Education and Science

Original Article

2025 Supplement 1 43-68

A Tool for Selecting and Writing Texts for Assessing Reading: Reader-Friendly Informative Text Rubric (ODBIMDEPA) *

Dilek Fidan [1], Tuğba Çelik Korat [2], Sercan Demirgüneş [3], Seçil Alaca [4], Nida Kırömeroğlu [5]

Abstract

This study focuses on the rubrics used for text evaluation, with the aim of introducing the Reader-Friendly Informative Text Rubric (ODBIMDEPA). This rubric, developed and validated through a reliability study, evaluates the reader-friendliness of informative texts used both in educational settings and in exams administered by institutions such as the Ministry of National Education (MoNE), the Assessment, Selection and Placement Center (ÖSYM), and the Yunus Emre Institute (YEE). The study also seeks to share insights gained from the analysis of evaluated texts. To achieve these goals, the document review method was employed, resulting in the creation of two distinct datasets. The first dataset, analyzed during the initial stage of the study, comprised 50 informative texts taken from the 5th, 6th, 7th, and 8th-grade textbooks used in MoNE schools in 2018. The second dataset, analyzed in the subsequent stage, included 37 informative texts selected from the YKS (University Entrance) exams administered by ÖSYM in 2023 and 2024. In the first stage, two independent academic raters scored the texts, while in the second stage, 11 expert raters for 4 subjects, 13 raters for 1 subject assessed them across eight dimensions defined by ODBIMDEPA. In the initial analysis, the weighted kappa coefficient was used to evaluate inter-rater reliability, indicating that the weighted average was acceptable and demonstrating that ODBIMDEPA exhibited significant consistency. In the second stage, the dataset was analyzed using the many-facet Rasch model, revealing a rater reliability of 0.95. This high reliability underscores the robustness of the developed rubric. Based on these findings, ODBIMDEPA provides a reliable framework for quantifying the reader-friendliness of informative texts used in assessments administered by institutions such as MoNE and ÖSYM, as well as those in educational textbooks. The rubric facilitates a quantitative assessment of whether these texts are suitable for use in exams or educational settings, offering insights into text selection and evaluation.

Keywords

Educational assessment
Informative text
Rubric
Reader-friendly text
ODBIMDEPA

Article Info

Received: 10.13.2024
Accepted: 01.02.2025
Published Online: 03.03.2025

DOI: 10.15390/EB.2025.14166

* A part of this study was presented at the International Symposium on Measurement, Selection and Placement held between 4-6 October 2024 as an oral presentation.

[1] Kocaeli University, Faculty of Education, Department of Turkish and Social Sciences Education, Türkiye, dilek.fidan@kocaeli.edu.tr

[2] Bahçeşehir University, Faculty of Educational Sciences, Department of Primary Education, Türkiye, tugba.celikkorat@bau.edu.tr

[3] Niğde Ömer Halisdemir University, Faculty of Education, Department of Social Sciences and Turkish Education, Türkiye, sdemirgunes@ohu.edu.tr

[4] Hacettepe University, Turkish and Foreign Language Teaching, Application and Research Center, Türkiye, secilalaca@hacettepe.edu.tr

[5] İstanbul University, School of Foreign Languages, Türkiye, nidakiromeroglu@gmail.com

Introduction

"Reader-friendliness" refers to a text's ability to be more easily processed and interpreted by its readers. The term "reader-friendly" was first introduced in the academic literature by Gray and Leary (1935), who identified four key factors contributing to a text's readability: content, style, format, and overall organization. In addition, text type and the presence (or absence) of genre-specific textual features significantly influence how effectively a text is received and processed by readers. Research, such as Brewer (1980) and Smith (1982), divides texts into two primary types: "narrative" and "informational." In the context of text quality evaluation, Çelik, Demirgüneş, and Baştuğ (2014) developed the Reader-Friendly Text Evaluation Rubric (ODMDR) to assess the stylistic and semantic qualities of both informational and narrative texts. Their rubric was informed by the work of Dreher and Singer (1989), who outlined the essential characteristics of "reader-friendly" texts. ODMDR was structured around seven dimensions: Text organization, time usage, coherence-cohesion, signaling, conceptual density, explanation, and discourse. Subsequent studies, such as those by İşeri, Çelik, and Demirgüneş (2015); Çelik, Demirgüneş, and Fidan (2015); Yıldırım and Ocak (2016); Aktaş and Alıcı (2018); and Soyuçok and Kartal (2018), have used ODMDR to evaluate the appropriateness of textbooks for students. These studies explore the suitability of texts from various perspectives, further contributing to the literature on reader-friendly text evaluation. Demirgüneş (2015) argued that each text type has distinct characteristics, making it difficult for the dimensions of the ODMDR to adequately address the specific needs of every text type (p. 67). Therefore, the scoring rubric must be modified to accommodate various text types. Research in the literature also indicates that informative texts employed in educational settings are frequently inferior to narrative texts in both quantitative and qualitative aspects. The findings indicate that ODMDR alone is inadequate for assessing the reader-friendliness of all text types, necessitating the development of scoring rubrics tailored to the specific requirements of each text type (Müldür and Şimşek, 2020).

The extensive classifications of texts in literature have resulted in diverse subtypes of informative texts. Duke and Bennett Armistead (2003) characterize informative texts as works produced exclusively to disseminate information. These texts do not contain characters or narratives but may involve elements such as indexes, photos, footnotes, graphs, and charts. Informative texts are classified as a type of non-fiction writing. Texts designed for informational purposes are categorized into four primary types: *informative, expository, explanatory*, and *argumentative*. Informative texts present encyclopedic material without incorporating commentary or criticism. Expository texts, in contrast, provide information through "explanation and elaboration" as well as "cause-and-effect relationships." These texts generally have a coherent organization, segmenting content into sections and utilizing vocabulary pertinent to the subject matter. Studies (Beck, McKeown, Sinatra, & Loxterman, 1991; Britton & Gulgoz, 1991; McNamara, 2001) indicate that coherent and logically consistent relationships within a text markedly improve readers' capacity to retain and comprehend the information conveyed. Explanatory texts, on the other hand, aim to answer questions such as "*How*?" (e.g., "How does a faucet work?" "How does a computer function?" "How are mountains formed?") and "*Why*?" (e.g., "Why do some objects sink while others float?" "Why is the ozone layer depleting?"). Finally, argumentative texts begin with the identification of the subject matter, followed by sections for discourse, principal arguments, illustrations, or viewpoints, and conclude with an open-ended or inconclusive conclusion.

The Turkish Ministry of National Education (MoNE) classifies text genres in its textbooks into three primary categories: narrative, informative, and poetic, and the Turkish language textbooks are structured according to this classification (MoNE, 2024). In this study, the rubric incorporates four distinct "text types intended for informational purposes"—informative, expository, explanatory, and argumentative—identified in the literature, collectively categorized as "informative text type." As a result, the rubric developed in this study was named as the "Reader-Friendly Informative Text Rubric (ODBİMDEPA)." The mandatory terminological integration of these subtypes illustrates the structural and content attributes of all four text kinds within this category. The principal objective of the developed rubric is to provide a pragmatic and pedagogical assessment instrument for use in educational settings and by stakeholders engaged in the education and training processes, including educators, centralized examination boards, and learners.

To assess the characteristics of the informative text type within the scope of the ODBİMDEPA, the "concrete" and "abstract" features that comprise this text type have been identified and summarized in Table 1, along with their corresponding sources.

Table 1. Concrete and abstract features of informative texts

Basic Qualification	Typical features	Source
Concrete features	**Markers** Headings, abstracts, preface, indicators, tables of contents, agendas (historical organizers of information presentation) ...	Lorch, Lemarie, and Grant (2011) Fisher, Frey, and Lapp (2008)
	Universal qualities Numbers, line spacing, margins	Klink Dengel, and Keininger (2000)
	Engaging content Subtitles, footnotes, bracket explainers, diagrams, graphs, and other descriptive visuals...	Fisher et al. (2008)
	Typographic features Font information (font style, font size... etc.)	Fisher et al. (2008) McNamara, Kintsch, Songer, and Kintch (1996)
Abstract features	**Structure and Rhetoric**	Yeung (2007) Aidinlou, Khodamard, and Azami (2012) Mann and Thompson (1988)
	Coherence	McNamara, Louwerse, McCarthy, & Graesser (2010) Mann & Thompson (1988)
	Lexical and grammatical content Semantic content Syntactic content	Aidinlou et al. (2012) McNamara et al. (2010)

A review of Table 1 indicates that information alone is not sufficient for a text to be classified as informative. The presentation of this information must incorporate concrete elements—such as titles, summaries, graphics, writing styles, and font sizes—that complement the abstract elements, including structure, rhetorical arrangements, coherence, and lexical and grammatical content. This study, based on the literature presented in Table 1 and the updated dimensions of the ODMDR (2014) for informative texts, identified eight dimensions for the evaluation of informative texts in educational settings: *Text Organization, Presentation of information, Formal patterns/Cohesion, Semantic patterns/Coherence, Explanation and reality, Argumentation, Style, and Universality.*

Significance and Aim of the Study:

Although the literature contains a wide range of text classification research, the number of studies that concentrate on rubrics focusing on the evaluation of the quality of texts used in educational settings is limited (Çelik, Demirgüneş, & Baştuğ, 2014; Çıralı Sarıca & Koçak Usluel, 2016). The Reader-Friendly Text Evaluation Rubric (ODMDR), developed in 2014 for the assessment of both narrative and informative texts (Çelik, Demirgüneş, & Baştuğ, 2014), was designed for all text categories. However, it fails to acknowledge the essential distinctions between narrative and informative text types, such as the reduced informational content in narrative texts contrasted with their more robust discourse characteristics. This disparity has underscored the necessity for new assessment instruments specifically designed for informative texts. This study developed the Reader-Friendly Informative Text Rubric (ODBİMDEPA) to assess the formal and semantic appropriateness of informative texts created for or modified for educational contexts. This rubric provides a numerical assessment of the readability of informative texts utilized in examinations administered by organizations like MoNE and the

Assessment, Selection and Placement Centre (ÖSYM), as well as texts featured in educational textbooks. It offers an objective and data-driven approach for evaluating the appropriateness of these texts for use assessment or instruction. In this regard, the main research question of the study is whether ODBİMDEPA is a reliable tool for assessing reader-friendly informative texts. A three-stage study was conducted to answer this question. A rubric was developed in the initial stage. In the second stage, the rubric's reliability was evaluated by having two academic exports rate texts from MoNE textbooks. In the third and final stage, the rubric was employed to evaluate the reader-friendliness of 37 informative texts from the YKS (University Entrance) exam administered by ÖSYM, based on the premise that texts used for assessment purposes must be reader-friendly. Research on reading comprehension (Kintsch, 1982; Kintsch, 1988; Dreher & Singer, 1989; Beck et al., 1991) indicates that comprehension is not solely affected by elements such as the presentation of propositions, information, or syntax; discourse also significantly contributes to the process. Texts used for assessing reading must possess high readability and be reader-friendly. This ensures the development of text-based items and allows for meaningful insights into test-takers' exam performance. Assessment tools including texts that are not reader-friendly are unlikely to provide meaningful results, as test-takers may have cognitive challenges arising from text-related issues throughout the reading comprehension process. The main objective of the final stage of this study was to assess the reader-friendliness of the text rather than rater reliability.

Method

Type of Research

This study aims to develop a rubric for evaluating the reader-friendliness of informative texts and evaluate how reader-friendly the texts used to assess reading comprehension are. The study was structured into three stages as previously described. The criteria, scores, and components of the rubric were defined in the initial stage. In the second stage, the reliability of the rubric developed for the specified purpose was evaluated. In the third stage, the reader-friendliness of 37 texts in the database was assessed using the developed rubric. A database was established to assess the validity and reliability of the rubric, comprising 50 informative texts from MoNE textbooks and 37 informative texts from YKS exam. The document review method, a qualitative research approach, was employed in the development of this database. According to Yıldırım and Şimşek (2008), this approach entails the examination of written documents that encompass information regarding the phenomenon or phenomena under investigation (p. 187).

As part of the method, two distinct sets of documents were generated. The first set comprised the subsequent textbooks used by MoNE in 2018: "Science Textbooks for Grades 4, 5, 6, 7, and 8." "Social Studies Textbooks for Grades 4, 5, 6, and 7," "Human Rights, Citizenship, and Democracy Textbook for Grade 8," and "Turkish Language Textbooks for Grades 4, 5, 6, and 7." The texts featured in these books were reviewed, resulting in a collection of 308 informative texts. Subsequently, 50 texts (16.3% of the collection) were randomly selected from this collection using SPSS (2015).

The second document set was derived from a total of 528 items included in the YKS-Field Proficiency Test and the YKS-Basic Proficiency Test administered by ÖSYM in 2023 and 2024, as published on ÖSYM's official website. From this collection, 37 items featuring instructive texts were selected based on the criteria outlined in the "Limitations" section, and these texts constituted the sample. The sample comprised 10 texts from Turkish Language and Literature, 10 from Philosophy, 7 from Religious Studies, 6 from Sociology, and 4 from History.

Selection of the Raters

The investigation on the reliability of ODBİMDEPA was performed using two distinct rater groups and two separate sets of documents. First, to assess the efficacy of the rubric, the first document set was evaluated by two academic experts, both associate professors in Turkish language education with a minimum of 15 years of professional experience. Subsequently, to assess the reader-friendliness of the texts, a total of 11 expert raters for 4 subjects, 13 raters for 1 subject evaluated the second set of documents using ODBİMDEPA. All 13 raters had graduate degrees in language instruction, and experience in rubric-based assessments, with a minimum of five years of professional experience.

Data Collection Instruments

The *Reader-Friendly Informative Text Rubric/ODBIMDEPA*, which was developed by the researchers to assess whether the informative texts are reader-friendly and is based on ODMDR developed by Çelik, Demirgüneş, and Baştuğ (2014), was used as a data collection instrument in the study. This instument consists of a total of 8 dimensions. These dimensions are respectively "*Text organization, Presentation of information, Formal patterns (Cohesion), Semantic patterns (Coherence), Exposition and reality, Argumentation, Style,* and *Universality*".

An explanation of each dimension in ODBIMDEPA is provided below, and Annex 1 presents the current form of the developed rubric. The rubric considers the following in the investigation of how reader-friendly a text is:

Dimension 1: Text Organization

Informative texts can convey multiple pieces of information or concepts. All information and concepts within the text must be coherent and mutually supportive, avoiding contradictions. Inconsistent statements, opposing perspectives, or a series of disconnected information can alienate the reader and decrease their capacity to engage with and comprehend the text. The central concept must be prominent throughout the text, with all auxiliary details and ideas supporting it. If the primary idea or concept does not permeate the text and is masked by irrelevant details or thoughts, the text may deviate from its intended purpose. A convergence of diverse information, opinions, and perspectives is essential to establish a coherent main idea.

Dimension 2: Presentation of Information

Two approaches are used to present information in informative texts. In the initial approach, a "**single main piece of information/idea**" is conveyed, and the reader is assisted in conceptualizing the main information through the use of various techniques for developing ideas, such as elaboration, comparison, and explanation. The second method involves the presentation of a "**bundle of information/ideas.**" In this method, the author makes the primary information the dominant element, while other pieces of information are positioned as its supporters. Concrete clarifications and explanations regarding the presence of the primary and supporting information must be provided when presenting information. Furthermore, the organization of information should be step-by-step, progressing from general to specific and from simple to complex, and it should be classified or grouped in a manner that enables the reader to comprehend it.

Dimension 3: Formal Patterns (Cohesion)

In terms of structure, informative texts are distinguished from other text types by their unique characteristics. The text should contain a section that presents theoretical/conceptual information related to the intended message, another section that develops or discusses this information, and, finally, a section that provides clues for resolving the issues or questions mentioned in the text, even if it is not explicitly divided into sections with subheadings.

The reader should be able to distinguish the primary information and its supporting structures within these sections by using elements such as font styles and sizes. Furthermore, it is feasible to incorporate various types of explanatory visuals, including tables and graphs. Linguistic devices such as anaphora and cataphora, ellipsis, pronominal drop, and repetition may also be employed, depending on the typological capabilities of the language.

The primary objective is to employ all available linguistic clues to assist the reader in comprehending the primary information. The text can be made more comprehensible and reader-friendly by highlighting key words in a variety of ways to remind the reader of the content, given that summarizing, interpreting, and comprehending informative texts is more difficult than other types of text.

Dimension 4: Semantic Patterns (Coherence)

In addition to formal coherence, there should also be semantic coherence in the text. If the presentation of information is managed by conveying **"only one main piece of information/idea"**, relationships such as exemplification, comparison, evidence-citing, definition, etc. should be interwoven with arguments supporting the main piece of information/idea. Achieving semantic coherence this method of I information presentation is relatively easier compared to presenting a **"bundle of information/ideas"**. However, presentation of the main piece of information by supporting or reinforcing it with "**a bundle of information/ideas**" could prove a more effective way. In this case, other information/ideas that will support the main information in the text should be compatible with each other and should not contradict the main piece of information/idea.

Dimension 5: Exposition and Reality

Every piece of information or idea presented to the reader in a text (particularly words that the reader is assumed not to know) must have an explanation/definition inside or outside the text (footnotes, end-of-text references, dictionary, etc.). These explanations/definitions must be given in a way that will not create new uncertainties for the reader, through concretization, exemplification, etc. In informative texts, the most effective way to facilitate the understanding of information pieces/ideas is to reach the reader's world and to bring the information/idea into the reader's own reality/world. Therefore, the writer must integrate the main information in the text with the reader's world knowledge.

Dimension 6: Discussion

Every piece of knowledge originates from either a sense of curiosity or a coincidence. It is possible to access information without coincidence during the instruction process by arousing curiosity in the reader. The bundle of information or ideas bundle presented in the text can arouse the reader's curiosity at varying levels. The role of the writer is not to leave the reader amidst conflicting ideas or debates, but to create a discussion about the main information in the text and to arose reader's curiosity without necessarily providing a clear or single resolution. At this point, there is a need for multiple perspectives and viewpoints in the text to foster discussion. In this context, the discussion does not necessarily have to reach a "**single resolution**"; rather, it should aim to trigger a feeling of curiosity. Discussion stimulates curiosity; and curiosity serves as the most critical motivating factor to comprehend the main information/idea.

In this dimension, the main point to consider is ensuring that the reader evaluates the main information using different pieces of information/ideas. On the other hand, the dispersion of the main information/ideas, attempts to conclude the discussion with a single solution, and the creation of texts that present a single solution/idea are indicators of a weak discussion.

Dimension 7: Style

The primary objective of all text types is to establish a robust relationship between the author and the reader. In informative texts, the author ought to earn the reader's trust by providing objective perspectives and the relevant information. The author of an informative text is required to maintain objectivity by relying on factual and scientific foundations. In this context, the author is required to interact with the reader's world without employing subjective constructs. The text's credibility is undermined by the use of sharp and subjective expressions such as "in my opinion," "regardless of what others say," "not open to discussion," "the most important," or "the only condition." Conversely, the use of precise language, including "according to another perspective," "numerically speaking," "if we take a realistic perspective," "evidence indicates that," "more than 1000 meters," or "the only migratory bird among them," illustrates proficiency and precision in the language. Additionally, the author should establish a connection between the reader and the text by employing summarizing phrases, such as "in conclusion," "as previously mentioned," or "according to the reports." Furthermore, the reader is compelled to engage with the text by the use of compelling examples. The author's use of linguistic expressions such as "Another point is (…)," "(…) in winter temperatures (…)," "For Norwegians (…),"

or "In Ottoman culture as well (…)" suggests that the writer employs a language that is open to different cultures, works to be easily understood, relies on data-supported evidence, and effectively employs comparisons.

Dimension 8: Universality

Knowledge is a universal phenomenon that remains constant in the presence of a variety of variables. Using terminology that is consistent with international agreements (e.g., Human Rights, Children's Rights, etc.) is beneficial for the development of a universal perspective in informative texts. For example, the author should refrain from drawing discriminations based on language, religion, gender, or other similar factors, as these distinctions may result in readers feeling alienated as a result of negative expressions in the text that contradict their culture, beliefs, or gender identity. Furthermore, the author should exhibit a perspective that fosters curiosity regarding various human behaviors and cultures. For example, the text could provide information about countries with populations with particularly high life expectancies, highlight the rights of various animals, emphasize multiculturalism, or address topics such as geography and botany while emphasizing sustainability or environmentalism. It could also include details about diverse regions within a country and the characteristics of their inhabitants. The author may incorporate a universal idea or piece of knowledge into a framework of non-universal knowledge or ideas. Conversely, the author should avoid using expressions that cultivate prejudices against individuals, normalize any form of discrimination or negative emotions, thoughts, or knowledge, or incorporate statements that could potentially foster discriminatory attitudes.

Validity and Reliability of the Data Collection Instrument

According to Tuncel (2011), rubric reliability is defined as "the assignment or performance of a student being evaluated receiving the same score in every evaluation and from each rater" (p. 222). The validity of a rubric should be evaluated in terms of "content, structure, and criteria," according to Moskal and Leydens (2000). The referenced study explains these validity aspects as follows:

***Content*:** Whether the assessment criteria identify any off-topic content; whether the criteria identify all aspects of the intended content; and whether there are any unidentified content areas of the assignment or activity to be assessed with the rubric.

***Structure*:** Whether all important aspects of the intended structure are evaluated by scoring criteria, and whether all evaluation criteria are related to the relevant structure.

Criteria: How the rating criteria reflect relevant performance or elements of future success; whether there are any aspects of relevant performance that are not reflected in the rating criteria.

The validity ODBİMDEPA was also checked using the three-stage process based on the aspects of "content, structure and criteria" proposed by Moskal and Leydens (2000):

In the first stage, the criteria were developed based on studies (Gray & Leary, 1935; Dreher & Singer, 1989; Nagabhand, Nation, & Franken, 1993) that investigated the characteristics of reader-friendly texts, which ensured that the content aspect was checked.

The second stage, as mentioned previously, was carried out with two academic experts rating a total of 50 texts in the first document set (informative texts in Science; Social Sciences; Human Rights, Citizenship and Democracy; courses and Turkish language course books used by MoNE in 2018) to check the functionality of the rubric, and the reliability results of the developed rubric were obtained by performing Cohen's Kappa (Weighted kappa) analyses due to the ordinal nature of the scoring in the rubric. After the completion of scoring, feedback was received from the raters regarding the content and structure, and then the study was presented as a report at UTEOK 2018 and opinions were received from the audience. The content of the rubric was revised based on the feedback, thus the instrument was improved in terms of content, structure and criterion.

The third stage was conducted with 11 raters in the second research group using a total of 37 exam texts (informative texts in the exams of *Turkish Language and Literature; Philosophy; Religious Studies; Sociology* and *History* courses selected from YKS-TYT (Field Proficiency Test) and YKS-AYT (Basic Proficiency Test) in 2023 and 2024), in the second document set, and two different statistical approaches were applied to evaluate the reliability among raters in the analysis of the scoring results: "Fleiss' kappa" performed with the *irr* package in R and "many-facet Rasch model" performed with the TAM package.

According to Eser and Aksu (2022, pp. 58-59), Fleiss' kappa is a method to investigate the level of agreement between three or more raters, and since it is influenced by prevalence and bias, it can lead to the paradox of high agreement but low kappa value. In addition, in the same study, it is stated that the ranges regarding the results obtained within the scope of Fleiss' kappa are used in the interpretation of the results regarding Cohen's kappa.

Fleiss (1971) states that this method shows how consistently raters evaluate categorical data and is suitable for measuring how individuals score rubrics based on specific criteria. The Kappa statistic takes into account agreement that might occur by chance, and results range from -1 to 1. **1** indicates perfect agreement, **0** indicates agreement by chance, and **negative values** indicate agreements less than agreements due to chance.

On the other hand, the many-facet Rasch measurement model (MFRM) is a linear model that converts observations on an ordinal scale to an equally spaced logit scale by calibrating all parameters (Bond & Fox, 2015).

The Rasch model has a number of superior features compared to traditional measurement methods (Rasch, 1960; Linacre, 1993). These features can be summarized as follows: Instead of raw scores, the many-facet Rasch model uses measurement values (cleared of measurement errors) belonging to individuals. MFRM also determines the rules for a linear connection that should exist among the facets for each facet (e.g., the quality of scientific research assignments, characteristics of assessment questions, rater behavior in this study) and creates these connections (Hambleton and Swaminathan, 1985). MFRM standardizes the facets by combining them on a common scale to obtain an unbiased and effective measurement and offers the opportunity to compare individuals' scientific research skills, the difficulty of the questions, and the "strictness" or "generosity" behaviors of the raters at the same time. In order to determine whether the texts used as measurement tools are reader-friendly, which is one of the main purposes of the study, the focus was on the consistency of the "severe" or "lenient" scoring of the raters throughout the evaluation process rather than the agreement of the raters with each other. Since no training was given before the use of the rubric and no standardization study was conducted, the low agreement between the raters was ignored and the evaluation was based on the consistency of severity/leniency.

Data Collection

During the presentation of the study at UTEOK 2018, informative texts from 5th, 6th, 7th and 8th grade textbooks (Science, Human Rights and Citizenship, Social Studies, Turkish) were used; the expressions and explanations of the dimensions of ODBİMDEPA were improved and updated based on the feedback and suggestions from the raters of the rubric.

In the second stage of this study, firstly, informative reading texts (from ÖSYM's 2023 and 2024 YKS exams) were selected for the validity of ODBİMDEPA by document review method as explained above. Subsequently, each of the raters selected by convenience sampling method was given "ODBIMDEPA User Guide" in advance and informed about how to score the texts using the rubric. Then, each rater was given evaluation forms including the texts they would score and the 8 dimensions of the rubric.

After the scoring, the data of all raters were brought together and analyzed in order to perform the necessary statistical analyses.

Data Analysis

Since the scoring was performed by two raters in the first stage, the Cohen's Kappa (Weighted Kappa) method, which provides the values related to the reliability of the two raters, was selected. Şencan's (2005) suggestions were used in the interpretation of the findings related to the analyses conducted with Cohen's Kappa:

No agreement	*= < .20*
Non to slight	*=.20-.40*
Moderate	*=.40-.60*
Substantial	*=.60-.80*
Almost perfect agreement	*= .80-1.00*

The coefficient ranges above indicate that reliability between the raters is present when the coefficient is.20 or higher. Reliability is considered to be at its highest level when the coefficient is between.80 and 1.00.

In the Cohen's Kappa method, agreement between raters is achieved when both raters assign the same score. Nevertheless, in ODBMR, the disagreement that arises when one rater assigns a score of 2 and the other assigns a score of 3 is distinct from the disagreement that arises when one rater assigns a score of 1 and the other assigns a score of 4. When the raters' scores were analyzed, it was observed that the level of inconsistency was mostly a difference of 1 point in the scores. Consequently, modifications were made to enhance the construct validity in response to the raters' feedback following the initial evaluation process. It was subsequently decided that a reliability study with an increased number of raters would be conducted. Furthermore, the objective was to evaluate the average scores of the texts and to explore the general trend regarding the degree of reader friendliness.

The second stage of the study involved the selection of additional raters and the scoring of the texts from YKS exams using the rubric, as determined by the aforementioned evaluations. The Fleiss Kappa model was chosen in situations where there were more than two raters who scored, as the Cohen's Kappa (Weighted Kappa) method was not feasible for reliability purposes. The *irr* package in R was employed to perform Fleiss Kappa analyses. It was noted that the raters' reliability was not at the desired level. In the Fleiss Kappa method, which is comparable to Cohen's Kappa, the raters' harmony is maintained by each assigning the same score. In the second set of documents, the score differences between the raters were primarily 1 point, similar to the scoring in the first stage. The low score range in the reliability studies conducted for the rubric is also a significant indicator. Determining the internal consistency and "severity" / "leniency" tendencies of the raters, rather than the consistency between the raters, was deemed appropriate for this reason. Additionally, the texts with average scores should be examined if the internal consistency is high. MFRM analysis was conducted using the *TAM* package in R to obtain the values associated with the internal consistency of the raters. MFRM incorporates variability across raters, in contrast to Fleiss' kappa. It also considers factors such as the "leniency" or "severity" of raters when predicting measured performance (Rasch, 1960). This implies that the model accounts for any discrepancies between raters, which leads to increased reliability values.

Another perspective on the reliability of the rubric was provided by the high predictive reliability of the model achieved with MFRM. Correlations were reviewed to investigate the EAP reliabilities and inter-rater consistency in the ratings of various subject areas.

To examine the relationships between raters, texts, and dimensions, a three-facet MFRM analysis was conducted using the FACET program, with raters, texts, and dimensions serving as the facets.

In order to understand raters' leniency-severity tendencies, a rater severity index (RSI) was calculated for each rater.

RSI=Mean score - overall average / overall standard deviation

The objective was to determine the extent to which the ratings of each rater deviated from the overall average by converting their average score into a z-score. Z-score standardization enabled the comparison of rater bias on a common scale (Bond & Fox, 2015; Field, 2013).

Once it was established that the raters are consistent, the goal was to assess whether the texts were reader-friendly. To achieve this, the average of the raters' scores (M) was calculated to assess the reader-friendliness of the texts. The scores provided by the raters in each dimension for each text were averaged, followed by the calculation of the overall average across all dimensions. The overall score was calculated by summing the average ratings of the evaluators across each dimension.

A text scored using ODBİMDEPA can receive a maximum of 4 points from each dimension and a maximum of 32 points in total. The total score of a text can reveal whether an informative text has the characteristics of a "reader-friendly" text by reviewing the following score ranges:

25 points and above **:** *Reader-friendly text*

16-24 points **:** *Moderately reader-friendly text*

15 points or less **:** *Text that is not reader-friendly*

The scores listed above were established by calculating percentages derived from the maximum and minimum possible scores for each dimension. A text that scores 1 point from each dimension totals 8 points; a text that scores 2 points from each dimension totals 16 points; and a text that scores 3 points from each dimension totals 24 points. In assessment systems, a score of 4/4 is viewed as excellent, 3/4 signifies success and adequacy, 2/4 reflects a moderate level of success, and 1/4 is regarded as insufficient. According to this framework, a text that achieves 25 points or more on ODBİMDEPA is deemed "reader-friendly," while a text scoring 15 points or less is regarded as not reader-friendly.

Limitations

The primary limitation of the study is that it focused solely on informative texts, excluding narrative texts from its scope. The second limitation relates to the criteria for excluding the document sets. In the initial document set, poetry texts, texts with visuals, and incomplete or unfinished texts were omitted. The second document set excluded texts with paragraphs that required fill-in blank completions, single-sentence paragraphs, visuals, poetry, disrupted flow or inserted sentences, paragraphs solely consisting of definitions, bullet-point information, and items from the Religious Studies exam that were entirely composed of verses from the Quran.

Results

This study centers on informative text rubrics and introduces ODBİMDEPA, a rubric designed to assess reader-friendly informative texts, while also offering insights into its application in educational, instructional, and assessment settings. The results derived from the procedures carried out to assess the validity and reliability of ODBİMDEPA for these objectives are detailed sequentially below:

1. Cohen's Kappa (Weighted Kappa) analysis findings obtained from the first study group

The weighted average kappa value for the scores assigned to the informative texts in 5th, 6th, 7th, and 8th-grade textbooks across ODBİMDEPA's eight dimensions was calculated as 0.53 [95% CI (0.46, 0.60)] in the study's first group (two independent academic raters). A total of 50 texts were utilized for scoring.

The Cohen kappa coefficients derived from the initial study group suggest a moderate degree of agreement, as noted by Cohen (1960). In this context, it was noted that ODBİMDEPA demonstrates a balanced and consistent distribution—indicating moderate reliability—but attaining a greater level of reliability would necessitate the inclusion of additional raters. Consequently, the researchers decided to carry out the study with an additional group of raters. The results from the second study group will be presented in the following sections.

In addition to the general agreement between the raters, the agreement of the raters in each dimension was also examined. The findings regarding the kappa values obtained for each dimension of ODBIMDEPA are presented in Table 2.

Table 2. Cohen's (Weighted) Kappa Coefficients by dimensions

Dimension	Criteria	Cohen (Weighted) Kappa Coefficient
1	Text organization	0.20
2	Presentation of Information	0.17
3	Formal Patterns/Cohesion	0.34
4	Semantic Patterns/Coherence	0.53 *
5	Explanation and Reality	0.55 *
6	Argument	0.60 *
7	Style	0.57 *
8	Universality	0.14

* p<.05

Table 2 reveals that the inter-rater reliability for each dimension is notably low, reflecting minimal agreement for the 2nd and 8th dimensions, whereas moderate agreement is noted for the 1st and 3rd dimensions. While the Cohen's Kappa coefficients for two dimensions show low rater reliability, the agreement among raters is satisfactory for the other dimensions.

1.1. Distribution of the Mean Scores from the Informative Texts in the First Set of Documents by Grades

Table 3 presents the number of texts along with their mean scores categorized by grade level in the database derived from the first document set.

Table 3. Distribution of Mean Scores of Informative Texts by Grades

Grade	Number of Texts	Mean Score
Grade 4	N=16	M= 2.3
Grade 5	N=8	M= 2.4
Grade 6	N=11	M= 2.3
Grade 7	N=11	M= 2.0
Grade 8	N=4	M=2.2
SUM	**50**	**2.24**

Table 3 demonstrates that the highest number of informative texts (N=16) appears at the 4th-grade level, whereas the least amount (N=4) is found at the 8th-grade level. Conversely, the highest average score stands at 2.4 in the 5th grade, while the lowest average score is 2.0 in the 7th grade. These findings allow for an examination of whether the readability and, as a result, the reader-friendliness of texts vary across grade levels, taking into account that the texts are tailored for different age groups. Although one might anticipate that texts aimed at younger audiences would be more organized and easier to understand, the average scores suggest that the degree of reader-friendliness remains relatively consistent across different grade levels.

1.2. Distribution of the Mean Scores from the Informative Texts in the First Set of Documents by Courses

The distribution of a total of 50 texts in the first set of documents according to the courses and the mean scores of each course are presented in Table 4.

Table 4. Distribution of Mean Scores from Informative Texts by Courses

Lesson	Number of Texts	Mean Score
Science	N=17	M=2.2
Human Rights and Citizenship	N=2	M=1.9
Social Studies	N=15	M=2.2
Turkish	N=16	M=2.4
SUM	**50**	**2.18**

Table 4 suggests that the informative text type is used most frequently in Science with 17 texts and in Turkish with 16 texts, while it is used the least frequently in Human Rights and Citizenship course (N=2). Although the number of informative texts used in Science course is higher than the number of informative texts used in the Turkish language course, the Turkish language course has the highest average score (2.4). On the other hand, the Human Rights and Citizenship course, featuring the least amount of informative texts, records the lowest average score at 1.9.

1.3. The Mean Scores of All Informative Texts in the First Set of Documents According to ODBIMDEPA Dimensions

The mean scores of the total 50 texts in the first document set from 8 dimensions of ODBİMDER are presented in Table 5.

Table 5. Mean scores of all informative texts by dimensions

Dimension	Mean Score
1- Text organization	2.6
2- Presentation of information	2.24
3- Formal patterns/Cohesion	2.17
4-Semantic patterns/Coherence	2.17
5- Explanation and reality	2.31
6- Argument	1.37
7- Style	1.94
8- Universality	2.89
SUM	**2.22**

Table 5 reveals that the best performing dimension in the first version of ODBIMDEPA in 2018 was the "Universality" dimension with a mean score of 2.9 and the "Text Organization" dimension with a mean score of 2.7. On the other hand, it was found that the least functioning dimension with a mean score of 1.3 was "Discussion" and "Style" with 1.9.

2. Findings of Fleiss' Kappa and MFRM approaches from the second study group

The second study group (eleven independent raters for 4 subjects, 13 independent raters for 1 subject) scored a total of 37 informative texts in the 2023 and 2024 YKS tests of ÖSYM as per the 8 dimensions in ODBİMDEPA. In the analysis of the scoring results, two different statistical approaches were used to assess the **inter-rater reliability**: **Fleiss' kappa**, performed with the *irr* package in R, and the **MFRM** performed with the *TAM* package. These two methods approach the data from different perspectives, and their results should be interpreted in the context of the assumptions and strengths of each model. Below are explanations of the findings, their effects on validity and inter-rater reliability, and explanations on how these values should be interpreted.

2. 1. Fleiss' Kappa: Assessing Consistency among Raters

Table 6 presents the Fleiss' kappa results obtained from the rating of the informative texts in the ÖSYM database consisting of YKS exams using ODBİMDEPA.

Table 6. Fleiss' Kappa Values of Informative Texts in YKS Exams Scored using ODBIMDEPA

Text Content	Kappa Value
Religious studies exam texts	0.09 *
Philosophy exam texts	0.03 *
Sociology exam texts	0.02
Turkish Language and Literature exam Texts	0.05 *
History exam texts	0.03 *

* p<.05

According to Table 6, obtained Fleiss' kappa values ranging between 0.02 and 0.09 indicate that there is no agreement between the raters. Although some results are statistically significant ($p < 0.01$), low kappa values indicate that the raters are not consistent in applying the rubrics in the same way. The kappa value of 0.09 (texts from the Religious Studies exam) is quite low and does not show a significant difference from chance-based agreement. In some cases (Sociology, kappa = 0.02), the agreement is not statistically significant, which shows that the agreement between the raters is random. The explanation made for the inconsistencies regarding the first stage also applies here. It should be taken into consideration that the degree of inconsistency of the raters is also important in a 4-point and 8-criteria scoring key. For example, in the scoring of the YKS Sociology section 1st text by 11 raters in Table 7, it is observed that the raters generally scored closely for the 8 dimensions.

Table 7. Rater (R) Scores of the Informative 1st Text in the YKS Sociology Test According to ODBIMDEPA

Dimension	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11
1	4	4	4	4	4	4	2	4	4	4	4
2	3	4	3	3	4	4	2	3	4	4	3
3	4	3	2	4	2	4	3	4	4	4	2
4	4	3	3	4	3	4	2	4	4	4	3
5	3	3	3	3	4	4	3	3	4	3	4
6	4	4	2	4	4	2	1	3	4	4	3
7	2	4	3	4	3	1	1	4	4	4	3
8	4	3	4	3	4	3	4	3	4	4	4

As can be seen in Table 7, all raters scored 4 points to the 1st dimension except R7. In the 2nd dimension, the raters generally scored 3 and 4 points, and there is a similar scoring tendency in the 3rd, 4th, 5th and 8th dimensions. The score range is broader in the 6th and 7th dimensions.

2.2. Many-Facet Rasch Model: Correction of Rater Effects

In the analyzes, similar results were obtained in all branches (Turkish Language and Literature, Philosophy, Sociology, Religious studies, History). Table 8 shows the many-facet Rasch results:

Table 8. Many-facet Rasch Results of Informative Texts in YKS Tests Scored According to ODBIMDEPA

Many-facet Rasch Elements	Rasch value
EAP Reliability	0.95
Variance	0.07
Standard deviation	0.27

A review of Table 8 suggests that the EAP reliability (A Posteriori Expected Reliability) of **0.95** shows that the model can predict the texts' compliance with the reader-friendliness criteria **with high reliability**. The variance (0.07) and standard deviation (0.27) values show that there is some variation between the scores, but the model is able to **compensate for these differences** and thus **obtain reliable estimates**.

This result shows that the Rasch model can make **valid inferences** in the evaluation of texts even if there are inconsistencies between raters **in terms of scoring**. In other words, the Rasch model assumes that the rubric raters are inconsistent but compensates for these inconsistencies by modeling the biases of each rater. This analysis suggests that the scores of the evaluated texts regarding whether they are reader-friendly are reliable measures despite the rater variances.

With the many-facet Rasch analyses, EAP reliabilities and correlations were also determined according to the scorings made by the raters for each dimension in the texts related to different YKS disciplines.

Table 9. Rater reliability of YKS texts according to disciplines

YKS Texts	EAP	Correlation
Sociology	0.88	0.45
History	0.95	0.27
Turkish language and literature	0.88	0.40
Religious culture and moral knowledge	0.92	0.62
Philosophy	0.92	0.59

Table 9 shows that EAP reliability is high in texts in all disciplines. The high EAP reliability shows that the model can make reliable predictions. The reliability among the raters, considering all raters, shows that the correlation is particularly low in history texts. Upon examining the data set, it is found that two raters are “outliers”. The low correlation values are influenced by these raters.

The map showing the general situation of the evaluators regarding the texts, dimensions and their scoring in different disciplines is shown in Figure 1. Two raters who scored only Turkish language and literature texts were not included in this analysis and the relationship of 11 raters with all texts and dimensions was examined. The analyses were conducted using the FACET program for the following subjects: History (T), Philosophy (F), Sociology (S), Religious Studies (D), and Turkish Language and Literature (TE). The reliability for the rater dimension was calculated as 0.98, with a separation index of 6.60.

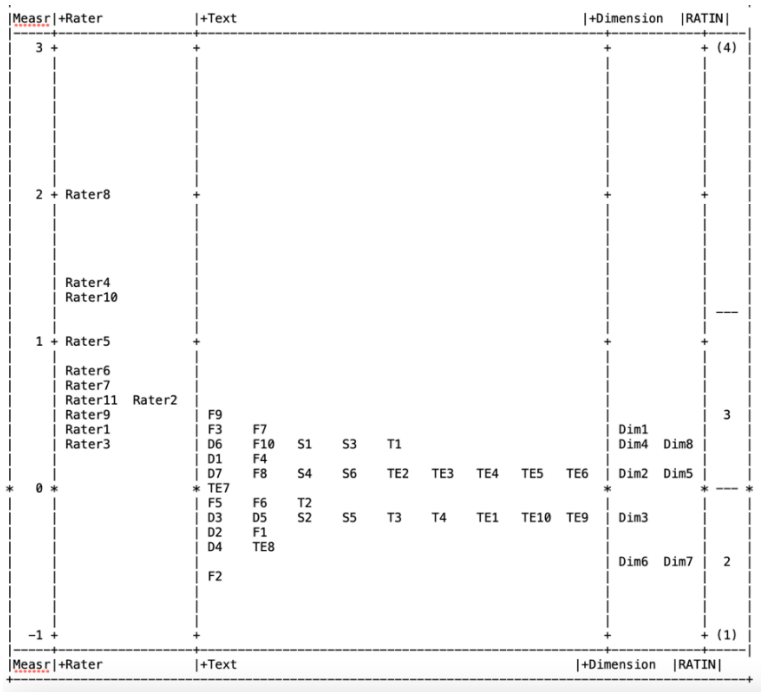

Figure 1. Rater, text, dimension correlation map

Figure 1 shows that the logit values for the raters have varied between 0 to 2. Among the texts, it is evident that F2, one of the Philosophy texts (F1-F9), received lower scores compared to the others and remained below the average. In contrast, F3, F7, and F9 achieved higher average scores, indicating a greater level of reader-friendliness. Additionally, the raters were observed to assign higher scores in dimension 1, lower scores in dimensions 6 and 7, and generally preferred scores of 2 or 3 across all dimensions. This tendency to avoid the lowest and highest scores may be attributed to the rubric's focus on abstract evaluations of text qualities and the raters' inclination to remain within a "safe zone." Although Figure 1 shows that raters generally tend to score more generously, examining individual rater tendencies in terms of severity and leniency is crucial for understanding their consistency and identifying potential outliers.

To determine the individual tendencies of the raters, the average score of each evaluator was converted into a z-score, and their severity/leniency tendencies analyzed. Table 10 presents the values of raters' severity/leniency tendencies.

Table 10. Raters' severity/leniency values: Rater Severity Index (RSI)

Rater	Sociology	History	Religious Studies	Philosophy	Turkish Language and Literature
R1	-0.48	-0.68	-0.36	-0.01	-0.36
R2	-0.12	0.64	-0.25	-0.58	-0.05
R3	-0.57	-0.49	-0.32	-0.47	-0.37
R4	0.62	0.29	0.02	0.29	0.45
R5	0.10	0.14	0.28	0.20	-0.12
R6	-0.01	-0.04	-0.05	-0.10	-0.04
R7	-0.34	0.29	0.24	0.32	-0.62
R8	0.55	0.67	0.24	0.64	0.69
R9	-0.03	-0.52	0.36	-0.70	-0.44
R10	0.53	0.42	0.55	0.16	0.03
R11	-0.24	-0.74	-0.70	0.25	0.03
R12	-	-	-	-	0.37
R13	-	-	-	-	0.42

Table 10 shows the values reflecting each rater's inclinations towards severity or leniency when scoring texts across various subject areas. As the values move further away from 0, the inclinations towards severity or leniency grow, with negative values reflecting severity and positive values reflecting leniency. Upon examining Table 10, it is noted that R5 typically shows a lenient tendency, yet they display severity in their evaluation of Turkish Language and Literature texts. Nonetheless, in both instances, the values stay near 0. Thus, it can be concluded that this transition does not considerably harm the consistency of R5's severity/leniency tendency. It is not the case for R7 and R9. The tendencies of these two raters regarding severity and leniency seem to fluctuate unpredictably depending on the subject area of the texts. As noted earlier, these raters might be viewed as "outliers." While R11's tendencies vary across subject areas, the situation is not the same as that of R7 and R9. Although R11 typically exhibits a severe tendency, they display a degree of leniency in their evaluation of philosophy texts. While R11 shows a lenient tendency in Turkish Language and Literature texts, their RSI value stays near 0, suggesting that the effect on the consistency of their severity/leniency tendency is not notably negative. Nonetheless, the tendency towards leniency in philosophical texts does seem to influence consistency to a certain degree. However, taking into account their general trend, R11 ought not to be categorized as an "outlier" in the same way as R7 and R9. For R12 and R13, who evaluated texts from just one subject area, it is not feasible to provide a significant observation about their tendencies towards severity or leniency.

2.3. Mean Scores of Texts

This section presents results concerning the mean and total scores of informative texts featured in the YKS exams texts for Sociology, History, Turkish Language and Literature, Religious studies, and Philosophy.

Six informative texts (N=6) were identified from the YKS Sociology exam and scored by 11 raters (R=11). Table 11 presents the findings from these ratings.

Table 11. Mean scores of sociology exam texts

YKS Sociology Texts	Mean Score	Total Score
1	3.38	27.09
2	3.09	24.72
3	3.42	27.36
4	3.28	26.27
5	3.11	24.89
6	3.26	26.09

Table 11 suggests that all informative texts in the sociology test have a score of 25 or above and can be classified as reader-friendly.

A total of four informative texts (N=4) were selected in the YKS History exam and scored by 11 raters (R=11). The results of these ratings are given in Table 12.

Table 12. Mean scores of history test texts

YKS History Texts	Mean Score	Total Score
1	3.40	27.27
2	3.12	25.09
3	3.07	24.63
4	3.06	24.54

A review of Table 12 suggests that it is possible to classify the informative texts in the History exam as reader-friendly, similar to the texts in the Sociology exam.

A total of ten informative texts were selected from the YKS Turkish Language and Literature exam and scored by 13 raters. Table 13 presents the results of these ratings.

Table 13. Mean scores of Turkish Language and Literature exam texts

YKS Turkish Language and Literature Texts	Mean Score	Total Score
1	3.13	25.08
2	3.32	26.54
3	3.40	27.23
4	3.39	27.15
5	3.35	26.77
6	3.38	27
7	3.24	25.92
8	2.93	23.46
9	3.16	25.31
10	3.22	25.77

The Turkish Language and Literature exam features a greater number of informative texts than any other test, with the exception of the Philosophy exam. Based on Table 13, every scored text is categorized as reader-friendly.

Furthermore, seven informative texts (N=7) were selected from the YKS Religious Studies exam and scored by 11 raters (R=11). Table 14 presents the results of these ratings.

Table 14. Mean scores of Religious Studies exam texts

YKS Religious Studies Texts	Mean Score	Total Score
1	3.37	27
2	3.03	24.27
3	3.10	24.81
4	2.93	23.45
5	3.09	24.72
6	3.40	27.22
7	3.27	26.18

Upon reviewing Table 14, it becomes obvious that the third text in the Religious Studies exam is categorized as borderline reader-friendly, whereas the fourth text is designated as "moderately reader-friendly." All the other texts can be classified as reader-friendly.

Furthermore, ten informative texts (N=10) were selected from the YKS Philosophy exam and scored by 11 raters (R=11). The results from these assessments are shown in Table 15.

Table 15. Mean scores of Philosophy exam texts

YKS Philosophy Texts	Mean Score	Total Score
1	2.97	23.81
2	2.81	22.54
3	3.44	27.54
4	3.35	26.81
5	3.18	25.45
6	3.17	25.36
7	3.46	27.72
8	3.30	26.45
9	3.52	28.18
10	3.43	27.45

The philosophy exam has the same number of informative texts as the Turkish language and literature exam and more than the other tests. Table 15 reveals that the 1st and 2nd texts are in the "moderately reader-friendly" text class, while the other texts are in the "reader-friendly" text class.

Discussion and Conclusion

The present study aimed to introduce the *Reader-Friendly Informative Text Rubric-ODBIMDEPA,* which was developed to investigate whether the informative texts used both in educational environments and in the exams of institutions such as MoNE, ÖSYM, YEE, are reader-friendly, and to report the findings about the texts analyzed.

ODMDR, first developed by Çelik, Demirgüneş, and Baştuğ (2014), was designed to be used in all text types. As this rubric (ODMDR) was used by different researchers in the literature, the feedback obtained led to the idea that separate rubrics could be developed for different types. One of the seven dimensions of ODMDR, "text organization", has been preserved in ODBİMDEPA, but its content has been customized for informative texts. The "time usage" dimension in ODMDR has been removed; because in informative texts, time periods are not clear and followable/absent as in narrative texts. The "Coherence-Cohesion" dimension, which was presented by combining them in the same dimension in ODMDR (Çelik, Demirgüneş, & Baştuğ, 2014), has been divided into two dimensions, "Formal Patterns (Cohesion)" and "Semantic Patterns (Cohesion)", and has been expanded in more detail in ODBİMDEPA. The reason for this is that cohesion and coherence are achieved with different strategies and linguistic expressions in narrative texts, and it is necessary to describe how these two elements can be made reader-friendly in informative texts. Similarly, the dimension titled "signaling" in ODMDR was excluded in ODBIMDEPA, and the elements of this dimension for informative texts were transferred to the aforementioned "Formal Patterns/Cohesion" dimension. The "conceptual density" dimension in ODMDR was not used in ODBIMDEPA, and instead the "presentation of information" dimension was added, detailing how the presentation of information in informative texts can make the text reader-friendly. The "explanation" dimension in ODMDR was preserved, but the "reality" dimension was added because the perception of reality in an informative text should be strongly activated in the reader. The last dimension of ODMDR, "discourse", was presented as the "style" dimension in ODBIMDEPA. The "universality" and "argumentation" dimensions introduced to ODBIMDEPA were not originally present in ODMDR.

The first version was presented in 2018 and later, with feedback from raters, the dimensions of the ODBİMDEPA were updated with more comprehensible expressions in terms of form and content. Thus, it has become clearer to future users of the rubric what they should pay attention to when evaluating any informative text according to the dimensions of the rubric.

In the next section, the findings from each of the two stages of the methodology are reported and discussed.

Comparison of Findings: Kappa and Rasch Model

The main difference between Fleiss' kappa and Rasch model is how they handle rater variability. Fleiss' kappa directly measures the agreement of raters and assumes that raters should be consistent with each other. The results show that raters have low agreement and are inconsistent in applying the rubrics.

On the other hand, Rasch model still achieves **high reliability** by accounting for the biases of each rater. This shows that even if raters are not consistent in scoring, Rasch model **compensates** for these differences and produces reliable estimates.

As mentioned earlier in the study, the score differences among evaluators typically float around 1 point, and the score descriptors in the developed rubric are associated with abstract traits, complicating the rating process. Considering these factors, it was concluded that the analysis results derived from MFRM are **satisfactory** and that the scores given to the texts can offer meaningful insights.

Conclusion and Discussion on the First Set of Documents

Upon examining the database of textbooks for 5th, 6th, 7th, and 8th grades used in courses by MoNE in 2018, it was found that the allocation of informative texts among the grades was not balanced. Consequently, out of the 50 informative texts in the database, it was observed that the classification system was inadequate and required revision. The program should also incorporate distinctions based on subgenres found in the literature, including descriptive, explanatory, argumentative, and persuasive, among others.

In the first version of ODBİMDEPA in 2018, the dimension with the highest score was "universality" with a mean score of 2.9, closely followed by the "text organization" dimension with a mean score of 2.7. This can be attributed to the way the selected texts accurately represented the sub-elements of these dimensions and that the content pertaining to the dimensions was valid. On the other hand, the "argumentation" dimension recorded the lowest mean score of 1.3, which can be explained in two ways. Firstly, students lacked exposure to argumentative texts in educational settings, hindering their ability to connect the text world with the real world. Secondly, the components of the "argumentation" dimension in the textbooks were inadequately designed regarding content. Consequently, the sub-elements of the "argumentation" dimension were redefined and updated by researchers, leading to a valid score for the "argumentation" dimension in the second study. The second least effective dimension received a score of 1.9, labelled as "Style." The identical approach was used for the "Style" dimension too.

Conclusion and Discussion on the Second Set of Documents

In the second phase of the study, a total of 37 informative texts randomly selected from the tests of Turkish language and literature, philosophy, religious culture and moral knowledge, sociology and history courses in the YKS tests conducted by ÖSYM in 2023 and 2024 were scored using ODBIMDEPA by 11 raters for the texts of philosophy, religious studies, sociology, and history exams and with the addition of 2 more raters for the Turkish language and literature texts, amounting to a total of 13 raters.

In order to test the validity and reliability among raters, Fleiss' kappa and MFRM values were examined. In the analyses, Fliess' kappa values were found to be low. Although this finding can be interpreted as indicating a need for rater training or improvement of the rubric due to the low agreement between raters, the following considerations should also be taken into account: Kappa statistics measure the non-chance agreement among raters, and for the agreement coefficient to be high, raters must give the same score in each dimension. However, it should be noted that the difference between the scores given in ODBIMDEPA, where the raters are expected to make decisions on abstract qualities related to a text, is also important. It was observed that there was mostly a 1-point difference among the scores given by the raters in both document sets, and the difference in scores was not high in most dimensions as indicated in Table 2. The table shows that raters mostly assigned a score of 3 and 4, especially in the 1st dimension. An examination of the descriptors for scores 3 and 4 in the first dimension of the rubric in Appendix 1 reveals raters agree that the "**text fulfils the criteria by incorporating at least one main idea along with at least one supporting subordinate idea**". However, they do not agree on the statement that "*all parts of the text to reinforce both the primary and supporting ideas*." It is not an easy task to find numerical equivalents of qualitative descriptions within the rubric, especially for abstract qualities such as "supporting ideas" in a text. For this reason, practices such as rater training and standardization studies are implemented before using any rubric. Since the aim of the study was to follow the initial procedures of rubric development, the content of the training provided to the raters was limited. Only the content that included descriptors about the dimensions included in the rubric was shared with the raters before scoring. Therefore, inconsistency between the raters is an expected result. Also, a low Fleiss' kappa value may indicate that the raters interpret the rubric criteria differently. However, this does not mean that there is a problem with the rubric itself; on the contrary, as mentioned above, it may indicate that more training or clarification should be provided on how the raters interpret the rubric. On the other hands, the low agreement among raters as indicated by Fleiss' kappa values can be explained by the fact that linguistic expressions vary according to the background knowledge

and mental schemas of individuals, and that readers can evaluate the text from different perspectives, and that a reader can evaluate the (same) text differently even in different reading occasions. In this regard, it would be appropriate to focus on whether the rubric works on the text in general, going beyond the Kappa agreement among the raters. To this end, following the analysis of the Fleiss' kappa values, MFRM values were also analyzed.

Unlike the Kappa value, the Rasch model shows how well the scale measures the structure it is intended to measure. MFRM results show that the rubric is well-structured and that the scores reflect a certain difficulty or quality progression. This model confirms the one-dimensionality of the measurement tool and that the items perform as intended within this structure. MFRM results show that the rubric generally works according to its purpose and that the measurement structure is solid as evidenced by the EAP reliability, which was found to be 0.951. In other words, high EAP reliability shows that the Rasch model can make accurate measurements through the use of the rubric, even if there is rater variability. Consequently, since the raters—who are educators with different subject expertise and training—might employ diverse methods when reading, interpreting, and rating informative or any kind of text, the Rasch model has been demonstrated as a suitable instrument to check the reliability of this scoring rubric.

To examine the general relationship between raters, texts, and dimensions, MFRM with 3-facets analyses conducted through the FACET program revealed that the inter-rater logit values ranged from 0 to 2, with raters tending to use scores of 2 and 3 more frequently. The tendency of raters to avoid the lowest and highest scores was interpreted as a desire to remain within a "safe range." However, the limited word count of the texts in the YKS, as well as the use of only partial text excerpts (one or two paragraphs), may also contribute to this tendency. In fact, the descriptions indicating 4 points—the highest score for each dimension in the rubric—are typically very broad and inclusive. ODBIMDEPA was developed to evaluate entire texts, and it may not be feasible for the text fragments used in the YKS to meet the criteria, such as " **All parts of the text support the main and supporting ideas**," when assessing a section extracted from the full text.

To examine the individual tendencies of each rater, MFRM with 4-facets analyses were performed using the TAM program in R. The analysis results have allowed for multiple interpretations across various contexts. Firstly, the analysis of the YKS texts across various subject areas was performed, showing high EAP reliability in all instances. Nonetheless, the correlation values reflecting inter-rater reliability were typically moderate and especially low for History texts. Upon examining the evaluators' tendencies towards leniency or severity, it was observed that their tendencies were largely consistent. Two evaluators were found as exhibiting random tendencies towards leniency or severity and were consequently classified as outliers.

In the second set of documents, the total scores assigned to the texts and their classification according to the ODBİMPEYA detailed in the methodology section revealed that the majority of the texts in this set are "reader-friendly," while a few are categorized as "moderately reader-friendly." Developing reader-friendly texts aids in minimizing adverse effects on the reading comprehension process. In other words, reader-friendly texts may ensure that the items developed from these texts have lower risks of carrying construct-irrelevant characteristics which otherwise may arise from low text comprehensibility. This approach may ensure that that no other characteristics interfere with the evaluation of the intended competencies. Upon reviewing the tables demonstrating the mean and total scores of informative texts across various disciplines in the YKS exams, it was found that the exams predominantly feature texts that are reader-friendly. In this context, employing these texts for assessment purposes and developing items derived from them can improve validity by reducing the impact of low comprehensibility on the assessment process. To confirm this hypothesis, the analyses of the items developed from these texts must be examined first. Furthermore, it is important to evaluate if there are variations in parameters like difficulty and discrimination among items created from texts

classified as reader-friendly compared to those viewed as moderately reader-friendly or not reader-friendly.

In conclusion, the analyses have demonstrated the validity of ODBİMDEPA as a scoring rubric for the investigation of how reader-friendly informative texts are. With the completion of the rubric and finalization of the initial reliability analyses, the rubric can be argued to be used reliably by raters. Employing the rubric upon rater-training could enhance inter-rater agreement. ODBİMDEPA serves as a scoring rubric that facilitates a broad classification of the reader-friendliness of texts. To use it as an effective scoring rubric in the selection and development of texts for educational instruction and assessment materials and instruments, the researchers plan to apply it to different sets of documents with enhanced rater training, thereby continuing reliability analyses in the future.

Suggestions

It has been found that the distribution of informative texts in the database of this study, which consists of 5th, 6th, 7th and 8th grade textbooks in courses used by MoNE in 2018, is not balanced across the grades. It would be useful to review the distribution of informative texts in the textbooks in the curriculum updated with the Century of Türkiye Maarif Model, which started to be implemented in 2024, across the grades and courses with further studies; and to evaluate the texts in these books with ODBIMDEPA and to investigate how reader-friendly the selected texts are.

In addition, in the new Model, which is implemented in 1st, 5th and 9th grades as of 2024, further studies should explopre adequacy of the classification of informative texts in textbooks and presence of a distinction in the Model between the sub-types texts (explanatory, argumentative, assertive, etc.) in the literature. This will guide both textbook authors and educators.

Since the informative texts in the textbooks examined in the first version of ODBIMDEPA had the lowest means, especially in the "argumentation" and "style" dimensions, there is a need for further studies to investigate how these dimensions function in the textbooks in the new curriculum implemented with the Century of Türkiye Maarif Model.

Following the analysis of the informative texts in the second document set with the revised ODBİMDEPA, the results indicated that although the MFRM suggested high internal consistency for the rubric, low Fleiss' kappa values underscore the necessity for enhanced training for raters or refinements to the rubric. Thus, application of ODBİMDEPA to broader contexts (in YKS items, textbook texts, etc.), should require enhanced rater training. Monitoring and incorporating these trainings and content revisions into the literature will offer valuable insights for researchers.

There is a need for the review of whether the texts used for assessment purposes are reader-friendly with a larger data set and designing comparative studies incorporating the review of item parameters such as difficulty and discrimination. Therefore, conclusions can be drawn about how text comprehensibility impacts the assessment of reading comprehension, and the types of standards required for selecting texts can be identified.

Acknowledgements

We would like to thank our raters who score the texts day and night during the development of ODBIMDEPA. We would also like to thank Fahri Yılmaz and Prof. Dr. Nuri Doğan for their contributions in the study with their expertise.

References

Aidinlou, N., Khodamard, N., & Azami, J. (2012). The effect of textual cohesive reference instruction on the reading comprehension of Iranian EFL students. *International Journal English Linguistics*, *2*(5), 18-24. doi:10.5539/ijel.v2n5p18

Aktaş, M., & Alıcı, D. (2018). Yazılan hikâyeyi değerlendirmeye yönelik analitik rubrik geliştirme: Geçerlik ve güvenirlik çalışması. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *14*(2), 597-610. doi:10.17860/mersinefd.424198

Beck, I. L., McKeown, M. G., Sinatra, G. M., & Loxterman, J. A. (1991). Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. *Reading Research Quarterly*, 26, 251-276.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3th ed.). London: Routledge/Taylor & Francis Group. doi:10.4324/9781315814698

Brewer, W. F. (1980). Literary theory, rhetoric, and stylistics: Implications for psychology. In R. J. Shapiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 221-239). Hillsdale, NJ: Erlbaum.

Britton, B. K., & Gulgoz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, 83, 329-404.

Çelik, T., Demirgüneş, S., & Baştuğ, M. (2014). Metin oluşturma ve değerlendirme çalışmalarına yönelik "okur dostu" metin değerlendirme rubriği. *International Journal of Eurasia Social Sciences*, *5*(14), 65-83.

Çelik, T., Demirgüneş, S., & Fidan, D. (2015). Okur dostu metin olma özelliği ile okuduğunu anlama başarısı arasındaki ilişkinin incelenmesi. *Başkent Üniversitesi Eğitim Fakültesi Dergisi*, *2*(1), 119-127.

Çıralı Sarıca, H., & Koçak Usluel, Y. (2016). Eğitsel bağlamda dijital hikâye anlatımı: Bir rubrik geliştirme çalışması. *Eğitim Teknolojisi: Kuram ve Uygulama*, *6*(2), 65-84.

Demirgüneş, F. (2015). *Türkçe ders kitaplarındaki metinlerin okur dostu metin bağlamında değerlendirilmesi* (Unpublished master's thesis). Niğde University, Niğde.

Dreher, M. J., & Singer, H. (1989). Friendly texts and friendly teachers. Theory into practice. *Learning Through Text*, *28*(2), 98-104.

Duke, N., & Bennett Armistead, V. (2003). *Reading and writing informational text in the primary grades: Research-based practices*. New York: Scholastic.

Eser, M. T., & Aksu, G. (2022). *R programlama dili ile puanlayıcılar arası uyum ve güvenirlik* (N. Doğan, Ed.). Ankara: Pegem.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Thousand Oaks, CA: SAGE Publications.

Fisher, D., Frey, N., & Lapp, D. (2008). Shared reading: Modeling comprehension, vocabulary, text structures, and text features for older students. *The Reading Teacher*, *6*(*7*), 584-556. doi:10.1598/RT.61.7.4

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382.

Gray, W. S., & Leary, B. E. (1935). *What makes a book readable*. Chicago: University of Chicago Press.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.

İşeri, K., Çelik, T., & Demirgüneş, S. (2015). Türkçe ders kitaplarında okur dostu metinlerin incelenmesi. *E-Dil Dergisi*, *1*(5), 227-242.

Kintsch, W. (1982). Aspects of text comprehension. *Advances in Psychology*, 9, 301-312.

Kintsch, W. (1988). The role of knowledge in discourse processing: A construction-integration model. *Psychological Review*, 95, 163-182.

Klink, S., Dengel, A., & Kieninger, T. (2000). *Document structure analysis based on layout and textual features*. Proceedings of International Workshop on Document Analysis Systems, DAS2000.

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

Lorch, R., Lemarié, J., & Grant, R. (2011). Signaling hierarchical and sequential organization in expository text. *Scientific Studies of Reading, 15*(3), 267-284. doi:10.1080/10888431003747535

Mann, W., & Thompson, S. (1988). Rethorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse, 8*(3), 243-281. doi:10.1515/text.1.1988.8.3.243.

McNamara, T. (2001). Language assessment as social practice: challenges for research. *Language Testing*, 18, 333-349. doi:10.1177/026553220101800402.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1-43. doi:10.1207/s1532690xci1401_1.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing Linguistic Features of Cohesion. *Discourse Processes*, *47*(4), 292–330. doi:10.1080/01638530902959943

Ministry of National Education. (2024). Ortaokul Türkçe dersi öğretim programı (5, 6, 7 ve 8. sınıflar). Retrieved from https://mufredat.meb.gov.tr

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10). Retrieved from https://www.researchgate.net/publication/242637290_Scoring_Rubric_Development_Validity_and_Reliability

Müldür, M., & Şimşek, N. D. (2020). Türkçe ders kitaplarındaki bilgilendirici metinlerin ve yapılarının incelenmesi üzerine bir araştırma. *Ana Dili Eğitimi Dergisi, 8*(3), 843-867.

Nagabhand, S., Nation, I., & Franken, M. (1993). Can text be too friendly?. *Reading in a Foreign Language, 9*(2), 895-907.

Rasch, G. (1960). *Probabilistic model for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research.

Smith, F. (1982). *Writing and the writer.* London: Heinemann.

Soyuçok, M., & Kartal, H. (2018). İlkokuma-yazma ders kitaplarının metin oluşturma basamağında oluşturulan metinlerin değerlendirilmesi. *Mehmet Akif Ersoy Üniversite Eğitim Fakültesi Dergisi,* 47, 435-460.

Şencan, H. (2005). *Sosyal ve davranışsal ölçmelerde güvenirlik ve geçerlilik*. Ankara: Seçkin Yayınları.

Tuncel, G. (2011). Sosyal bilgiler dersinde rubriklerin etkili kullanımı. *Marmara Coğrafya Dergisi*, 23, 213-233.

Yıldırım, H., & Ocak, G. (2016). İlkokul sosyal bilgiler dersi "okur dostu" metinlere yönelik rubrik hazırlama çalışması. In *XV. Uluslararası Sınıf Öğretmenliği Sempozyumu bildiri özet kitabı* (pp. 422-423). Ankara: Pegem Akademi.

Yeung, L. (2007). In search of commonalities: Some linguistic and rhetorical features of business reports as a genre. *English for Specific Purposes*, *26*(2), 156-179.

Yıldırım, A., & Şimşek, H. (2008). *Sosyal bilimlerde nitel araştırma yöntemleri* (6. ed.). Ankara: Seçkin Yayıncılık.

Appendix 1

READER-FRIENDLY INFORMATIVE TEXT EVALUATION RUBRIC (ODIMDEPA)

Dimensions	1	2	3	4
Text Organization	The text clearly presents **only one piece of information or idea** to the reader, while also **containing multiple elements that contradict that information or idea.**	The text clearly presents **multiple pieces of information or ideas** to the reader. There are multiple elements in the text that contradict these pieces of information or ideas.	The text includes **at least one main idea and at least one supporting idea**. There are elements in the text that contradict the main information/idea and supporting information/ideas.	The text includes at least one main idea and at least one supporting idea. **All parts of the text support the main and supporting ideas.**
Presentation of Information	**The information in the text is listed randomly, with no attention paid to the relationship between priority and posteriority**. More than one information/idea is **not fully explained, exemplified, or made concrete.**	**Some** information/ideas in the text can be classified. More than one information/idea is **not fully explained, exemplified, or made concrete.**	The information in the text is presented in a gradual order (e.g., general to specific, simple to complex, old to new). At most, one information is **not fully explained or made concrete**.	Supporting information/ideas for the main idea to be conveyed in the text are presented in gradual stages and **there is no abstract or unexemplified information**.
Formal Patterns/Cohesion	The text includes **only** the main stylistic arrangements (e.g., font size, typeface, line spacing). Basic formal elements (e.g., font size, typeface, punctuation, spelling) **contain errors or omissions.**	The text includes the main stylistic arrangements and basic markers (e.g., bold, italic) for highlighting the main information/idea to be conveyed. In the text, the fundamental formal arrangements (font size, typeface, line spacing, spelling, punctuation, etc.) have been applied correctly. However, detailed and content-specific customized subheadings, markers, and other similar elements have not been applied.	The text incorporates not only the main formal arrangements but also text-specific subheadings and sections (such as theoretical background, problem statement, study area, problem resolution, examples, recommendations, etc.). The main idea intended to be conveyed, along with the supporting information and thoughts, is structured using **various markers (e.g., bold, italicized words, footnotes, etc.)**.	The text effectively utilizes the main formal arrangements while incorporating text-specific headings and sectioning. The typological cohesion mechanisms of the language (such as pronoun omission, anaphoric and cataphoric references, etc.) have been used flawlessly.

Semantic Patterns/Coherence	**There is no unity between the sentences of the text.** Different pieces of information are not presented in a complementary way.	**There is partial unity** between the sentences of the text, with **some use of connectors (e.g., prepositions, conjunctions)**. Information bundles show partial consistency.	**There is unity** between the sentences of the text. Connectors are used effectively, and **semantic integrity is maintained** even without connectors. **The text partially allows the inference of the main information as a whole.**	**There is semantic integrity between the sentences** of the text. The supporting information reinforcing the main idea **contributes to the coherence of the tex**t. As a whole, the text allows for **the inference of the main idea** in an integrated manner.
Explanation and Reality	The main information/idea and other relevant details (e.g., people, places, time, terms, general information bundle) **are not explained** in the text.	Explanations of additional details (e.g., people, places, time, terms, general information bundle) are **partially** included. **No** examples, comparisons, or concretizations are provided to enhance understanding.	In the text, the explanations of the information provided in support of the intended message (such as people, place, time, terms, general knowledge clusters, etc.) are largely reinforced through examples, comparisons, and concretization. However, the core information or ideas presented **do not have direct correspondences** that align with the reader's real-life experiences.	In the text, the main idea and supporting details are presented in a way that resonates with the reader's real-life experiences. This is achieved through examples, comparisons, concretization, highlighting differences and similarities, problem-solving approaches, and explanations of specific terms.
Argumentation	The information/ideas are not presented progressively, and the text **does not create** a discussion environment that arouses curiosity.	The information/ideas arouse some curiosity and **partially** lead the reader to engage in a discussion. No suggestions or clues are provided to resolve the discussion. **There are no** suggestions/tips for resolving the discussion.	The main information and supporting ideas arouse curiosity and lead the reader to engage in a mental discussion. **No** suggestions or clues are provided to resolve the discussion.	The main information initiates a mental discussion in the reader, supported by clear suggestions and clues to help the reader reach a resolution.

Style	The author has **failed to** communicate with the reader. Information or ideas are conveyed **without** presenting any unique opinions or perspectives.	The author has established **partial** communication with the reader. Frequent sharing of personal experiences and the use of multiple subjective expressions such as "in my opinion," "actually," and "as I see it" have led to a **departure from** objectivity. While the author has presented their knowledge or ideas, they have done so in a way that is closed to discussion and alternative perspectives.	The author has engaged in communication with the reader, making their stance on the presented information and ideas explicit through expressions such as "as can be seen" and "on the other hand." Rather than sharing personal experiences, the author has conveyed factual, proven information in a clear manner. Subjective expressions such as "in my opinion," "actually," "it seems," and "as I see it" are present **to some extent** but do not dominate the text.	The author has established strong communication with the reader. This communication is achieved through clear, evidence-based examples and comparisons, without relying on subjective expressions such as "in my opinion," "actually," "it seems," or "as I see it." Instead, the author conveys definitive knowledge, ideas, and judgments using objective indicators.
Universality	The text contains multiple instances of situations, events, and language usage that contradict universal sensitivities such as human rights, animal rights, sustainability, and environmentalism. These include elements such as discrimination based on language, religion, or gender, as well as the glorification of war and all forms of violence.	The text **does not employ language** that conveys sensitivities related to universal values such as human rights, animal rights, sustainability, or environmentalism. However, there is at least one instance of a statement or idea that includes discrimination based on language, religion, or gender, or that expresses praise for war or any form of violence.	The text does not contain any elements that violate human, animal, or environmental rights. However, **it does not evoke** interest, curiosity, or appreciation for different cultures or living beings. **No content expansions have been made** to engage the reader's interest in different cultures or relationships.	The text includes elements that promote interest, curiosity, and appreciation for diverse cultures and species, with universally acceptable content. In terms of content, the text holds universal value with subject matter that is acceptable across all cultures.

ODBIMDER's Score Ranges

25 points and above **:** Reader-friendly text

16-24 points **:** Moderately reader-friendly text

15 points or less **:** Text that is not reader-friendly