International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 470–491
https://doi.org/10.21449/ijate.974214
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Reliability of Ratings of Multidimensional Fluency Scale with Many-Facet Rasch Model

Cigdem Akin Arikan[1,*], **Pinar Kanik Uysal**[2], **Huzeyfe Bilge**[1], **Kasim Yildirim**[2]

[1]Ordu University, Faculty of Education, Department of Measurement and Evaluation in Education, Türkiye
[2]Ordu University, Faculty of Education, Turkish Language Teaching Department, Türkiye
[3]Kafkas University, Faculty of Education, Turkish Language Teaching Department, Türkiye
[4]Muğla Sıtkı Koçman University, Faculty of Education, Primary School Teaching Department, Türkiye

ARTICLE HISTORY

Received: July 25, 2021
Revised: Feb. 04, 2022
Accepted: Apr. 23, 2022

Keywords:
Rater bias,
Rubric,
Reading fluency,
Prosody,
Many-facet Rasch model,
Rater reliability

Abstract: The aim of this study was to determine whether the reliability of raters was provided by assessing reading prosody using the Multidimensional Fluency Scale (MDFS). The study was completed with a cross-sectional design, and in line with this, the prosodic reading skills of 41 fifth-grade students were rated by elementary school classroom teachers and Turkish language arts teachers using the MDFS. Data obtained from the ratings were analyzed with the many-facet Rasch model (MFRM). When the findings are investigated, the reading prosody rubric used in the research served the purposes of the reading prosody criteria, the sub-dimensions of the rubric could be reliably differentiated, the determined criteria were reliable, and the criteria categories appear to be adequate. Additionally, the severity and leniency of raters were found to differ, and Turkish language arts teachers were found to perform more severe ratings than classroom teachers. It was found that raters were ranked reliably in terms of severity/leniency, and that their levels of severity/leniency differed from each other. Another result obtained is that the prosody criterion that students completed with the most difficulty was phrasing. Therefore, it was concluded that the MDFS is a reliable rubric and that researchers and teachers can reliably use it to assess prosodic reading skills.

1. INTRODUCTION

The interest shown in reading fluency has increased in recent years. One of the most important reasons for this is the understanding of the significant correlation between reading fluency and academic success (Baştuğ & Keskin, 2012; Buck & Torgesen, 2003; Hallman, 2009; National Reading Panel, 2000; Rasinski, 2004; Yıldız et al., 2019). While initially, it was common to deal with speed and accuracy in reading fluency explained with the automaticity theory, in recent years, the definition of fluency has expanded and begun to include different concepts (Godde et al., 2019). According to the accepted view, reading fluency comprises speed, accuracy, and prosody, with researchers (Godde et al., 2019; National Reading Panel, 2000;

*CONTACT: Cigdem Akin-Arikan ✉ akincgdm@gmail.com Ordu University, Faculty of Education, Department of Measurement and Evaluation in Education, Türkiye

e-ISSN: 2148-7456 /

Rasinski, 2004, 2010; Schwanenflugel et al., 2004; Ulusoy et al., 2011) stating the need to deal with these three elements when defining reading fluency (Benjamin & Schwanenflugel, 2010).

There are differences at the point of measurement and assessment when dealing with these three essential elements of reading fluency, with prosody being accepted as relatively more challenging to assess compared to speed and accuracy (Baştuğ, 2021; Valencia et al., 2010). Rubrics in which rater judgments come into play are used in the evaluation of prosody. This situation involves a variety of difficulties. Perhaps the most important of these is the degree to which raters are consistent in giving scores, the degree to which they are severe or lenient when giving scores, and the purpose served by the rubric used. The Multidimensional Fluency Scale (MDFS) (Zutell & Rasinski, 1991), commonly used in Turkey to assess prosodic reading, appears to have been examined for consistency between several raters to show reliability in general (Ceyhan, 2019; Kanık Uysal & Duman, 2020). However, it is accepted that this type of assessment is deficient in many aspects (Eckes, 2015), and it is recommended to note the severity of the raters (Bond & Fox, 2015). Though more comprehensive reliability studies were performed for the original version of the MDFS (Moser et al., 2014; Smith & Paige, 2019), these types of comprehensive analysis are not encountered for the version adapted to Turkish.

1.1. Prosody

Prosody is a comprehensive and well-established term used since very ancient times (Couper-Kuhlen, 1986; Crystal, 2008; Sinambela, 2017; Spafford et al., 1998; VandenBos, 2015; Xu & Liu, 2012). In literature, prosody is defined as the melody of a language (Sinambela, 2017), the flow of rhythm involving intonation, words and sentence length, and the stress patterns of a language (Spafford et al., 1998). It is a phonological feature of speaking related to a phoneme sequence, like stress, intensity or duration rather than a single section (VandenBos, 2015) and is a term used to express variations in pitch, loudness, tempo and rhythm in suprasegmental phonetics and phonology (Crystal, 2008). It is defined as the ability of readers to use phrasing and expression appropriately (Rasinski, 2004) and is a general language term describing rhythmic and tonal features of speech (Dowhower, 1991). While some of these definitions consider prosody as a speech term, some attribute it to a dimension of reading aloud. In addition to speech, prosody has been mentioned frequently in relation to reading skills in recent years and prosodic reading, also called expressive reading, means reading using the voice well with appropriate phrasing and reflecting the emotions in the text (Rasinski, 2004). Prosody includes elements like intonation, rhythm involving syllable-word-sentence length, stress, pitch, and tempo (Crystal, 2008; Dowhower, 1991; Palmer, 2010; Rasinski, 2004; Spafford et al., 1998; VandenBos, 2015; Xu & Liu, 2012).

The importance given by people to prosody, or their attempts to understand based on prosody, begin in infancy. A child develops a sensitivity to the mother tongue and prosodic features used by the mother, and these prosodic features play an essential role in early reading development (Godde et al., 2019). Research shows that even infants younger than 12 months use prosody as a primary clue to syntactic structures and that their babbling contains prosodic features (Kuhn & Stahl, 2013). Based on these results, it is possible to say that prosody is one of the most important elements affecting understanding, even from very young ages. People's contact with prosody begins when they are young and has an effect on their comprehension skills in future periods (Godde et al., 2019; Kuhn & Stahl, 2013). This relationship between prosody and comprehension, between understanding what is read and prosody, is frequently revealed and widely accepted (Çetinkaya et al., 2016; Godde et al., 2019; Schwanenflugel et al., 2004). Prosody has an important place in determining reading competence and identifying reading fluency (Keskin, 2012; Schreiber, 1991). However, the multiple dimensions of prosody and rubrics for measurement require great care in the measurement and assessment process.

Assessment of prosody is more difficult compared to measuring speed and accuracy (Grosjean & Collins, 1979; Moser et al., 2014; Valencia et al., 2010). Prosody involves reading by paying attention to many elements like intonation, stress, pauses, syntax and semantic groups, causing prosody to be the most difficult variable to measure among reading skills (Godde et al., 2019). After many years of measuring speed and accuracy, it was emphasized that measurement of fluency without prosody was not sufficient (Dowhower, 1991; Kuhn, 2007; Schreiber, 1991). After awareness of this deficiency, rubrics were developed to measure prosody. The most common among these are the MDFS (Zutell & Rasinski, 1991) and the Oral Reading Fluency Scale (U.S. Department of Education, 2002), with these two rubrics accepted as being the most commonly used rubrics to assess prosody (Morrison & Wilcox, 2020; Smith & Paige, 2019).

After Allington (1983), Zutell and Rasinski (1991) were the first to develop rubrics to assess prosody. Prepared as a task-specific rubric (Brookhart, 2013), in the MDFS the researchers dealt with three dimensions of phrasing, smoothness and pace, with each dimension organized on four levels. This rubric was updated by Rasinski (2004), and expression and volume was added to bring it to four dimensions. This update was completed to allow a separate assessment of the four basic features included in prosody. This multidimensional rubric comprises the dimensions of 'expression and volume', 'phrasing', 'smoothness' and 'pace'. Statements for each point are included in the rubric and raters perform the assessment in line with these statements. It was shown to be among the best scales to assess prosody (Benjamin et al., 2013) in many studies using this rubric (Aşıkcan, 2019; Morrison & Wilcox, 2020; Overstreet, 2014; Rasinski et al., 2017; Young & Rasinski, 2009).

In spite of the common use of rubrics to assess prosody, a variety of problems are encountered in using rubrics. Zutell and Rasinski (1991) stated that they developed the MDFS for use in class. In other words, the rubric is oriented toward in-class application. Additionally, these rubrics require judgments mediated by the assessor, and the professional experience of those evaluating the reading may affect the scores obtained from the rubric (Moser et al., 2014). To overcome these problems, it is recommended to provide training or directions to raters who will use the rubric (Zutell & Rasinski, 1991). Based on these views, the use of rubrics to assess prosody involves more than just listening to oral reading and giving scores, and requires many precautions to be taken in relation to reliability.

Studies show that interrater agreement is not sufficiently high in assessments made without training (Godde et al., 2017; Haskins & Aleccia, 2014). Though agreement rates were significant in these studies, the significance was not high enough to ensure use in a common fashion. Among studies using the MDFS, studies are encountered where the prosody assessment was excluded due to the lack of statistical agreement between two raters (Bilge, 2019); where there were high levels of agreement between scores given by two experts (Ceyhan, 2019; Kanık Uysal & Duman, 2020; Overstreet, 2014; Paige et al., 2021); and where assessments were made by a single rater without examining interrater reliability (Aşıkcan, 2019; Esmer, 2019; Kaya Tosun, 2019; Kızıltaş, 2019, Rasinski et al., 2017; Zimmerman et al., 2019). Based on these differences in the relevant literature, it is understood that there are different forms used when assessing prosody despite widespread use, and a need to investigate the interrater reliability related to the use of rubrics.

1.2. Rater Reliability

Most measurement processes in behavioral science involve errors; however, this problem is observed more frequently when measurement is made by raters (Shrout & Fleiss, 1979) and the fallibility of human raters has led to serious concerns related to the psychometric quality of scores given to those entering exams (Eckes, 2015). Many studies revealed that in situations where it is not possible to perform automatic rating for assessment of the performance of individuals, evaluation of student responses by several raters would ensure that more

definite/accurate results are obtained. However, in this situation, the scores of an individual are not just linked to their performance or the difficulty of the task, they are also related to rater behavior, or in other words, errors due to the raters. Personal bias error, one of the rater errors, occurs in three different ways. Raters may have a tendency to give lower scores than deserved by the student's performance in severity error (excessive negativity error), they may tend to give higher scores in the leniency error (excessive positivity error) or they may avoid giving low or high scores and give moderate scores, called the central tendency error (McMillan, 2017). If the scores given to the same individual are similar during assessment by several raters, it means reliability is provided or sufficient between raters. However, this situation is not always present in practice. As raters generally comprise an important source of variance, they threaten the validity of inferences made from the results (Eckes, 2015). For this reason, it is important to investigate the effect of raters on the assessment of the performance of individuals. The effects of raters can be identified by two general methods: generalizability theory (G-Theory) and the many-facet Rasch model (MFRM). However, there are some differences between the two approaches. G-theory provides information at the group level, while the MFRM provides individual-level information about all the variability sources (Barkaoui, 2008; Linacre, 1993). The MFRM, one of the item response theory (IRT) models, includes assessments of the effects of other possible sources of systematic error (raters, ratings, tasks, and items) (Sudweeks et al., 2004). The MFRM, from a micro perspective, has the advantages of simultaneously assessing the difficulty of test items, a student's ability, the severity/leniency of raters, and the consistency of scores on the same scale (Li et al., 2021). In addition, the MFRM exceeds G-Theory by presenting quality control fit statistics as well as calculating a measure and a standard error for each source (Linacre, 1993). Using the MFRM, each facet's contribution is analyzed independently of the other facets (Engelhard & Myford, 2003), which allows it to make more accurate estimates than scores obtained with G-Theory. It can be said that the MFRM should be preferred primarily in assessments where it is not possible to score objectively (İlhan, 2015).

As discussed earlier, prosody is a critical component of reading fluency. Teachers and researchers evaluate prosodic skills frequently. While the MDFS was prepared for in-class usage (Zutell & Rasinski, 1991), it is common to use this rubric in scientific research. Since the MDFS depends on human rating, and thus, errors are expected to occur, the question of whether there are significant differences between raters' ratings is important because decisions about prosodic skills are made depending on these assessments. The reason why the MFRM is used for this scale is that it is a stronger psychometric model than classical test theory (Haiyang, 2010) in terms of its ability to detect interactions between different error sources, and is recommended by researchers (Baird et al., 2013) to avoid the limitations of classical approaches. In the literature, there are studies that tested the interrater reliability of assessments performed using prosody rubrics. Moser et al., (2014) examined the interrater reliability of Zutell and Rasinski's (1991) MDFS, and found that the rubric was reliable. In the study by Smith and Paige (2019), it was determined that the MDFS was more reliable compared to the Oral Reading Fluency Scale (U.S. Department of Education, 2002). However, there is no study encountered in the literature revealing how reliability between multiple raters is provided for the version of the MDFS adapted to Turkish. As the use of the reading prosody rubric occurs in the Turkish lesson, the MDFS developed by Zutell and Rasinski (1991), updated by Rasinski (2004) and adapted to Turkish by Yıldız et al., (2009) appears to have been used by researchers in primary school teaching (Aşıkcan, 2019; Ceyhan, 2019; Kaya Tosun, 2019) and Turkish education in secondary schools (Armut & Türkyılmaz, 2017; Kanık Uysal & Duman, 2020). Based on this, in this study the aim was to identify the degree to which interrater reliability was provided for assessment of reading prosody by elementary school classroom teachers and Turkish language arts teachers using the MDFS. In line with this aim, answers to the following questions were sought:

1) Do teachers display differences in terms of severity/leniency when assessing students' prosodic reading?
2) Are there differences in terms of severity/leniency during assessment of students' prosodic reading according to teaching branch?
3) What are the results for task/criterion difficulty analysis related to students' prosodic reading?
4) What are the outcomes of central tendency behavior and bias analysis of raters?

2. METHOD

This research aimed to identify the reliability of results obtained from different raters using the MDFS (Zutell & Rasinski, 1991). In line with the aim of the research, the assessment results for prosodic reading of students by different raters were investigated with the MFRM. Descriptive research, which is a type of quantitative design, was used in the study (Fraenkel & Wallen, 2009).

2.1. Participants

2.1.1. *Students*

Prosodic reading data were obtained from fifth-grade students in a state middle school located in the central district of a metropolitan city in Turkey. The criterion sampling method was used for the selection of students included in the research. The reason for determining the criterion as fifth-grade level was that this class level is known by both elementary school classroom teachers and Turkish language arts teachers. Until the 2012-2013 academic year in Turkey, elementary school classroom teachers continued teaching until fifth grade, and the reading skills of students at this grade level were assessed by elementary school classroom teachers. However, since the 2012-2013 academic year, a 4+4+4 educational system has been implemented, and the fifth grade was moved to the middle school level. For this reason, the reading skills of students in fifth grade are currently assessed by Turkish language arts teachers. To find answers to one of the problems in this research, namely "Are there differences in terms of severity/leniency during assessment of students' prosodic reading according to teaching branch?" the study included both elementary school classroom and Turkish language arts teachers. Comparisons were made between branches in assessing the oral reading prosody of students in fifth grade, known by both departments. In order to meet this criterion, in other words, to include teachers who had taught at fifth-grade level, teachers with at least ten years of teaching experience were included in the study. All fifth-grade students attending the school in which the research was performed were invited, and 41 students from eight different classes who volunteered to participate and whose parents signed consent forms were included in the study. Of the students in the study group, 24 were girls (58%) and 17 were boys (42%).

2.1.2. *Raters*

Ten teachers participated in an assessment of oral reading records. For the determination of the teachers, the criterion sampling method of purposive sampling was used. In line with this, the criterion was that all teachers participating in the study had been employed for at least ten years. The situation leading to this criterion was that oral reading skills of students are assessed by elementary school classroom teachers and by Turkish language arts teachers in middle school. Before the data files were sent to the raters, they were given rater training. The raters were given information about the sub-dimensions of reading prosody, how it is assessed, and what requires attention during assessment, and the rubric to be used was described. In addition to the information given during training, a written form related to the rubric and the elements that require attention during the rating process was prepared and given to the raters. The raters assessed the voice recording for each student using the MDFS and included their scores in an

Excel table. Information related to the demographic characteristics of the raters is given in Table 1.

Table 1. *Information related to raters.*

Rater	Gender	professional experience (years)	Branch
R1	Male	20	Turkish Language Arts
R2	Female	10	Turkish Language Arts
R3	Female	14	Turkish Language Arts
R4	Female	18	Turkish Language Arts
R5	Female	15	Turkish Language Arts
R6	Male	15	Elementary School
R7	Female	17	Elementary school
R8	Female	10	Elementary school
R9	Female	30	Elementary school
R10	Female	17	Elementary school

Table 1 shows that the raters included five Turkish language arts and five elementary school classroom teachers and that their years of experience varied from 10 to 30 years.

2.2. Measurement Tools

In order to identify the prosodic reading skills of students, a narrative text was chosen. The MDFS was used to assess recordings of oral readings of this text.

2.2.1. *Narrative Text*

In order to measure reading prosody, a text was chosen in line with expert opinion from the Turkish textbook used in previous years and permitted by the Ministry of National Education and Board of Education and Discipline (MoNE, 2016). When choosing texts, opinions were sought from three Turkish language arts teachers, three elementary school classroom teachers and an academic in the field of Turkish education. The selected text was a story containing 275 words. When deciding on the type of text, again expert opinion was sought. It was concluded that stories were more suitable in reflecting prosodic reading elements (reflecting mutual dialogue and emotional variations). Students read the text aloud, a voice recording was made for each student, and a one-minute portion of the reading was assessed. In the literature, one-minute voice recording samples were stated to be sufficient for assessment of prosody (Rasinski et al., 2017; Zimmerman et al., 2019), with no significant difference found between assessments of one minute and three minutes (Valencia et al., 2010).

2.2.2. *Multidimensional Fluency Scale*

The MDFS was developed by Zutell and Rasinski (1991), updated by Rasinski (2004) and adapted to Turkish by Yıldız et al. (2009). It comprises four dimensions. These are “expression and volume”, “phrasing”, “smoothness” and “pace”. These are rated from one (1) to four (4) points with a graded rating key. Scores obtained from the four dimensions comprise the total prosody scores, and so the lowest score that can be obtained is 4, while the highest score is 16. The rubric contains statements for each point and raters perform the assessment in line with these statements. For example, for the ‘phrasing’ dimension, ‘1’ point is equivalent to the statement “reading is monotone, reader does not pay attention to units of meaning or word groups, mostly reads word-by-word”, while ‘4’ points are equivalent to the statement “generally reads by paying attention to word groups and units of meaning, reveals the emotional features of expressions in appropriate phrasing”. Additionally, if total scores obtained from the rubric at the end of the assessment are less than 10, it means reading is inadequate in prosodic terms and requires development, while scores of 10 or more are accepted as adequate prosodic reading

(Rasinski et al., 2017). Many studies in the research stated that valid and reliable measures were obtained by rating using this rubric (Ceyhan, 2019; Kaya Tosun, 2019; Moser et al., 2014; Rasinski et al., 2017; Rasinski et al., 2009; Valencia et al., 2010; Zimmerman et al., 2019).

2.3. Procedure

The research data were collected in the fall semester of the 2019-2020 educational year. Permission for the research was granted by the Provincial Directorate of National Education (Number: 1 88023 89-44-E.22033 447) and an ethics committee report was obtained (Decision number: 2020-39). Additionally, detailed information was given to parents about the research regarding making voice recordings of the students, and the necessary permission was obtained by signing the 'Parental Consent Form'. Data for the research were collected from fifth-grade students in a state school in the center of the city. The school administration, guidance and counseling service and the Turkish language arts teachers in classes in which data would be collected were given detailed information about the content and aims of the research. When collecting data, care was taken to ensure students were in an environment in which they felt comfortable, with voice recordings made in the school meeting room to ensure a quiet environment. When taking voice recordings, each student was talked to for a few minutes before reading to minimize the student's agitation, and attempts were made to overcome problems with breath and agitation control.

2.4. Data Analysis

The MFRM was used for data analysis related to the MDFS. The MFRM is a member of the Rasch model family. The Rasch model was originally used for dichotomous data, while later, the Rasch model also began to be used for polytomous data. Parameters for the MFRM are expressed on a common scale called a 'logit scale'. The logit scale unit makes it possible to compare units on every facet with others (Linacre, 1994). The Rasch model is linked to the difficulty level of items for the competence levels of individuals. In addition to competence and item difficulty levels, the MFRM is a model paying attention to the potential effects on performance outcomes of performance criteria, measurement time, raters, and other sources of variance and their interactions (Eckes, 2015; Linacre, 2002). Additionally, the MFRM provides information about how well the values predicted by the model created by performance analysis for each individual, rater or task match the expected values (Sudweeks et al., 2004). The MFRM analysis was conducted using the FACETS program (Linacre, 1994). The facets in the MFRM are calibrated simultaneously on a single linear scale. Thus, it is possible to measure the severity or leniency of a rater on the same scale as the difficulty of tasks/items for the competence of individuals (Eckes, 2015).

2.4.1. *Separation index and separation index reliability*

The separation index and reliability are separately calculated for each facet in the model (Schumacker & Smith, 2007). The reliability of raters on a facet represents ratings being different in a reliable way, rather than similar in a reliable way (Haiyang, 2010). For this reason, it should not be considered a measure of the reliability of raters or consistency between raters (Sudweeks et al., 2004). If the aim is to separate individuals in terms of performance, the separation reliability value should be high (Myford & Wolfe, 2003). While the separation index has values from 1 to ∞, the reliability coefficient has values from 0 to 1 (Sudweeks et al., 2004). For individual facets, the reliability of the separation index may be interpreted similarly to the Cronbach alpha coefficient (Engelhard & Myford, 2003; Myford & Wolfe, 2003).

2.2.2. *Fit statistics and chi-square statistic*

For each facet in the research, two statistics are obtained: 'infit' and 'outfit' mean squares. Wright and Linacre (1994) and Linacre (2002) stated that the lower limit for these values was 0.5 and the upper limit was 1.5. Bond and Fox (2015) stated that if the infit and outfit mean

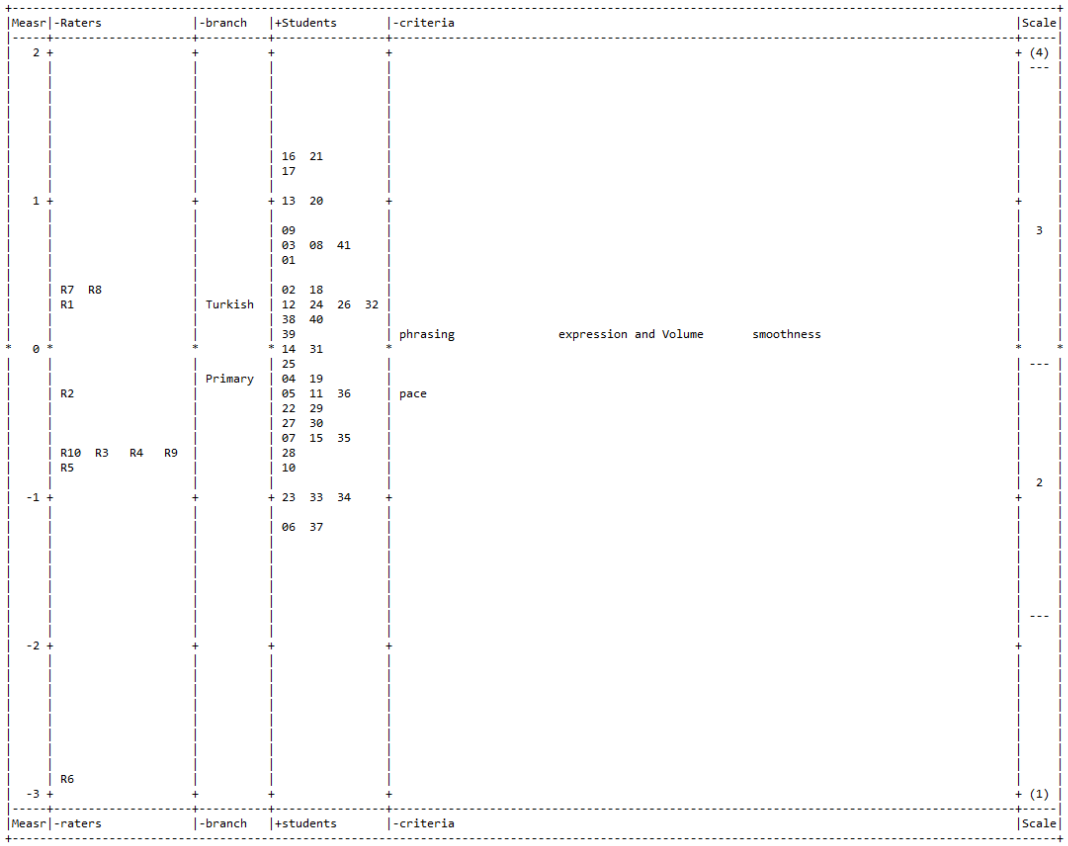
squares values were larger than 1.30, there was no fit and values below 0.70 represented overfitting. Fit values larger than 1 show more variability than expected between the scores of raters, while values below 1 are interpreted as showing less variability than expected (Eckes, 2015). Additionally, Linacre (2011) stated that the Rasch-Cohen kappa value may be used to assess whether raters acted independently or not, and that this value should be close to 0. The chi-square statistic (fixed) provides information about whether there is a significant difference present or heterogeneity between elements/levels of a facet. If the chi-square statistic is significant, it is interpreted that there is a difference between levels of the facet. A significant chi-square statistic for the rater facet shows that at least two raters do not share the same parameter (Eckes, 2005). In this research, there were four facets of rater, branch, student and criterion (dimension). The infit and outfit mean squares statistics, reliability values and separation rates were interpreted for each facet.

3. RESULTS

3.1. Findings of MFRM

This section includes findings obtained from analyzing reading prosody with the MDFS in practice. The logit-scale obtained as a result of the MDFS used by a total of 10 raters as Turkish and elementary school classroom teachers, assessing 41 students and four criteria on three facets is presented in Figure 1.

Figure 1. *Logit Scale for Four Facets*



The first column in Figure 1 shows the logit, the measurement unit of the logit scale. The ability levels of the students included in the study, difficulty level of the criteria, rater branches and rater severity/leniency levels are interpreted based on this measurement unit. The second column in the figure contains measurements belonging to raters and gives the opportunity to make interpretations about the severity/leniency of raters. This means the rater with the highest

logit score in this column performs the most severe rating and the rater with the lowest logit score performs the most lenient rating. When the figure is investigated, the most severe rating was given by the seventh rater (.42), while the most lenient rating was given by the sixth rater (-2.87). The third column of the figure lists the performance for the criteria in terms of the branch of the rater; in other words, it represents the raters' ability to give scores. The logit values for the Turkish teaching branch (.32) were found to be higher than the values for primary teachers (-.16). The fourth column of the figure lists the students from highest to lowest in terms of performance for the criteria found in the graded rating key. Thus, students number 16 and number 21 had the highest performance (1.3) and students number 4 and number 37 had the lowest performance (-1.25). The positive and negative values on the logit scale for student performance of criteria included on the reading prosody scale, in other words the spread over a wide range, shows that students could be well differentiated from each other. The final column of the logit scale lists the difficulty level of the dimensions. Accordingly, the 'phrasing' and 'expression and volume' dimensions (0.14) were the most difficult, while the 'pace' dimension (-.32), where students obtained the highest scores, was the easiest.

The logit scale provides important information about all facets and is included in measurement reporting in order to obtain more detailed information about all facets. Firstly, the findings for the measurement report related to raters are included in Table 2.

Table 2. *Measurement report related to raters.*

Raters	Mean	Measure	Model Standard error	Infit MnSq	Outfit MnSq
7	2.44	.42	.10	1.33	1.33
8	2.44	.41	.10	1.33	1.33
1	2.27	.33	.10	.65	.64
2	2.62	-.30	.10	1.08	1.10
4	2.83	-.68	.11	1.04	1.03
10	3.01	-.69	.11	.80	.79
3	2.84	-.69	.11	.58	.58
9	3.04	-.75	.11	.77	.80
5	2.88	-.77	.11	1.29	1.27
6	3.76	-2.87	.18	.92	1.14
Population		.88	.02	.26	.27
Sample		.93	.02	.28	.28
Model population RMSE = 0.11		S.S.=.88	Separation index = 7.64		Reliability = .98
Model sample RMSE = 0.11		S.S.=.92	Separation index = 8.06		Reliability = .98
Model fixed chi-square = 384.4	*df*=9		*p*=.00		
Model random chi-square = 8.8	*df*=8		*p*=.36		

When Table 2 is investigated, the reliability and separation index values related to the rater facet were .98 and 8.06, respectively. As the separation index approaches zero, the severity/leniency of raters is accepted as being more similar. The separation index reliability value is interpreted as showing how well separated raters are on the rater facet (Eckes, 2015). The high reliability and separation index values can be said to show that raters differed in their rating. In other words, it means there was unwanted variance between raters and this variance contributed to measurement error (Engelhard, 2002). Additionally, the chi-square statistic (χ^2=384.4, *df*=9, *p*=.00) shows that there was a statistically significant difference between raters in terms of severity and leniency. The negative values in the measurement column show that there were raters giving more generous scores compared to others, while positive values show more severe raters. Severe raters have a tendency to assign lower evaluations/scores consistently compared to other raters, while more lenient raters have a tendency to assign higher evaluations/scores (Myford & Wolfe, 2004). In this research, the Rasch-Cohen's kappa statistic

was -.13 and this shows that raters gave consistent scores, though only partially. The infit and outfit mean squares (.58/1.33) obtained in the study appear to be within the desired interval (Linacre, 2014).

Additionally, when the mean total scores given by the raters are examined, apart from one teacher (3.76), the other teachers had a tendency to give category means from 2.44 to 3.04. Accordingly, it is understood that students displayed prosodic skills at moderate and close to adequate levels. The general mean given to students was 2.81, which may be interpreted as showing that teachers viewed the prosodic skills of students as partly adequate. The measurement report related to the branches of the raters is given in Table 3.

Table 3. *Measurement report according to branch of raters.*

Branch	Measure	Standard error	Infit MnSq	Outfit MnSq
Turkish	.32	.05	.93	.92
Primary	-.16	.05	1.06	1.08
Population	.13	.00	.07	.08
Sample	.18	.00	.10	.11
Model population RMSE = .05		S.S.=.16	Separation index = 3.21	Reliability = .91
Model sample RMSE = .05		S.S.=.23	Separation index = 4.65	Reliability = .96
Model fixed chi-square = 22.6	*df*=1	*p*=.000		

When Table 3 is investigated, the reliability and separation index values related to the branch of the raters were found to be .96 and 4.65, respectively. The reliability and separation index values show that the scores given by raters differed in terms of branch. Additionally, the chi-square test results (χ^2=22.6, *df*=1, *p*=.00) reflect the significant difference between scores given by Turkish and primary teachers. According to Table 3, the infit (.93/1.06) and outfit (.92/1.08) mean squares appear to be within the desired interval. The measurement report related to the student facet is given in Table 4.

In Table 4, the reliability and separation index values for the student facet were .90 and 2.97, respectively. Additionally, the significant results of the chi-square test (χ^2=370.8, *df*=40, *p*=.00) and the high separation index and reliability indicate the students displayed differences in reading prosody skill levels. Additionally, when the infit and outfit mean squares are investigated, only individuals numbered 2, 3, 33 and 39 had fit values larger than 1.5, in other words, outside the accepted interval. Linacre (2003) stated that infit and outfit mean squares between 1.5 and 2.0 were not productive but not harmful, while values above 2.0 disrupted the model. For this reason, these individuals did not break the model. The measurement report related to the criterion facet is given in Table 5.

According to Table 5, the reliability and separation index values related to the criterion facet were .90 and 3.02, respectively. Additionally, when the 'there is no difference between criteria difficulty levels' hypothesis was tested with the chi-square test, it was significant (χ^2=29.3, *df*=3, *p*=.00). This means that the tasks on the MDFS differed in terms of difficulty levels.

Table 4. *Measurement report for students.*

Ind. No	Scores	Logit Score	S.E.	Infit MnSq	Infit MnSq	I. no	Scores	Logit Score	S.E.	Infit MnSq	Infit MnSq
16	137	1.30	.26	1.06	1.49	25	112	.21	.2	1.35	1.31
21	137	1.30	.26	.93	1.41	4	109	.21	.2	.94	.96
17	136	1.23	.26	1.24	1.10	19	108	.21	.21	.69	.66
13	132	.98	.25	.96	.86	5	106	.21	.21	1.14	1.10
20	132	.98	.25	.96	.89	11	106	-.33	.21	.86	.84
9	129	.80	.25	.98	.1.03	36	106	-.33	.21	1.25	1.23
3	128	.74	.24	1.52	1.55	22	105	-.38	.21	0.6	0.59
8	128	.74	.24	.53	.62	29	105	-.38	.21	1.18	1.19
41	127	.69	.24	.82	.77	30	103	-.46	.21	.79	.77
1	126	.63	.24	1.37	1.35	27	102	-.51	.21	.77	.78
2	122	.47	.23	1.58	1.64	7	101	-.55	.21	.97	.96
18	121	.37	.23	.49	.64	15	101	-.55	.21	.65	.66
24	120	.32	.23	.62	.57	35	101	-.55	.21	.74	.72
12	119	.27	.22	.86	.82	28	98	-.68	.21	.68	.68
26	119	.27	.22	.49	.47	10	95	-.81	.21	.75	.76
32	119	.27	.22	1.25	1.28	23	91	-.98	.21	.71	.68
38	118	.12	.22	1.22	1.17	33	91	-98	.21	1.91	1.83
40	117	.17	.22	.93	.94	34	90	-1.03	.21	.82	.81
39	116	.13	.22	1.82	1.73	6	85	-1.25	.21	.47	.48
14	114	.03	.22	.81	.76	37	85	-1.25	.21	.1.26	1.26
31	113	-.02	.22	.58	.56						

Model population RMSE = .22	S.S.=.65		Separation index = 2.93	Reliability = .90
Model sample RMSE = .22	S.S.=.66		Separation index = 2.97	Reliability = .90
Model fixed chi-square = 370.8	*df*=40	*p*=.00		
Model random chi-square = 36.02	*df*=39	*p*=.60		

Ind. No: Individual Number, S.E.: Standard Error

Table 5. *Measurement report for criterion facet.*

Criteria	mean	Logit	S.E.	Infit MnSq	Outfit MnSq
2 phrasing	2.74	.14	.06	1.09	1.14
3 smoothness	2.74	.13	.07	.91	.94
1 expression and volume	2.78	.06	.07	.91	.93
4 pace	2.98	-.32	.07	1.05	1
Mean	2.81	.00	.07	.99	1
Population		.19	.00	.08	.08
Sample		.22	.00	.09	.10

Model population: RMSE = .07	S.S.=.18	Separation = 2.57	Reliability = .87
Model sample: RMSE = .07	S.S.=.21	Separation = 3.02	Reliability = .90
Model fixed chi-square = 29.3, *df*=3, *p*=.00			
Model random chi-square = 2.7, *df*=2, *p*=.25			

3.2. Central Tendency Behavior

One of the rater errors frequently encountered with scores given with a graded key is central tendency behavior. Central tendency behavior indicates that when raters are assigning scores they tend to avoid giving high or low scores and give central scores. Values related to the rating categories (from 1-4) in the graded rating key used for this were investigated. The measurement report related to rating categories is presented in Table 6.

Table 6. *Statistics related to scale structure in rating categories.*

Branch	Category	*f*	%	Mean measure	Outfit MnSq	Rasch–Andrich threshold value	
						Measure	S.E.
Primary	1	74	9%	-.30	.8		
	2	162	20%	-.07	.6	-.84	.13
	3	326	40%	.61	1	-.37	.09
	4	258	31%	1.46	.9	1.21	.09
Turkish	1	74	9%	-0.36	.9		
	2	250	30%	-0.09	.7	-1.34	.13
	3	356	43%	0.39	.8	-0.14	.08
	4	140	17%	0.73	1	1.47	.10

When the table is investigated, the outfit mean squares varied from .6 to 1; in other words, they were within the desired interval (Linacre, 2014). Additionally, as the rating categories increase (from 1 to 4), a monotonous increase in threshold values is expected (Eckes, 2015). It appeared that the Rasch-Andrich threshold values monotonously increased and that all values were smaller than the logit value 5. When the frequency and percentage values related to rating categories are investigated on the table, it was identified that elementary school classroom teachers used rating categories 3 and 4, and that Turkish language arts teachers used rating categories 2 and 3 more often.

3.3. Bias Interaction

If the *t* values obtained from the interaction tables are outside the ±2 interval, they should be investigated for interaction effects. When the branch interaction of raters is investigated, the t values were within the desired interval. However, there was no statistically significant interaction effect according to branch. This study included 10 raters and 4 criteria. For this reason, there were a total of 40 interactions. The table related to the bias interaction according to criteria and raters is given in the appendix (Appendix). When the findings obtained for the rater-criteria interaction are investigated, the fifth, seventh and eighth raters were identified to have t values for the pace criterion outside the ±2 interval, while the third and sixth raters had *t* values for expression and volume outside the ±2 interval. Negative *t* values for the scores of these raters show that their scores were lower than expected, while positive values show that their scores were higher than expected. In other words, there was a difference between the scores expected and the scores observed for the rating by these raters, and bias was present. The eighth, seventh and sixth raters had positive bias and were more lenient raters, while the third and fifth raters had negative bias and were more severe raters. As the t values for the other raters were within the expected interval, rating bias can be ignored. Additionally, the interaction effect for rater and criteria facets was statistically significant ($\chi^2 = 58.1$, $df = 40$, $p = .03$).

4. DISCUSSION and CONCLUSION

The research aimed to investigate the reliability of results obtained with the MDFS assessed by elementary school classroom teachers and Turkish language arts teachers with the MFRM. When the results related to the first problem in the research, “Do teachers display differences in terms of severity/leniency when assessing students’ prosodic reading?” are investigated, the rating behavior of the raters included in the study was reliable, they were reliably ranked in terms of severity and leniency, and their severity/leniency levels were different from each other. When examined from the perspective of evaluators, the difference in severity/leniency levels is an unwanted situation and will restrict the ability of evaluators to take each other’s places (Eckes, 2015). Linacre (2012) emphasized that infit and outfit mean squares smaller than 0.5 show overfitting of the model and give misleading results, and indicate that evaluators did not use the full interval in the rubric. Additionally, values between 0.5 and 1.5 indicate that the values are efficient for measurement. In the results of the research, though there were

differences between teachers in terms of severity and leniency, the infit and outfit mean squares were within the values proposed by Linacre (2012). It was concluded that the sixth rater was the most generous and the seventh rater was that strictest among raters. Goodwin (2016) stated that rating judgments of raters may be affected by the examples rated previously. Additionally, halo effects, described as bias in rating caused by different aspects other than the judged dimension of a person, can lead to misjudgments (American Psychological Association, 2015). In this study, therefore, raters may have been affected by the reading performance of the previous student or by the halo effect when evaluating students. This may have caused differences in the severity/leniency levels of raters.

When the results related to the second problem in the research, "Are there differences in terms of severity/leniency during assessment of students' prosodic reading according to teaching branch?" are investigated, the two raters who were the most severe and most lenient were observed to be classroom teachers. In other words, raters who were Turkish language arts teachers performed more consistent and similar ratings when giving scores for the rubric, while classroom teachers gave different ratings compared to each other. This situation may have caused the rating reliability of raters to be low. When choosing evaluators, it is necessary to find those with appropriate educational backgrounds and experience in the field (Myford & Wolfe, 2003). The prejudices, attitudes, and personality traits of evaluators and the purpose of the assessment may cause a tendency to evaluate more severely (Eckes, 2015). Classroom teachers giving scores at the extremes compared to Turkish language arts teachers may be interpreted with a variety of variables. Turkish language arts teachers take many courses directly and indirectly related to language skills during undergraduate education, while classroom teachers take courses in different areas like mathematics, music, etc. In fact, Taşkaya and Muşta (2008) identified that classroom teachers felt they were inadequate and that the education they received was inadequate with regard to Turkish teaching. In this situation, it may be expected that the general knowledge and judgments related to language skills of classroom teachers will be different to those of Turkish language arts teachers. Some studies in the literature showed differences (Coşkun & Coşkun, 2014; Doğan, 2013) and similarities (Benzer & Eldem, 2013; Doğan, 2013; Saracoğlu et al., 2011) in terms of a variety of features between Turkish language arts teachers and classroom teachers. However, there are findings showing that teachers in both branches do not use rubrics often (Acar & Anıl, 2009; Benzer & Eldem, 2013). It may be considered that when teachers perform assessments with rubrics that they do not use much, their lack of familiarity with these tools may affect the judgments made and may lead to more personal behavior when giving scores. This situation may be interpreted as showing that teachers do not have adequate experience with the use of rubrics. Stevens and Levi (2005) stated that as experience is gained with rubrics, reliable ratings will increase. Mathson et al. (2006) considered that the lack of training on the assessment and teaching of fluency limits the use of rubrics in the classroom. From this perspective, the difference in scores between teachers may be explained by the lack of experience related to rubrics. It is possible that another source of difference between scores is the students they are in constant contact with in both branches. While the Turkish curriculum expects students to have upper-level reading skills, the Turkish curriculum used by classroom teachers requires more basic skills. In this situation, teachers have different expectations and it is probable that this difference was reflected in the assessment of prosody.

When the results related to the third problem in the research, "What are the results for task/criterion difficulty analysis related to students' prosodic reading?" are investigated, the prosody criteria can be ranked in order from more difficult to easier for students as phrasing, smoothness, expression and volume, and pace. This situation overlaps with the ranking accepted in the literature. According to researchers, students begin with letter-sound relationships and move toward higher units (for example, units of meaning), before completing

accurate, paced and prosodic reading in order (Baştuğ, 2021; Keskin, 2012; Mathson et al., 2006; Samuels, 2006). The findings in this study show that students received the highest scores for the pace dimension on the rubric. In the study by Godde et al. (2019), the dimension of prosody completed with the most difficulty was intonation and expressive reading (Godde et al., 2019). Considering that intonation and reading in meaningful units are closely related to each other (Godde et al., 2019), students who read quickly were determined not to understand the text because reading in meaningful units requires implicit clues in the text to be solved by readers (Rasinski, 2004; Schreiber, 1991). Based on this, the degree of difficulty emerging in student assessment according to the MDFS by teachers is supported by previous findings. However, the findings conflict with the finding of Godde et al. (2019) that expression was the most difficult. In this study, 'expression and volume' (mean 2.78) was the dimension with the second highest scores received by students and was easier than smoothness (mean 2.79). The lack of a large difference between these may be interpreted as showing that the difficulty of the two dimensions may occasionally change places; in all cases, the most difficult dimension was reading with intonation and meaningful groups, while pacing was the easiest dimension. In fact, Daane et al. (2005) stated that students who read quickly may develop reading skills for word groups.

When the results related to the fourth problem in the research, "What are the outcomes of central tendency behavior and bias analysis of raters?" are investigated, when the mean total scores given by the raters are examined, the general mean was 2.81. The classroom teachers were found to use categories 3 and 4 more often, while Turkish language arts teachers used categories 2 and 3 more often. In other words, it may be said that Turkish language arts teachers chose central categories as raters. One of the reasons for this may be that when raters gave scores to students, they saw students as having partially adequate prosodic skills; in other words, the raters displayed central tendency behavior (Myford & Wolfe, 2004). Results from assessments by teachers using rubrics are similar to the study by Daane et al. (2005). In the study by Daane et al. (2005), 83% of students were grouped in the 2nd and 3rd categories, dominantly in the 3rd category, and scores were closer to the 3rd category. For this reason, the scores given by teachers to students were similar to previous findings in terms of distribution. Another result of the research is that the eighth, seventh and sixth raters had positive bias, while the third and fifth raters had negative bias. The reliability of these raters for assessment of prosodic reading by students was lower compared to other raters. The findings obtained in the research are consistent with the results of other studies researching rater bias (Baştürk & Işıkoğlu, 2008; Köse et al., 2016; Özbaşı & Kumandaş-Öztürk, 2020; Şata & Karakaya, 2020; Yüzüak et al., 2015). Goodwin (2016) stated that additional training would be beneficial for rater bias. When the findings obtained in the research are investigated, it was concluded that the reading prosody rubric used in the research served the purpose of measuring reading prosody of students, the sub-dimensions on the rubric could reliably differentiate, the criteria determined were reliable, and the criteria categories were suitable and adequate for measurement. Another result is that the prosody criterion where students experienced the most difficulty was phrasing. When examined generally, it was concluded that the MDFS is a reliable rubric for use by researchers and teachers to evaluate prosodic reading skills. This result overlaps with the findings of two studies found in the literature using the generalizability theory to determine the reliability of the scale (Moser et al., 2014; Smith & Paige, 2019). Additionally, it may be said that it is necessary to train people for rating prosody using the MDFS (Bilge, 2019; Erguvan & Dünya, 2020; Kaya Uyanık et al., 2019; Smith & Paige, 2019; Zutell & Rasinski, 1991); however, training may not be adequate all the time (Barrett, 2001; Eckes, 2015; Yan, 2014). For this reason, to ensure the reliability of MDFS ratings, the use of at least two but preferably three texts (Moser et al., 2014), and the presence of at least two raters (Smith & Paige, 2019) are recommended. However, the reliability obtained for measurements with two raters may be

seriously misleading and it should not be forgotten that high rater reliability obtained with two raters does not always mean accurate rating (Eckes, 2015). For this reason, more reliable results will be obtained with the MFRM so as not to ignore the severity of raters in adapted or prepared rubrics; otherwise, the problems that could arise are actually ignored (Bond & Fox, 2015). In this study, the rubric developed by Zutell and Rasinski (1991), updated by Rasinski (2004) and adapted to Turkish by Yıldız et al. (2009) was used. Similar studies may investigate rater behavior using different rubrics/scales measuring prosody.

In this research, a fatigue effect may be present in the results obtained by raters evaluating forty students. For this reason, future studies may perform investigations on reading data obtained with more texts and fewer students. In this research, the rater reliability of the MDFS was researched. In addition to elements contributing to the literature in this way, there are a number of limitations of the study, including assessment of reading prosody with one text, and the inclusion of forty students and ten raters in the study.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Ordu University, 27.05.2020, 2020-39.

Authorship Contribution Statement

Cigdem Akin Arikan: Investigation, Review of Literature, Introduction, Methodology, Discussion and Conclusion. **Pinar Kanik Uysal**: Investigation, Introduction, Review of Literature, Methodology, Discussion and Conclusion. **Huzeyfe Bilge:** Introduction, Review of Literature, Discussion and Conclusion. **Kasim Yildirim:** Supervision.

Orcid

Cigdem Akin Arikan https://orcid.org/0000-0001-5255-8792
Pinar Kanik Uysal https://orcid.org/0000-0003-1208-9535
Huzeyfe Bilge https://orcid.org/0000-0001-7664-488X
Kasim Yildirim https://orcid.org/0000-0003-1406-709X

REFERENCES

Acar, M., & Anıl, D. (2009). Sınıf öğretmenlerinin performans değerlendirme sürecindeki değerlendirme yöntemlerini kullanabilme yeterlilikleri, karşılaştıkları sorunlar ve çözüm önerileri [Classroom teacher evaluation methods to use in the performance assessment process qualification of able, they comparison problems and solution proposals]. *Journal of TUBAV Science*, *2*(3), 354-363.

American Psychological Associaton. (2015). Halo effect. In APA dictionary of psychology (2nd ed., p. 667).

Allington, R.L. (1983). Fluency: the neglected reading goal. *The Reading Teacher*, *36*(6), 556-561.

Armut, M., & Türkyılmaz, M. (2017). Ortaokul öğrencilerinin okuma becerileri üzerine bir inceleme [An investigation on reading skills of middle school students]. *Erzincan University Journal of Education Faculty*, *20*(1), 217-236. https://doi.org/10.17556/erziefd.330587

Aşıkcan, M. (2019). *Üçüncü sınıf öğrencilerinin akıcı okuma becerilerinin geliştirilmesine yönelik bir eylem araştırması* [*An action research on ımproving fluent reading skills of third-grade primary school students*] [Unpublished doctoral dissertation]. Necmettin Erbakan University.

Baird, J.A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). Marker effects and examination reliability. A Comparative exploration from the perspectives of

generalisability theory, Rash model and multilevel modelling. Oxford: University of Oxford for Educational Assessment. Retrieved from http://dera.ioe.ac.uk/17683/1/2013-01-21-marker-effects-and-examinationreliability.pdf

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* [Doctoral Dissertation, Available from ProOuest Dissertations and Theses database]. UMI No: 304360302.

Barrett, S., (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49-58.

Baştuğ, M., (2021). *Akıcı okumayı geliştirme: kavramlar, uygulamalar, değerlendirmeler [Developing reading fluency: concepts, practices, assessments]*. Pegem Akademi Yayıncılık

Baştuğ, M., & Keskin, H.K. (2012). Akıcı okuma becerileri ile anlama düzeyleri (basit ve çıkarımsal) arasındaki ilişki [The relationship between fluent reading skills and comprehension level (literal and ınferential)]. *Ahi Evran University Journal of Kırşehir Education Faculty*, *13*(3), 227-244.

Baştürk, R., & Işıkoglu, N. (2008). Analyzing process quality of early childhood education with many facet rash measurement model. *Educational Sciences:Theory and Practice, 8*(1), 25-32.

Benjamin, R.G., & Schwanenflugel, P.J. (2010). Text complexity and oral reading prosody in young readers. *Reading Research Quarterly, 45*(4), 388-404. https://doi.org/10.1598/RRQ.45.4.2

Benjamin, R.G., Schwanenflugel, P.J., Meisinger, E.B., Groff, C., Kuhn, M.R., & Steiner, L. (2013). A spectrographically grounded scale for evaluating reading expressiveness. *Reading Research Quarterly, 48*(2), 105-133. https://doi.org/10.1002/rrq.43

Benzer, A., & Eldem, E. (2013). Türkçe ve edebiyat öğretmenlerinin ölçme ve değerlendirme araçları hakkında bilgi düzeyleri [Level of the information about Turkish and literature teachers' measurement and assessment materials]. *Kastamonu Education Journal, 21*(2), 649-664.

Bilge, H. (2019). *Okuma, yazma ve konuşma akıcılık ile okuduğunu anlama ve kelime hazinesi arasındaki ilişki [The relationships between reading, writing and speaking fluencies, reading comprehension and vocabulary*] [Unpublished doctoral dissertation]. Gazi University.

Bond, T.G., & Fox, C.M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3th ed.). Routledge.

Brookhart, S. M. (2013). *How to Create and Use Rubrics for Formative Assessment and Grading*. ASCD.

Buck, J., & Torgesen, J. (2003). The relationship between performance on a measure of oral reading fluency and performance on Florida comprehensive assessment test. FCRR Technical Report# 1. *Florida Center for Reading Research.*

Ceyhan, S. (2019). *Etkileşimli sesli okumanın öğrencilerin okuduğunu anlama, okuma motivasyonu ve akıcı okumalarına etkisi [The effect of interactive reading aloud on the reading comprehension, reading motivation and reading fluency of students]* [Unpublished doctoral dissertation]. Gazi University.

Coşkun, E., & Coşkun, H. (2014). İlkokul ve ortaokullardaki bitişik eğik yazı uygulamalarına ilişkin öğretmen, öğrenci ve veli görüşleri [teachers', students' and parents' views on cursive italic handwriting]. *Mustafa Kemal University Journal of Social Sciences Institute, 11*(26), 209-223.

Couper-Kuhlen, E. (1986). *An Introduction to English Prosody*. Edward Arnold.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics* (6th ed.). Blackwell Publishing.

Çetinkaya, Ç., Ateş, S., & Yıldırım, K. (2016). Prozodik okumanın aracılık etkisi: Lise düzeyinde okuduğunu anlama ve akıcı okuma arasındaki ilişkilerin incelenmesi [The mediation effect of reading prosody: exploring the relations between reading fluency and reading comprehension at high school level]. *Turkish Studies, 11*(3). https://doi.org/10.7827/TurkishStudies.9339

Daane, M.C., Campbell, J.R., Grigg, W.S., Goodman, M.J., & Oranje, A. (2005). *The nation's report card: fourth-grade students reading aloud: NAEP 2002 special study of oral reading*. Washington, D.C.: U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics.

Doğan, B. (2013). Türkçe ve sınıf öğretmenlerinin okuma güçlüğüne ilişkin bilgileri ve okuma güçlüğü olan öğrencileri belirleyebilme düzeyleri [Determining Turkish language and elementary classroom teachers' knowledge on dyslexia and their awareness of diagnosing students with dyslexia]. *Research in Reading & Writing Instruction, 1*(1), 20-33.

Dowhower, S.L. (1991). Speaking of prosody: fluency's unattended bedfellow. *Theory Into Practice, 30*(3), 165-175. https://doi.org/10.1080/00405849109543497

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal, 2*(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments* (Vol. 22). Peter Lang Edition. https://doi.org/10.1080/15366367.2018.1516094

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. A. Tindal & T. M. Haladyna (Eds.), *Large scale assesment for all students: Validity, technical adequacy, and implementation* (pp. 261-287). Erlbaum.

Engelhard, G., & Myford, C.M. (2003). Monitoring faculty consultant performance in the advanced placement English Literature and composition program with a many-faceted Rasch model. *ETS Research Report Series, 2003*(1), i-60.

Erguvan, İ.D., & Dünya, B.A. (2020). Analyzing rater severity in a freshman composition course using many facet Rasch measurement. *Language Testing in Asia, 10*(1), 1-20. https://doi.org/10.1186/s40468-020-0098-3

Esmer, B. (2019). *Okuduğunu anlama ile akıcı okuma, okur benlik algısı, okumaya adanmışlık ve okuyucu tepkisi ilişkileri [Direct and inferential relations among reading comprehension, silent and oral reading fluency, reading self-concept, reading engagement and response to picturebooks]* [Unpublished doctoral dissertation]. Gazi University.

Fraenkel, J.R., & Wallen, N.E. (2009). *How to Design and Evaluate Research in Education* (7th ed.). McGraw-Hill.

Godde, E., Bailly, G., Escudero, D., Bosse, M.L., & Gillet-Perret, E. (2017). Evaluation of reading performance of primary school children: objective measurements vs. subjective ratings.*WOCCI 2017-6th Workshop on Child Computer Interaction*, Nov 2017, Glasgow, United Kingtom.

Godde, E., Bosse, M.L., & Bailly, G. (2019). A review of reading prosody acquisition and development. *Reading and Writing, 33*(2), 399-426. https://doi.org/10.1007/s11145-019-09968-1

Goodwin, S., (2016). A many-facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing,* 30, 21-31. https://doi.org/10.1016/j.asw.2016.07.004

Grosjean, F., & Collins, M. (1979). Breathing, pausing and reading. *Phonetica, 36*(2), 98-114. https://doi.org/10.1159/000259950

Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics, 33*(2), 87-102.

Hallman, J. (2009). Reading aloud: comprehending, not word calling. In R. Stone (Ed.), *Best practices for teaching reading: what award-winning classroom teachers do* (pp. 39-43). Corwin Press.

Haskins, T., & Aleccia, V. (2014). Toward a reliable measure of prosody: an investigation of rater consistency. *International Journal of Education and Social Science, 1*(5), 102-112.

İlhan, M. (2015). *Standart ve solo taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli rasch modeli ile incelenmesi [The identification of rater effects on open-ended math questions rated through standard rubrics and rubrics based on the SOLO taxonomy in reference to the many facet Rasch model]* [Unpublished doctoral dissertation]. Gaziantep University.

Kanık Uysal, P., & Duman, A. (2020). The effects of fluency-oriented reading instruction on reading skills. *Pegem Journal of Education and Instruction, 10*(4), 1111–1146. https://doi.org/10.14527/pegegog.2020.034

Kaya Tosun, D. *(2019). Okuma çemberlerinin okuduğunu anlama, akıcı okuma, okuma motivasyonu ve sosyal beceriler üzerindeki etkisi ve okur tepkilerinin belirlenmesi [The effect of literature circles on reading comprehension, reading fluency, reading motivation and social skills and exploring of reader responses]* [Unpublished doctoral dissertation]. Gazi University.

Kaya Uyanık, G., Güler, N., Taşdelen Teker, G., & Demir, S. (2019). The analysis of elementary science education course activities through many-facet Rasch model. *Kastamonu Education Journal, 27*(1), 139-150. https://doi.org/10.24106/kefdergi.2417

Keskin, H.K. (2012). *Akıcı okuma yöntemlerinin okuma becerileri üzerindeki etkisi [Impact of reading fluency methods on reading skills]* [Unpublished doctoral dissertation]. Gazi University.

Kızıltaş, Y. (2019). *Ana dili farklı ilkokul öğrencilerinin akıcı okuma ve okuduğunu anlama becerilerinin incelenmesi [The study of reading fluency and reading comprehension skills of primary school whose mother tongue is different]* [Unpublished doctoral dissertation]. Gazi University.

Köse, İ.A., Usta, H.G., & Yandı, A. (2016). Sunum yapma becerilerinin çok yüzeyli Rasch analizi ile değerlendirilmesi [Evaluation of presentation skills by using many facets rasch model]. *Abant İzzet Baysal University Journal of Faculty of Education*, *16*(4), 1853-1864.

Kuhn, M.R. (2007). Effective oral reading assessment (or why round robin reading doesn't cut ıt). In J.R. Paratore & R.L. McCormack (Ed.), *Classroom literacy assessment: making sense of what students know and do* (pp. 101-112). The Guilford Press.

Kuhn, M.R., & Stahl, S.A. (2013). Fluency: developmental and remedial practices-revisited. In D.E. Alvermann, N.J. Unrau, & R.B. Ruddell (Ed.), *Theoretical models and processes of reading* (pp. 385-412). International Reading Association.

Li, G., Pan, Y., & Wang, W. (2021). Using generalizability theory and many-facet Rasch model to evaluate in-basket tests for managerial positions. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.660553

Linacre, J.M. (1993). Generalizability theory and many-Facet Rasch measurement, in *Paper presented at the Annual Meeting of the American Educational Research Association* (Atlanta, GA).

Linacre, J.M. (1994). *Many-Facet Rasch Measurement*. Chicago: Mesa Press.

Linacre, J.M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J.M. (2003). Size vs. significance: Infit and outfit mean-square and standardized chi-square fit statistic. *Rasch Measurement Transactions, 17*, 918.

Linacre, J.M. (2011). *Facets computer program for Many-Facet Rasch Measurement*. https://www.winsteps.com/facets.htm#:~:text=Facets%20is%20designed%20to%20handle,further%20measurement%20and%20structural%20facets

Linacre, J.M. (2012). *Many-facet Rasch measurement: Facets tutorials*. http://www.winsteps.com/tutorials.htm

Linacre, J.M. (2014). *A user's guide to FACETS Rasch-model computer programs*. http://www.winsteps.com/a/facets-manual.pdf

Mathson, D.V., Allington, R.L., & Solic, K.L. (2006). Hijacking fluency and instructionally informative assessments. In T. Rasinski, C. Blachowicz, & K. Lems (Eds.), *Fluency instruction: research-based best practices* (pp. 106-119). The Guilford Press.

McMillan, J.H. (2017). *Classroom Assessment. Principles and Practice that Enhance Student Learning and Motivation* (7th ed.), Pearson.

MoNE. (2016). *5. Sınıf Türkçe ortaokul ders kitabı [Turkish secondary school textbook for 5th graders*]. Devlet Kitapları.

Morrison, T.G., & Wilcox, B. (2020). Assessing expressive oral reading fluency. *Education Sciences, 10*(59). https://doi.org/10.3390/educsci10030059

Moser, G.P., Sudweeks, R.R., Morrison, T.G., & Wilcox, B. (2014). Reliability of ratings of children's expressive reading. *Reading Psychology, 35*(1), 58-79. https://doi.org/10.1080/02702711.2012.675417

Myford, C.M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C.M., & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189-227.

National Reading Panel. (2000). *Teaching children to read: an evidence-based assessment of the scientific research litearature on reading and its implications for reading instruction: reports of the subgroups*. Washington, DC: National Institute of Child Health Human Development.

Overstreet, T.B. (2014). *The effect of prosody instruction on reading fluency and comprehension among third-grade students* [Unpublished doctoral dissertation]. Andrews University.

Özbaşı, D. & Kumandaş-Öztürk, H. (2021). Öğretim materyallerinin çok yüzeyli Rasch analiziyle değerlendirilmesi [Evaluatıon of teaching materials with many-facet Rasch analysis]. *Trakya Journal of Education*, 11(1), 187-200.

Paige, D.D., Smith, G., Rupley, W., & Wells, W. (2021). Reducing high-attaining readers to middling: the consequences of inadequate foundational skills instruction in a high-ses district. *Literacy Research and Instruction, 60*(1), 81-106. https://doi.org/10.1080/19388071.2020.1780653

Palmer, M.L. (2010). *The relationship between reading fluency, writing fluency, and reading comprehension ın suburban third-grade students* [Unpublished doctoral dissertation]. San Diego State University.

Rasinski, T. (2004). *Assessing Reading Fluency*. Honolulu, Hawai: Pasific Resources for Education and Learning.

Rasinski, T. (2010). *The Fluent Reader*. Scholastic.

Rasinski, T., Paige, D.D., Rains, C., Stewart, F., Julovich, B., Prenkert, D., Rupley, W.H., & Nicholas, W.D. (2017). Effects of intensive fluency instruction on the reading proficiency of third-grade struggling readers. *Reading and Writing Quarterly, 33*(6), 519-532. https://doi.org/10.1080/10573569.2016.1250144

Rasinski, T., Rikli, A., & Johnston, S. (2009). Reading fluency: More than automaticity? More than a concern for the primary grades?. *Literacy Research and Instruction, 48*(4), 350-361. https://doi.org/10.1080/19388070802468715

Samuels, S.J. (2006). Reading fluency:ıts past, present, and future. In T. Rasinski, C. Blachowicz, & K. Lems (Ed.), *Fluency ınstruction research-based best practices* (pp. 7-20). The Guildgord Press.

Saracoğlu, S., Dedebali, N.C., Dinçer, B., & Dursun, F. (2011). Sınıf, fen ve teknoloji ile Türkçe öğretmenlerinin öğretmen stillerinin incelenmesi [Investigatıon of teaching styles of primary school, science and techonology, Turkish teachers]. *Education Sciences, 6*(3), 2313-2327.

Schreiber, P.A. (1991). Understanding prosody's role in reading acquisition. *Theory into Practice, 30*(3), 158-164. https://doi.org/10.1080/00405849109543496

Schumacker, R.E., & Smith, E.V. (2007). A Rasch perspective. *Educational and Psychological Measurement, 67*(3), 394-409. https://doi.org/10.1177/0013164406294776

Schwanenflugel, P.J., Hamilton, A., Kuhn, M.R., Wisenbaker, J.M., & Stahl, S.A. (2004). Becoming a fluent reader: reading skill and prosodic features in the oral reading of young readers. *Journal of Educational Psychology, 96*(1), 119-129. https://doi.org/10.1037/0022-0663.96.1.119

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. https://doi.org/10.1037/0033-2909.86.2.420

Sinambela, S.E. (2017). Prosody as a tool for assessing reading fluency of adult esl students. *Advances in Language and Literary Studies, 8*(6), 83-87. https://doi.org/10.7575/aiac.alls.v.8n.6p.83

Smith, G.S., & Paige, D.D. (2019). A study of reliability across multiple raters when using the NAEP and MDFS rubrics to measure oral reading fluency. *Reading Psychology, 40*(1), 34-69. https://doi.org/10.1080/02702711.2018.1555361

Spafford, C.S., Pesce, A.J.I., & Grosser, G.S. (Ed.). (1998). *The Cyclopedic Education Dictionary*. Delmar Publishers.

Stevens, D.D., & Levi, A.J. (2005). *Introduction to Rubrics: an Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning*. Stylus.

Sudweeks, R.R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239-261. https://doi.org/10.1016/j.asw.2004.11.001

Şata, M., & Karakaya, İ. (2020). Investigation of the use of electronic portfolios in the determination of student achievement in higher education using the many-facet Rasch measurement model. *Educational Policy Analysis and Strategic Research, 15*(7-21).

Taşkaya, S.M., & Muşta, M.C. (2008). Sınıf öğretmenlerinin Türkçe öğretim yöntemlerine ilişkin görüşleri [Teachers' opinions on Turkish teaching methods]. *Elektronik Sosyal Bilimler Dergisi, 7*(25), 240-251.

U.S. Department of Education. (2002). *National assessment of educational progress (NAEP) 2002 oral reading fluency study*. Washington, DC: Institute of Education Sciences, National Center for Education Statistics.

Ulusoy, M., Ertem, İ.S., & Dedeoğlu, H. (2011). Evaluating pre-service teachers' oral reading records prepared for the grades 1-5 considering the prosodic competences. *Gazi University Journal of Gazi Educational Faculty, 31*(3), 759-774.

Valencia, S.W., Smith, A.T., Reece, A.M., Li, M., Wixson, K.K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly, 45*, 270-291. https://doi.org/10.1598/RRQ.45.3.1

VandenBos, G.R. (2015). *APA Dictionary of Psychology*. American Psychological Association.

Xu, Y., & Liu, F. (2012). Intrinsic coherence of prosodic and segmental aspects of speech. In O. Niebuhr (Ed.), *Understanding prosody: the role of context, function and communcation* (Vol. 13, pp. 1-26). De Gruyter.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, *31*(4), 501-527. https://doi.org/10.1177/0265532214536171

Yıldız, M., Kanık Uysal, P., Bilge, H., Wolters, A.P., Saka, Y., Yıldırım, K., & Rasinski, T. (2019). Relationship between Turkish eighth-grade students' oral reading efficacy, reading comprehension and achievement scores on a high-stakes achievement test. *Reading Psychology*, *40*(4), 1-21. https://doi.org/10.1080/02702711.2018.1555363

Yıldız, M., Yıldırım, K., Ateş, S., & Çetinkaya, Ç. (2009). An evaluation of the oral reading fluency of 4th graders with respect to prosodic characteristic. *International Journal of Human Sciences*, *6*(1), 353-360.

Young, C., & Rasinski, T. (2009). Implementing readers theatre as an approach to classroom fluency instruction. *The Reading Teacher*, *63*(1), 4-13. https://doi.org/10.1598/RT.63.1.1

Yüzüak, A., Yüzüak, B., & Kaptan, F. (2015). Performans görevinin akran gruplar ve öğretmen yaklaşımları doğrultusunda çok-yüzeyli Rasch ölçme modeli ile analizi [A many-facet Rasch measurement approach to analyze peer and teacher assessment for authentic assessment task]. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(1), 1-11.

Zimmerman, B.S., Rasinski, T., Was, C.A., Rawson, K.A., Dunlosky, J., Kruse, S.D., & Nikbakht, E. (2019). Enhancing outcomes for struggling readers: Empirical analysis of the fluency development lesson. *Reading Psychology*, *40*(1), 70-94. https://doi.org/10.1080/02702711.2018.1555365

Zutell, J., & Rasinski, T. (1991). Training teachers to attend to their students' oral reading fluency. *Theory Into Practice*, *30*(3), 211-217. https://doi.org/10.1080/00405849109543502

APPENDIX

Bias Interaction

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias- Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	puanlayıcıla	Sq	Nu	pua	measr-	N	Ölçütler	measr-
161	156.59	41	.11	-1.01	.60	-1.69	40	.0984	1.1	1.4		36	6	P6	-2.87	4	hız	-.32
127	113.23	41	.34	-.64	.23	-2.86	40	.0067	.6	.7		3	3	P3	-.69	1	ifade ve ses düzeyi	.14
136	124.91	41	.27	-.60	.24	-2.45	40	.0190	1.5	1.4		35	5	P5	-.77	4	hız	-.32
106	96.86	41	.22	-.38	.20	-1.85	40	.0714	1.3	1.3		17	7	P7	.33	2	anlama üniteleri ve tonlama	.13
106	96.86	41	.22	-.38	.20	-1.85	40	.0714	1.3	1.3		18	8	P8	.33	2	anlama üniteleri ve tonlama	.13
130	123.05	41	.17	-.35	.23	-1.52	40	.1363	1.1	1.0		34	4	P4	-.68	4	hız	-.32
156	153.85	41	.05	-.29	.39	-.75	40	.4572	1.0	1.3		26	6	P6	-2.87	3	pürüzsüzlik	.06
121	114.80	41	.15	-.28	.22	-1.30	40	.2003	1.1	1.1		32	2	P2	-.30	4	hız	-.32
128	123.39	41	.11	-.23	.23	-1.01	40	.3177	.7	.8		29	9	P9	-.75	3	pürüzsüzlik	.06
125	120.44	41	.11	-.22	.22	-.98	40	.3315	1.0	.9		10	10	P10	-.69	1	ifade ve ses düzeyi	.14
96	91.64	41	.11	-.18	.20	-.89	40	.3793	.6	.6		21	1	P1	.29	3	pürüzsüzlik	.06
105	100.84	41	.10	-.17	.20	-.85	40	.4027	.7	.7		31	1	P1	.29	4	hız	-.32
123	121.74	41	.03	-.06	.22	-.27	40	.7850	.6	.6		9	9	P9	-.75	1	ifade ve ses düzeyi	.14
100	98.66	41	.03	-.06	.20	-.27	40	.7863	1.3	1.3		27	7	P7	.33	3	pürüzsüzlik	.06
100	98.66	41	.03	-.06	.20	-.27	40	.7863	1.3	1.3		28	8	P8	.33	3	pürüzsüzlik	.06
116	114.77	41	.03	-.05	.21	-.26	40	.7978	.7	.7		24	4	P4	-.68	3	pürüzsüzlik	.06
123	122.12	41	.02	-.04	.22	-.19	40	.8491	.7	.7		30	10	P10	-.69	3	pürüzsüzlik	.06
97	96.74	41	.01	-.01	.20	-.05	40	.9588	1.2	1.2		7	7	P7	.33	1	ifade ve ses düzeyi	.14
97	96.74	41	.01	-.01	.20	-.05	40	.9588	1.2	1.2		8	8	P8	.33	1	ifade ve ses düzeyi	.14
153	153.23	41	-.01	.03	.34	.08	40	.9372	.9	1.1		16	6	P6	-2.87	2	anlama üniteleri ve tonlama	.13
103	104.05	41	-.03	.04	.20	.22	40	.8307	1.3	1.3		2	2	P2	-.30	1	ifade ve ses düzeyi	.14
114	115.03	41	-.03	.05	.21	.22	40	.8297	1.3	1.3		5	5	P5	-.77	1	ifade ve ses düzeyi	.14
112	113.34	41	-.03	.06	.21	.28	40	.7811	.4	.4		13	3	P3	-.69	2	anlama üniteleri ve tonlama	.13
119	120.54	41	-.04	.07	.21	.33	40	.7405	.7	.7		20	10	P10	-.69	2	anlama üniteleri ve tonlama	.13
102	104.16	41	-.05	.09	.20	.44	40	.6608	1.0	1.0		12	2	P2	-.30	2	anlama üniteleri ve tonlama	.13
111	113.08	41	-.05	.09	.21	.43	40	.6659	.8	.8		14	4	P4	-.68	2	anlama üniteleri ve tonlama	.13
129	130.87	41	-.05	.10	.23	.43	40	.6660	.9	.9		39	9	P9	-.75	4	hız	-.32
87	89.86	41	-.07	.12	.21	.58	40	.5620	.4	.4		11	1	P1	.29	2	anlama üniteleri ve tonlama	.13
103	105.94	41	-.07	.12	.20	.60	40	.5502	.9	.9		22	2	P2	-.30	3	pürüzsüzlik	.06
118	121.84	41	-.09	.18	.21	.83	40	.4091	.8	.8		19	9	P9	-.75	2	anlama üniteleri ve tonlama	.13
111	115.13	41	-.10	.18	.21	.87	40	.3895	1.1	1.1		15	5	P5	-.77	2	anlama üniteleri ve tonlama	.13
126	129.74	41	-.09	.19	.22	.86	40	.3959	.8	.8		40	10	P10	-.69	4	hız	-.32
84	89.75	41	-.14	.24	.21	1.17	40	.2471	.9	.8		1	1	P1	.29	1	ifade ve ses düzeyi	.14
118	123.28	41	-.13	.25	.21	1.16	40	.2534	.5	.5		33	3	P3	-.69	4	hız	-.32
111	116.80	41	-.14	.25	.21	1.23	40	.2266	1.3	1.3		25	5	P5	-.77	3	pürüzsüzlik	.06
107	112.98	41	-.15	.26	.21	1.25	40	.2198	1.4	1.4		4	4	P4	-.68	1	ifade ve ses düzeyi	.14
108	115.03	41	-.17	.30	.21	1.48	40	.1476	.4	.4		23	3	P3	-.69	3	pürüzsüzlik	.06
97	107.79	41	-.26	.45	.20	2.21	40	.0329	1.3	1.3		37	7	P7	.33	4	hız	-.32
97	107.79	41	-.26	.45	.20	2.21	40	.0329	1.3	1.3		38	8	P8	.33	4	hız	-.32
147	153.19	41	-.15	.60	.29	2.09	40	.0427	.8	1.0		6	6	P6	-2.87	1	ifade ve ses düzeyi	.14
115.3	115.23	41.0	.00	-.02	.23	-.01			1.0	1.0	Mean (Count: 40)							
17.7	16.92	.0	.13	.31	.07	1.21			.3	.3	S.D. (Population)							
18.0	17.13	.0	.14	.31	.07	1.22			.3	.3	S.D. (Sample)							