*f*	*%*	*f*	*%*
Norm-referenced	Pass	348	90.40	37	9.60	385	100.00
	Fail	0	0	19	100.00	19	100.00
	Total percentage of accurate classification =90.8 percent						
Criterion-referenced	Pass	213	98.60	3	1.40	216	100.00
	Fail	0	0	188	100.00	188	100.00
	Total percentage of accurate classification = 99.30 %						

Table 2: The results for classification obtained through norm-referenced and criterion-referenced assessment systems for students assessed through criterion-referenced assessment

		Pass		*Fail*		*Total*	
	Actual Status	*f*	*%*	*f*	*%*	*f*	*%*
Norm-referenced	Pass	536	94.00	34	6.00	570	100.00
	Fail	0	0	33	100.00	33	100.00
	Total percentage of accurate classification = 94.4 percent						
Criterion-referenced	Pass	238	95.60	11	4.40	249	100.00
	Fail	8	2.30	346	97.70	354	100.00
	Total percentage of accurate classification = 96.8 percent						

When the students who had been assessed through the criterion-referenced assessment were assessed through their own system of assessment, 238 (95.6%) of the 249 students who were decided to pass and 346 (97.7%) of the 354 students who were decided to fail were classified "accurately". 11 out of 249 students for whom a pass decision was made and 8 out of 354 students for whom a fail decision was made were classified incorrectly. The discriminant analysis shows that the 11 students for whom a pass decision was made should be marked to fail, and the 8 students for whom a fail decision was made should be marked to pass. The total percentage of accurate classification for the discriminant function was 96.8 percent (Table 2).

When the students who had been assessed through the criterion-referenced assessment were assessed through the norm-referenced assessment, 536 (94%) of the 570 students for whom a pass decision was made and 33 (100%) of the 33 students for whom a fail decision was made were classified "accurately". 34 out of 574 students for whom pass decision was made were classified incorrectly, while none of the 33 students for whom the fail decision was made were classified incorrectly. Discriminant analysis demonstrates that 34 of the students who have been classified as successful according to criterion-referenced assessment are classified incorrectly according to norm-referenced assessment. The total percentage of accurate classification for the discriminant function was 94.4 percent.

According to these results, it might be said that making decisions for the 603 students on the basis of criterion-referenced assessment, which was their own system of assessment, seems to produce more accurate results.

The classification decisions made for all of the 1007 students constituting the study group through norm-referenced and criterion-referenced assessment systems, respectively, and the results for classification obtained through discriminant analysis are shown in Table 3.

When the students included in the study group were assessed through norm-referenced assessment, 891 (94.9%) of the 939 students who were decided to pass and 68 (100%) of the 68 students who were decided to fail were classified "accurately". Discriminant analysis shows that 48 students who have been evaluated as successful are incorrectly classified. The total percentage of accurate classification for the discriminant function was 95.3 percent (Table 3).

Table 3: The results for classification obtained through norm-referenced and criterion-referenced assessment systems for students included in the study group

		Pass		*Fail*		*Total*	
	Actual Status	*f*	*%*	*f*	*%*	*f*	*%*
Norm-referenced	Pass	891	94.90	48	5.10	939	100.00
	Fail	0	0	68	100.00	68	100.00
	Total percentage of accurate classification = 95.2 percent						
Criterion-referenced	Pass	450	96.80	15	3.20	465	100.00
	Fail	0	0	542	100.00	542	100.00
	Total percentage of accurate classification = 98.5 percent						

When the students in the study group were assessed through the criterion-based assessment, 450 (96.8%) of the 465 students for whom the pass decision was made and 542 (100%) of the 542 students for whom fail decision was made were classified "accurately". Yet, 15 out of 465 students for whom pass decision was made were classified incorrectly. Discriminant analysis shows that 15 students who have been evaluated as successful are incorrectly classified. The total percentage of accurate classification for the discriminant function was 98.5 percent.

Accordingly, when all of the students were assessed through the norm-referenced assessment, they were classified accurately at the rate of 95.2 percent, whereas they were classified accurately at the rate of 98.5 percent on being assessed through the criterion-referenced assessment. Thus, it may be said that making decisions for the 1007 students included in the study group on the basis of criterion-referenced assessment seems to produce more accurate results.

DISCUSSION

Many international testing and assessment systems today operate under uncompromising and tight timelines with high stakes for particular stakeholders. Such forces add stress to the assessment design and validation efforts, as well as to the preparation of test users at large for appropriate and meaningful information use. When the consequences of assessment results have high stakes decisions for individuals, programs or institutions (Proctor and Silverman 2011), validity issues must become a more important part of the public matters.

Measurement theorists have long understood that validity depends not only on how well tests and assessments are designed and validated, but also on how defensibly the resulting information is put to use in applied and policy settings (Cronbach 1971; Messick 1989; Kane 2006; Kane 2013). Test data uses today are more complex and multilevel than before. Interpretations and decisions made with test scores involve statistical procedures yielding results that involve several steps (Lissitz 2009). It must be decided how educational assessment works and which assessment criterion must be used in one of these steps. Because, according to the answers, students' judgments may vary. From this point of view, validity and assessment system issues can't think separately and studies about them must be examined.

Determining how to make the classification accuracy of a measurement process has come up as a sub-problem in the educational, medical and especially psychiatric studies prepared for examining any measurement tools' reliability and validity (Rosenberg et al. 1998; Kan 2004; Guzeller 2005; Kelecioglu and Guzeller 2006; Gulec et al. 2007; Irmak et al. 2007; Yaprak 2007; Buyukozturk and Bokeoglu 2008; Gunuc 2009; Lubans et al. 2011; Atar 2012; Koutsouleris et al. 2012; Park et al. 2012; Janssen et al. 2013; ul Hassan et al. 2013). In such studies, some statistical methods were utilized to determine the classification accuracy or classification validity of measurement tools, and the results are interpreted according to these statistical outcomes. Gunuc (2009) aimed to develop a Turkish Internet Addiction Scale in his work. A Likert-type scale consisting of 35 items was used in this study. Internet attitude levels of students were classified as low, medium and high. However, the "Two Step Cluster Analysis" technique was applied for more detailed results about the individuals' attitude levels. As a result, the sample was separated into the four groups, namely, "addicted group", "addiction risk group" "threshold group" and "non-addicted group" rather than three groups. In addition to this, 76 (10.1%) of 754 individuals were addicted to the Internet and 199 (26.4%) individuals were addiction risk group. The remaining 222 (29.4 %) individuals were in the threshold group and 257 (34.1 %) individuals were in the non-addicted group with a total of 479 (63.5 %) not being classified as addicted.

In literature reviews, studies comparing the systems of norm and criterion-referenced assessment from certain perspectives were also found. These perspectives were mostly about the similarities and differences of the type of assessment and the areas of usage (Nartgun 2007; Toprakci et al. 2007; Astin 2012; Salvia et al. 2012; Rust and Golombok 2014). Nartgun (2007), for instance, analyzed comparatively to see whether or not criterion-referenced assessment and norm-referenced assessment brought about any differences in students' grades, which were the indicators of their levels of achievement. Considering the research findings as a whole, they were interpreted as the marks given on the basis of norm-referenced assessment moved away from representing the students' levels of achievement,

in contrast to the marks given on the basis of criterion-referenced assessment. According to Astin (2012), norm-referenced tests, such the GRE and SAT, are well-suited to the competitive value framework underlying the resources and reputational conceptions of excellence because they can be used in selection and screening and lend themselves readily to competitive comparisons. But they are ill suited to the talent development approach because they make it difficult to measure growth or change over time. Criterion-referenced tests, on the other hand, not only make it possible to establish absolute standards of performance but also allow researchers to assess how much students actually change with time. In short, reliance on norm-referenced tests promotes the values of selection and competition, whereas reliance on criterion-referenced tests promotes the values of teaching and learning.

Contrary to the above studies, any study, which investigates the validity of classification decisions taken about individuals according to assessment criteria couldn't find. In fact, it is emerged doing studies about how to class the students according to norm and criterion-referenced assessment because every classification decision was given based on the assessment criteria.

In this context, the goal of this study was to examine classification validity (a special type of validity) of the decisions made in relation to the same group of students assessed according to norm and criterion-referenced assessment approaches. From this examination's findings, it may be concluded that making decisions for classification of students on the basis of criterion-referenced assessment is more valid in placing the students into right groups (classes). It may be wrong to make recommendations as to what type of measure should be used in assessment by depending on only one research study. Yet, in deciding on the preference for the system of assessment especially in institutions of higher education, it may be recommended that the results of this current study and of such studies should be taken into consideration.

CONCLUSION

In consequence of transforming the same scores into grades according to both approaches of assessment, the following could be said:

1. When the students who have been assessed through norm-referenced assessment were assessed through their own system of assessment, they were classified accurately at the rate of 90.8 percent, while they were classified accurately at the rate of 99.3 percent on being assessed through the criterion-referenced assessment. Thus, it may be said that making decisions through criterion-referenced assessment for the students who are assessed through norm-referenced assessment would produce more accurate results.

2. When the students who had been assessed through criterion-referenced assessment were assessed through their own system of assessment, they were classified accurately at the rate of 96.8 percent, while they were classified accurately at the rate of 94.4 percent on being assessed through norm-referenced assessment. Thus, it may be said that making decisions for the students who are assessed through criterion-referenced assessment through their own system of assessment produces more accurate results.

3. When all of the participants were assessed through the norm-referenced assessment, they were classified accurately at the rate of 95.2 percent while they were classified accurately at the rate of 98.5 percent on being assessed through criterion-referenced assessment. Thus, it may be said that making decisions for the 1007 students included in the study group on the basis of criterion-referenced assessment seems to produce more accurate results.

RECOMMENDATIONS

In conclusion, since the decisions made in accordance with norm-referenced and criterion-referenced assessment are geared to the future of individuals, it is suggested that such comparisons should regularly be made between types of measures, the probable reasons for high validity of classifications made on the basis of decisions through criterion-referenced assessment should be established, and such points should be considered in similar research studies. Besides, it is also recommended that different statistical methods should be employed in classifying the individuals in the forthcoming research studies.

REFERENCES

Astin AW 2012. *Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation in Higher Education*. Lanham: Rowman and Littlefield Publishers.

Atar HY 2012. Egitim fakultelerinde uygulanan ozel yetenek sinavlarinin siniflama dogrulugu uzerine bir calisma. *Egitim ve Bilim*, 37(163): 268-282.

Atilgan H, Kan A, Dogan N 2009. *Egitimde Olcme ve Degerlendirme*. Ankara: Ani Publications.
Bloxham S, Boyd P, Orr S 2011. Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education*, 36(6): 655-670.
Buyukozturk S, Cokluk Bokeoglu O 2008. Diskriminant fonksiyon analizi: Kavram ve uygulama. *Eurasian Journal of Educational Research*, 33: 73-92.
Cohen RJ, Swerdlik ME 2004. *Psychological Testing and Assessment: An Introduction to Tests and Measurement.* New York: McGraw-Hill Book Company.
Cronbach LJ 1970. *Essentials of Psychological Testing.* New York: Harper and Row Publications.
Cronbach LJ 1971. Test validation. In: RL Thorndike (Ed.): *Educational Measurement*. Washington: American Council on Education, pp. 443–507.
Erkus A 2003. *Psikometri Uzerine Yazilar.* Ankara: Turkish Psychological Association Publications.
Erkus A 2004. The proposal of a new conceptualization for validity and criterion-referenced assessment. *Eurasian Journal of Educational Research*, 16: 113-117.
Erkus A 2006. *Sinif Ogretmenleri icin Olcme ve Degerlendirme: Kavramlar ve Uygulamalar.* Ankara: Ekinoks Publications.
Gulec H, Gulec MY, Kucukali CI 2007. Eriskin dikkat eksikligi hiperaktivite bozuklugu tanisi konmus erkek mahkumlarda Iowa kumar testi Turkce uyarlamasinin psikometrik ozellikleri. *Turkiye'de Psikiyatri*, 9(2): 91-97.
Gunuc S 2009. *Internet Bagimlilik Olceginin Gelistirilmesi ve Bazi Demografik Degiskenlerle Internet Bagimliligi Arasindaki liskilerin Incelenmesi.* Master Thesis, Unpublished. Van: Yuzuncu Yil University.
Guzeller CO 2005. *Ortaogretim Kurumlari Ogrenci Secme ve Yerlestirme Sinavinin Gecerligi.* PhD Thesis, Unpublished. Ankara: Hacettepe University.
Guzeller CO, Kelecioglu H 2006. Ortaogretim kurumlari ogrenci secme sinavinin siniflama gecerligi uzerine bir calisma. *Hacettepe University Journal of Education*, 30: 140-148.
Irmak T, Sutcu Y, Tekinsav S, Aydin A, Sorias O 2007. Otizm davranis kontrol listesinin (abc) gecerlik ve guvenirliginin incelenmesi. *Cocuk ve Genclik Ruh Sagligi Dergisi*, 14(1): 13-23.
Janssen X, Cliff DP, Reilly JJ, Hinkley T, Jones RA 2013. Predictive validity and classification accuracy of actigraph energy expenditure equations and cut-points in young children. *PLoS One*, 8(11): e79124.
Kan A 2004. OSS'nin siniflama gecerligi uzerine bir calisma. *Inonu University Journal of Education*, 5(8): 51-60.
Kane MT 2006. Validation. In: R Brennan (Ed.): *Educational Measurement.* Westport: American Council on Education and Praeger, pp. 17–64.
Kane MT 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1): 1-73.
Koutsouleris N, Davatzikos C, Bottlender R, Patschurek-Kliche K, Scheuerecker J, Decker P, Meisenzahl EM 2012. Early recognition and disease prediction in the at-risk mental states for psychosis using neurocognitive pattern classification. *Schizophrenia Bulletin*, 38(6): 1200-1215.
Lissitz RW 2009. *The Concept of Validity: Revisions, New Directions, and Applications*. NC: Information Age Publishing Inc.
Lubans DR, Hesketh K, Cliff DP, Barnett LM, Salmon J, Dollman J, Hardy LL 2011. A systematic review of the validity and reliability of sedentary behaviour measures used with children and adolescents. *Obesity Reviews*, 12(10): 781-799.
Messick S 1989. Validity. In: RL Linn (Ed.): *Educational Measurement.* New York: American Council on Education and Macmillan, pp. 13-103.
Murphy RK, Davidshofer OC 2005. *Psychological Testing: Principles and Application.* New Jersey: Prentice-Hall Book Company.
Nartgun Z 2007. Ayni puanlar uzerinden yapilan mutlak ve bagil degerlendirme uygulamalarinin notlarda farklilik olusturup olusturmadigina iliskin bir inceleme. *Ege University Journal of Education*, 8(1): 19-40.
Park YS, Abramson DM, Levin K 2012. Assessing the Reliability and Validity of the Evaluation Support Decision Tool. *Columbia University Academic Commons EDST Report No. 031113,* Columbia: National Center for Disaster Preparedness.
Proctor CP, Silverman RD 2011. Confounds in assessing the associations between biliteracy and English language proficiency. *Educational Researcher*, 40(2): 62-64.
Rankin NO, Vadum AC 2001. *Psychological Research: Methods for Discovery and Validation.* Boston: Mc Graw Hill Book Company.
Redelius K, Hay PJ 2012. Student views on criterion-referenced assessment and grading in Swedish physical education. *Physical Education and Sport Pedagogy*, 17(2): 211-225.
Rosenberg SD, Drake RE, Wolford GL, Mueser KT, Oxman TE, Vidaver RM, Luckoor R 1998. Dartmouth assessment of lifestyle instrument (DALI): A substance use disorder screen for people with severe mental illness. *The American Journal of Psychiatry*, 155(2): 232-238.
Rust J, Golombok S 2014. *Modern Psychometrics: The Science of Psychological Assessment*. Oxford: Routledge.
Sadler DR 2009. Indeterminacy in the use of preset criteria for assessment and grading. *Assessment and Evaluation in Higher Education*, 34(2): 159-179.
Salvia J, Ysseldyke J, Bolt S 2012. *Assessment: In Special and Inclusive Education*. Michigan: Cengage Learning Books.
Toprakci E, Baydemir G, Kocak A, Akkus O 2007. Egitim Fakultelerinin Egitim-Ogretim Ve Sinav Yonetmeliklerinin Karsilastirilmasi. *Paper presented in Congress on 16th National Educational Sciences* in Gaziosmanpasa University, Tokat, September 5 to 7, 2007.
Turgut MF, Baykul Y 1992. *Olcekleme Teknikleri.* Ankara: OSYM Publications.
ul Hassan E, Shahzeb F, Shaheen M, Abbas Q, Hameed Z 2013. Measuring validity of determinants of individual investor decision making investing in Islamabad stock exchange of Pakistan. *Middle-East Journal of Scientific Research*, 14(10): 1314-1319.
Yaprak B 2007. *lkogretim Ogrencilerinin Algiladiklari Anne-Baba Tutumunun Diskriminant Analiziyle Belirlenmesi ve Benlik Saygisi ile Olan liskisinin Degerlendirilmesi Uzerine Bir Uygulama.* Master Thesis, Unpublished. Eskisehir: Eskisehir Osmangazi University.

Paper received for publication on December 2014
Paper accepted for publication on March 2016