ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(3); 222-234

Rubrics in Terms of Development Processes and Misconceptions*

Fuat ELKONCA** Görkem CEYHAN*** Mehmet ŞATA****

Abstract

The present study aimed to examine the development process of rubrics in theses indexed in the national thesis database and to identify any misconceptions presented in these rubrics. A qualitative research approach utilizing document analysis was employed. The sample of theses was selected based on a literature review and criteria established by expert opinions, resulting in a total of 395 theses being included in the study using criterion sampling. Data were collected through a "thesis review form" developed by the researchers. Descriptive analysis was employed for data analysis. Findings indicated that approximately 27% of the 395 theses contained misconceptions, with a disproportionate percentage of these misconceptions (The rating scale was called rubric and the checklist was called rubric) being found in master's theses. Regarding the field of the thesis, the highest rate of misconceptions was observed in health, social sciences, special education, and fine arts, while the lowest rate was found in education and linguistics. Additionally, theses with misconceptions tended to possess a lower degree of validity and reliability evidence compared to those without misconceptions. This difference was found to be statistically significant for both validity evidence and reliability evidence. In theses without misconceptions, the most frequently presented validity evidence was expert opinion, while the reliability evidence was found to be the percentage of agreement. The findings were discussed in relation to the existing literature, and recommendations were proposed.

Keywords: rubric, document analysis, misconception, reliability, validity.

Introduction

In the field of social and educational sciences, the use of appropriate measurement tools and methods is crucial to ensure the consistency and accuracy of decisions made about test takers. These characteristics are often intangible and exist only through indirect measurement. Therefore, it is important to provide evidence of the reliability and validity of the measurements obtained from these tools. There are various classifications for measurement tools, but they can generally be divided into traditional and complementary/versatile categories. The shift towards a constructivist approach in education since 2005-2006 has led to increased use of complementary measurement tools.

Rubrics, a type of complementary measurement tool, have gained widespread use in education and training activities (Brookhart, 2018). This trend is largely attributed to the flexibility and appropriateness of rubrics in assessing 21st-century skills, which are higher-order cognitive abilities (Dochy et al., 2006). Rubrics must be designed with clear and well-defined criteria and performance level definitions to measure these skills effectively (Brookhart & Chen, 2015; Lane & Tierney, 2008). One of the main reasons for the popularity of rubrics in education and training is their high level of reliability and validity in measurement (Jonsson & Svingby, 2007). Several studies have explored the use of rubrics in education and have discussed the reliability and validity issues surrounding their use (Brookhart, 2018; Brookhart & Chen, 2015; Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Reddy & Andrade, 2010). These studies suggest that the development of rubrics should be approached in a systematic manner, with a focus on collecting evidence for their reliability and validity (Moskal, 2000; Moskal & Leydens, 2000).

* A part of this study was presented at 8th International Congress on Measurement and Evaluation in Education and Psychology. Ege University, İzmir, Turkey.

** Asst. Prof. Dr., Muş Alparslan University, Muş- Türkiye, f.elkonca@alparslan.edu.tr, ORCID ID: 0000-0002-2733-8891
*** Asst. Prof. Dr., Muş Alparslan University, Muş- Türkiye, g.ceyhan@alparslan.edu.tr, ORCID ID: 0000-0001-9342-6876
**** Assoc. Prof.Dr., Van Yüzüncü Yıl University, Faculty of Education, Van-Türkiye, mehmetwsata@gmail.com, ORCID ID: 0000-0003-2683-4997

To cite this article:
Elkonca, F., Ceyhan, G. & Şata, M. (2023). Rubrics in terms of development processes and misconceptions. *Journal of Measurement and Evaluation in Education and Psychology, 14*(3), 222-234. https://doi.org/10.21031/epod.1251470
Received: 15.02.2023
Accepted: 13.09.2023

Unlike checklists and rating scales, rubrics provide a clear definition for each performance level, which is essential for ensuring the validity of measurements. In the process of developing rubrics, it is crucial to seek input from experts in the field to ensure that the definitions accurately represent the relevant features being measured (Moskal & Leydens, 2000). Rubrics are widely used in both educational research and classroom evaluation practices, as they also measure psychological constructs. Therefore, evidence of construct validity is crucial for making accurate inferences. According to the literature, rubrics have several benefits, including higher rater reliability, improved measurement of complex performance tasks, and increased individual reasoning skills (Jonsson & Svingby, 2007; Morrison & Ross, 1998; Wiggins, 1998). These benefits can be realized by ensuring reliability and validity in the development process of rubrics.

It is evident in the national literature that the concept of rubrics is utilized in a variety of different concepts and meanings, indicating the presence of misconceptions. Misconceptions, defined as perceptions or understandings that deviate from the expert consensus (Zembat, 2010), are not solely indicative of errors or lack of knowledge, but rather emerge as a result of faulty cognitive structures. As misconceptions correspond to situations in which cognitive perception leads to systematic errors, individuals who hold misconceptions often exhibit resistance and are unwilling to accept their existence (Yenilmez & Yaşa, 2008). The literature is limited in terms of studies that specifically investigate misconceptions related to rubrics in detail (Brookhart, 2013; Brookhart, 2018; Reynolds-Keefer, 2010). Brookhart (2013) has highlighted that the most prevalent misconceptions include the belief that rubrics are solely used as assessment tools for products and that they serve to quantitatively measure student learning, as well as the conflation of rubrics with rating scale tools. These misconceptions limit the purpose of using rubrics and hinder the full realization of student learning. Therefore, it is crucial to identify and address misconceptions surrounding rubrics. In the present study, the prevalence of misconceptions surrounding rubrics is considered to be of equal importance to the development processes of rubrics.

Numerous studies in the academic literature have examined the use and analysis of rubrics. A commonality among these studies is the emphasis on the presentation of reliability and validity evidence in the development and utilization of rubrics. The present research endeavors to not only investigate this aspect but also to determine if misconceptions exist concerning the utilization of rubrics in master’s theses and dissertations (hereafter theses). The evaluation of both the development processes and correct use of rubrics, which are frequently employed in the precise and consistent assessment of 21st-century skills, highlights the significance of this study. Furthermore, while international literature offers a plethora of studies examining rubrics across various levels of education and educational research, the dearth of such studies in the national literature underscores the importance of this research. Additionally, the study aims to examine the rubrics used in theses conducted between 2005, when the constructivist approach was incorporated into the education system, and 2022.

In this study, the primary objective was to examine the development process of rubrics utilized in theses and to investigate any misconceptions surrounding their use. To achieve this goal, the research sought to address the following questions:

1. What is the distribution of rubrics used in theses according to the type and field of theses?
2. Are there misconceptions in the process of developing and using rubrics used in theses, and if so, what types of misconceptions exist?
3. Is there a difference in theses with and without misconceptions in relation to the field and type of theses?
4. Is there a difference in terms of presenting the validity and reliability evidence of theses with and without misconceptions?
5. What is the distribution of theses without misconceptions (the rating scale was called rubric and the checklist was called rubric)?
6. Is there a difference in the validity evidence and reliability evidence of theses without misconceptions according to the field of theses?

Method

Research Design

The study employs document analysis, a qualitative research method, to examine the development processes of rubrics used in theses and associated misconceptions. Document analysis is a systematic approach to evaluate both electronic and printed sources (Corbin & Strauss, 2008; Koyuncu, et al., 2018). While various processes have been reported in the literature, this study adhered to the five stages proposed by Forster (1995), namely, (1) accessing the documents, (2) verifying their authenticity, (3) comprehending the content of the documents, (4) conducting data analysis, and (5) utilizing the obtained data (as cited in Yıldırım & Şimşek, 2011).

Population and Sample

The population of theses was all dissertations and theses scanned in the YÖK (Council of Higher Education) thesis system. However, the criterion sampling method was used, and all theses included in the YÖK thesis system between January 1, 2005 and March 1, 2022, were selected as a sample. This selection was influenced by the fact that constructivist education and complementary measurement and evaluation approaches were commonly used after 2005. The search words such as rubric, rating scale, and checklist were used in the YÖK thesis system to identify relevant theses. A total of 512 theses and dissertations were found as a result of the search with the criteria of year and searching words, but 38 theses with duplicate ID numbers were removed, resulting in 474 theses being included in the examination. Of these, 79 theses were excluded from the study because they only mentioned the name of the rubric and did not use it, leaving a total of 395 theses examined.

Data Collection Tool

The thesis review form developed by the researchers was used as a data collection tool. This tool was created through an analysis of relevant literature and the development of a list of criteria that align with the characteristics and processes that rubrics should possess. Initially, a total of 15 criteria were established.

Validity and Reliability Evidence for the Data Collection Tool

The researchers collected evidence to establish the reliability and validity of measurements obtained based on the checklist developed in their study. To assess content validity, the researchers employed Lawshe's (1975) approach and solicited the opinions of eight experts in the field of Measurement and Evaluation in Education to determine the appropriateness and content validity of the criteria. The content validity ratio limit value for eight experts was set at .69, and one criteria that fell below this threshold were removed (Wilson et al., 2012). One criterion was also revised, resulting in a final data collection tool comprising 14 criteria. This criteria; type of thesis and dissertations, sample group, field of the thesis and dissertations, status of having misconceptions, type of misconception, validity evidence, reliability evidence, rubric type, originality, sample size, guided theory, number of rating scale levels, weighting and scoring.

Considering that the checklists and rating scales mentioned by Brookhart (2018) as misconception types are often referred to as rubrics, these misconceptions were expected to emerge.

To establish the reliability of the measurements obtained from the measurement tool, three experts independently coded 10 randomly selected theses and evaluated each one according to the 13 different criteria. Krippendorff's Alpha reliability coefficient was calculated to determine inter-coder agreement, yielding a coefficient of .93.

Data analysis

In studies conducted based on a qualitative research approach, there are two basic analysis processes: content and descriptive analysis. In this study, a qualitative research approach was adopted, and the method of descriptive analysis was selected as the primary technique for data analysis. The choice of this method was based on the pre-determined features of the rubric, which were established through a thorough examination of existing literature. Furthermore, the chi-square analysis was applied to investigate the incidence of misconceptions, while the z ratio test was utilized to ascertain the presence of significant differences between the categories of the criteria. All data analysis procedures were conducted with a significance level of 0.05.

Findings

The findings were presented according to the order of the research questions. Thus, Table 1 presented information about the thesis type, the field of the thesis, and the sample group of the documents analyzed.

Table 1

Distribution of theses according to their type and field, and sample group

Criterion	Category	f	%
Type of thesis	Master's thesis	241	61.0
	Dissertation	154	39.0
Sample group	Primary school	55	13.9
	Middle school	124	31.4
	High school	35	8.9
	Associate degree	2	0.5
	Undergraduate	120	30.4
	Teacher	29	7.3
	Other	30	7.6
Field of the thesis	Educational sciences	98	24.8
	Basic education	45	11.4
	Special education	3	0.8
	Science and math education	99	25.1
	Turkish and social education	67	17.0
	Science	4	1.0
	Health sciences	4	1.0
	Social sciences	5	1.3
	Fine arts	41	10.4
	Linguistics	29	7.3
Total		395	100

Regarding Table 1, most of the theses utilizing rubrics were master's theses. Furthermore, the primary sample population for these theses was composed of individuals at the secondary school and undergraduate levels. Upon examination of the distribution of theses by field, the majority were in the fields of science and mathematics education and educational sciences. Following the analysis of the distribution of rubrics according to the type and field of the thesis, Table 2 presented an examination of the prevalence of misconceptions and, if present, identified the specific misconceptions.

Table 2

Distribution of the presence of misconceptions in rubrics and identification of specific misconceptions

Misconception		**f**	**%**
Status of having misconceptions	Yes	104	26.3
	None	291	73.7
	Total	395	100.0
Type of misconception	The rating scale was called rubric	88	85.0
	The checklist was called rubric	16	15.0
	Total	104	100.0

Based on the distributions in Table 2, 104 rubrics had misconceptions while 291 (73.7%) did not. It is seen that in 88 (85.0%) of the theses with misconceptions, the rating scale was called as rubric, and the checklist was called as rubric in 16 (15.0%) of the theses with misconceptions. The comparison of field of the thesis in terms of having misconceptions was presented in Table 3.

Table 3

Comparison of the theses with and without misconceptions according to the field of the thesis

Category	**Misconception**				**Chi-square**			
	No		**Yes**		χ^2	**p**	**Compare Column Proportions**	
	f	**%**	**f**	**%**			**Misconception (No)**	**Misconception (Yes)**
Educational sciences (A)	79	80.6	19	19.4			A-F (p = .047) A-G (p = .031)	
Basic education (B)	31	68.9	14	31.1				
Science and math education (C)	78	78.8	21	21.2			C-G (p = .032)	
Turkish and social education (D)	49	73.1	18	26.9	17.01	.009*		
Linguistics (E)	23	79.3	6	20.7				
Fine arts (F)	24	58.5	17	41.5				F-A (p = .047)
Other (G)	7	43.8	9	56.3				G-A (p = .031) G-C (p = .032)
Total	291	73.7	104	26.3				

*p< .05

As demonstrated in Table 3, a significant difference (χ^2= 17.01; p < .05) was observed in the prevalence of misconceptions in the rubrics of the theses analyzed within the scope of the study, based on the fields of the theses. The lowest prevalence of misconceptions was found in the fields of Educational Sciences (80.6%), Linguistics (79.3%), and Science and Math Education (78.8%), while the highest prevalence of misconceptions was found in the fields of Fine Arts (41.5%) and other fields (Health, Social Sciences, Special Education, Science) (56.3%). In order to determine the source of the difference, column ratios were compared (z-test) and it was concluded that theses written in the fields of Educational Sciences and Science and Math Education contained fewer misconceptions than theses written in Fine Arts and other fields (health, social sciences, special education, science). The findings related to the comparison of the rubrics with and without misconceptions according to thesis type were presented in Table 4.

Table 4

Comparison of the theses with and without misconceptions according to the thesis type

Category	Misconception				Chi-square			
	No		Yes		χ^2	p	Compare Column Proportions	
	f	%	f	%			Misconception (No)	Misconception (Yes)
Master's thesis	174	72.2	67	27.8	0.69	.406	---	---
Dissertation	117	76.0	37	24.0			---	---
Total	291	73.7	104	26.3				

As exhibited in Table 4, an analysis was conducted to investigate the prevalence of misconceptions in the rubrics, based on the level of degree (master's thesis or dissertation). Results revealed that there was no statistically significant difference in the prevalence of misconceptions between the two groups (χ^2= 0.69; p >. 05). Specifically, it was found that 27.8% of the master's theses and 24% of the dissertations contained misconceptions, with similar ratios observed in both groups.

Table 5

Comparison of theses with and without misconceptions regarding validity and reliability evidence

Variable	Category	Misconception				Chi-square			
		No		Yes		χ^2	p	Compare Column Proportions	
		f	%	f	%			Misconception (No)	Misconception (Yes)
Validity evidence	No (A)	102	65.4	54	34.6	9.13	.003*	---	A-B (p = .003)
	Yes (B)	189	79.1	50	20.9			B-A (p = .003)	---
	Total	291	73.7	104	26.3				
Reliability evidence	No (A)	141	64.7	77	35.3	20.28	.000*	---	A-B (p = .000)
	Yes (B)	150	84.7	27	15.3			B-A (p = .000)	---
	Total	291	73.7	104	26.3				

*p < .05

Table 5 presented the results of a chi-square analysis comparing the presence of misconceptions in the rubrics of the theses within the scope of the research, in terms of the inclusion of validity and reliability evidence. The results indicated a statistically significant difference between the two groups (χ^2= 9.13; p < .05). The findings revealed that the proportion of theses containing misconceptions was higher among the group without validity evidence (34.6%) compared to the group with validity evidence (20.9%). The same pattern was observed when examining the presence of misconceptions in relation to reliability evidence, with 35% of theses without reliability evidence containing misconceptions, compared to 15% of theses with reliability evidence (χ^2= 20.28; p < .05).

Table 6

Distribution of the rubrics used in theses without misconceptions according to various characteristics

Criterion	Category	f	%
Rubric type	Analytic	248	85.2
	Holistic	43	14.8
Originality	Developed	227	78.0
	Adapted	13	4.5
	Original	51	17.5
Sample size	0-30 sample size	102	35.1
	31-100	115	39.5
	101-200	37	12.7
	201 and above	37	12.7
Guided Theory	No	135	46.4
	Classical test theory (CTT)	141	48.5
	Generalizability theory	6	2.1
	More than 1 theory	9	3.1
Number of rating scale levels	Three-level	74	25.4
	Four-level	121	41.6
	Five-level	59	20.3
	Six-level	11	3.8
	Seven-level and above	6	2.1
	Multiple different levels	17	5.8
	No level	3	1.0
Weighting	Criteria were weighted the same	255	87.6
	Criteria weighted differently	34	11.7
	Criteria were not scored	2	0.7
Scoring	Total score	239	82.1
	Median	1	0.3
	Mean	36	12.4
	Percentage	15	5.2

In the analysis of Table 6, 248 (85.2%) rubrics used were analytical, while 43 (14.8%) were holistic. 227 (78%) rubrics were created by the researchers themselves, 13 (4.5%) were adapted, and 51 (17.5%) were taken from another study. In terms of sample sizes, 102 (35.1%) of the rubrics used 0-30 samples, 115 (39.5%) used 31-100, 37 (12.7%) used 101-200, and 37 (12.7%) used 201 or more. 135 (46.4%) of the rubrics lacked theory-based steps, 141 (48.5%) included classical test theory, 6 (2.1%) included generalizability theory, and 9 (3.1%) included more than one theory. Considering the findings on how many levels the criteria of the DPAs were graded, 74 (25.4%) were graded in threes, 121 (41.6%) in fours, 59 (20.3%) in fives, 11 (3.8%) in sixes and 6 (2.1%) in sevens and above. In addition, the criteria were scored differently in 17 rubrics (5.8%), and 3 rubrics were not scored. Considering the different weighting of the criteria, equal weighting was used in the majority of the rubrics (f = 255; 87.6%) while 34 (11.7%) criteria were weighted differently, and 2 (0.7%) rubrics were not rated. Considering the methods used in the interpretation of the scores obtained from rubrics, 239 rubrics (82.0%) were interpreted by taking the total score, 36 (12.4%) by taking the mean score, 15 (5.2%) by taking the percentage, and 1 by taking the median score. Whether the rubrics used in theses without misconceptions contain validity evidence was compared according to their fields, and the findings were presented in Table 7.

Table 7

Comparison of the validity evidence of the rubrics without misconceptions according to the thesis fields

Category	Validity Evidence				Chi-square			
	No		Yes				Compare Column Proportions	
	f	%	f	%	χ^2	p	Validity Evidence (No)	Validity Evidence (Yes)
Educational sciences (A)	17	21.5	62	78.5	14.66	.023*		A-B (p = .031) A-C (p = .000)
Basic education (B)	13	41.9	18	58.1			B-A (p = .031)	
Science and math education (C)	38	48.7	40	51.3			C-A (p = .000) C-F (p = .040)	
Turkish and social education (D)	17	34.7	32	65.3				
Linguistics (E)	8	34.8	15	65.2				
Fine arts (F)	6	25.0	18	75.0				F-C (p = .040)
Other (G)	3	42.9	4	57.1				
Total	102	35.1	189	64.9				

*p < .05

As can be seen in Table 7, the presence or absence of validity evidence in the rubrics without misconceptions in the theses analyzed within the scope of the research was compared according to the thesis fields and a statistically significant difference was obtained (χ^2= 14.66; p < .05). Based on the findings, in the process of developing or using rubrics, the most validity evidence was presented in the fields of Educational Sciences (78.5%) and Fine Arts (75%), respectively. In addition, the least validity evidence was in the fields of Science and Mathematics education (51.3%), Other fields (57.1%) and Basic Education (58.1%). In order to determine the source of the difference, column ratios were compared (z-test). The rate of having validity evidence of rubrics in theses written in the fields of Educational Sciences and Fine Arts Education was significantly higher than Basic Education and Science and Math fields. The types of validity evidence presented for the rubrics used in theses without misconceptions were also analyzed and their distributions were presented in Table 8.

Table 8

Distribution of the types of validity evidence presented in the rubrics used in theses without misconceptions

Types of Validity Evidence		Yes		No	
		f	%	f	%
Validity Evidence		189	64.9	102	35.1
Factor Analysis		9	3.1	282	96.9
Content Validity		186	63.9	105	36.1
	Expert Opinion Only	178	61.2	113	38.8
	Lawshe-Davis	7	2.4	284	97.6
	Table of specification	5	1.7	286	98.3
Criterion Validity		2	0.7	289	99.3

According to the findings, validity evidence was reported in a total of 189 (64.9%) theses. The striking result of the study was that the evidence presented for content validity (f = 186; 63.9%) was quite high, but it was concluded that most of this evidence relied on expert opinion only (f = 178; 61.2%). For content validity, statistical analyses such as Lawshe-Davis (f=7; 2.4%) and table of specification (f=5; 1.7%) were involved in a minimal number of theses. Similarly, it was concluded that the evidence presented for factor analysis (f=9; 3.1%) and criterion validity (f=2; 0.7%) were very few. Within the scope of the research, whether the DPAs used in theses without misconceptions contain reliability evidence was compared according to the fields in which the theses were written and the findings obtained are given in Table 9.

Table 9

Comparison of the reliability evidence of the rubrics without misconceptions according to the thesis fields

Category	Reliability Evidence				Chi-square			
	No		Yes				Compare Column Proportions	
	f	%	f	%	χ^2	p	Reliability Evidence (No)	Reliability Evidence (Yes)
Educational sciences (A)	31	39.2	48	60.8	6.77	.343	---	---
Basic education (B)	16	51.6	15	48.4			---	---
Science and math education (C)	43	55.1	35	44.9			---	---
Turkish and social education (D)	23	46.9	26	53.1			---	---
Linguistics (E)	13	56.5	10	43.5			---	---
Fine arts (F)	10	41.7	14	58.3			---	---
Other (G)	5	71.4	2	28.6			---	---
Total	141	48.5	150	51.5				

As seen in Table 9, the presence or absence of reliability evidence in the rubrics without misconceptions in the theses was compared according to the thesis fields, and no statistically significant difference was found (χ^2= 6.77; p > .05). In general, 51.5% of the theses had reliability evidence, while 48.5% did not. Although, similar to the validity results, more reliability evidence was reported in the rubrics used in theses in the fields of educational sciences (60.8%) and fine arts (58.3%), this difference was not statistically significant. Within the scope of the research, the types of reliability evidence presented for the rubrics used in the theses without misconceptions were also analyzed and their distributions were given in Table 10.

Table 10

Reliability evidence presented in the rubrics used in theses without misconceptions

Types of Reliability Evidence		Yes		No	
		f	%	f	%
Reliability Evidence		150	51.5	141	48.5
Item Analysis (Difficulty, discrimination, t-test)		7	2.4	284	97.6
Test-retest		5	1.7	286	98.3
Cronbach Alpha		25	8.6	266	91.4
Inter-Rater Reliability		138	47.4	153	52.6
	Percentage agreement	53	18.2	238	81.8
	Intraclass correlation	44	15.1	247	84.9
	Cohen Kappa	31	10.7	260	89.3
	Kendall Tau	17	5.9	274	94.1
	Krippendorff's Alpha	7	2.4	284	97.6
	G study (generalizability)	4	1.4	287	98.6
	Rasch	3	1.1	288	98.9

According to Table 10, a total of 150 (51.5%) theses reported reliability evidence. The evidence presented for rater reliability was generally high (f = 138; 47.4%). Considering the types of rater reliability, reliability coefficient was reported using Percentage agreement in 53 (18.2%) theses, intraclass correlation coefficient in 44 (15.1%) theses, Cohen kappa in 31 (10.7%) theses, Kendall Tau in 17 (5.9%) theses, Krippendoff's Alpha in 7 (2.4%) theses, G coefficient in 4 (1.4%) theses, and Rasch method in 3 (1.1%) theses. In addition to these results, 7 (2.4%) theses reported evidence for item analysis, 5 (1.7%) theses reported test-retest and 25 (8.6%) theses reported Cronbach's Alpha reliability coefficient.

Discussion, Conclusion and Recommendations

This study aimed to examine the development process of rubrics used in theses and the misconceptions about use and consruction of rubrics in this process. Findings were discussed according to the research questions.

Most postgraduate theses that used rubrics as data collection tools were at the master's level, with the sample mostly from the secondary school and undergraduate levels. Most were used in science and math education and educational sciences. Document analysis studies showed similar results (Brookhart, 2018; Çolak-Ayyıldız, 2022; Ocak & Yeter, 2018; Reddy & Andrade, 2010). Brookhart (2018) examined the articles published between 2005-2017 and found that most rubrics were based on undergraduate students.

Regarding the findings related to the misconceptions and misconception types, it was found that one-fourth of the theses contained misconceptions. The majority of misconceptions were caused by the use of a rating scale as a rubric. Only a small number of theses used checklists as rubrics. In a similar study, Brookhart (2018) found that checklists were used as rubrics in only 7 of 51 articles. This misconception is present in both national and international literature but is more prevalent in national literature. This highlights a deficiency in the knowledge of researchers in national literature. The lack of addressing this issue in the literature presents a significant problem in practice.

The analysis of misconceptions according to discipline area revealed that the lowest number of misconceptions were in educational sciences, science and math education, and linguistics, while the highest number of misconceptions were in fine arts, which was found to be statistically significant. This may suggest lower reliability and validity of scores obtained through the use of rubrics in fine arts, compared to higher reliability evidence presented in educational sciences theses, which may be due to courses on scale development in postgraduate education. No significant difference was found in misconceptions according to thesis type (dissertation or master's). This indicates that misconceptions are similar in both levels, with 25% of theses having misconceptions, pointing to a high level of misconceptions. Despite regular monitoring of dissertations, this situation highlights a significant deficiency in practice and evaluation.

An analysis was conducted to differentiate the validity and reliability evidence of theses using rubrics as a data collection tool between those with and without misconceptions. Results showed a statistically significant difference between the two groups, with theses without misconceptions having greater validity and reliability evidence. Studies in the literature (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Reddy & Andrade, 2010; Rezaei & Lovorn, 2010) showed that reliability and validity evidence for measurements obtained from rubrics were presented. The validity evidence presented in theses without misconceptions was found to mostly be based on expert opinion (content validity), a non-statistical process. Review studies in the literature (Brookhart, 2018; Jonsson & Svingby, 2007; Panadero & Jonsson, 2013) reported similar results. (Brookhart, 2018; Jonsson & Svingby, 2007; Panadero & Jonsson, 2013). In Brookhart (2018), it was found that expert opinion (content validity) was the main form of validity evidence presented. Jonsson & Svingby (2007) found a lower frequency of content validity as validity evidence. Rater reliability was the most commonly reported form of reliability evidence when using rubrics without misconceptions (Brookhart, 2018; Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Reddy & Andrade, 2010; Rezaei & Lovorn, 2010). Jonsson and Svingby (2007) reported that over half of 76 articles used rater reliability. Rezaei and Lovorn (2010) similarly found that rater reliability was commonly reported. Brookhart (2018) argued that rater reliability was generally reported in studies using rubrics. This shows that studies using rubrics in the literature tend to present inter-rater reliability as evidence of reliability.

This study focused on theses in which misconception-free rubrics were used since it examined the properties of rubrics. It was found that analytical rubrics were the most commonly used type, primarily developed by researchers rather than adapting pre-existing rubrics (similar to findings in Brookhart, 2018). The widespread use of analytical rubrics in academic studies can be attributed to the specificity of such rubrics. In the development processes of the rubrics, generally a sample size of 100 or less was used, and CTT-based analyses were conducted. In the study conducted by Brookhart (2018), small samples were used more. The prevalence of analytical rubrics in academic studies is largely due to their specificity and the demands of the evaluation process. Analytical rubrics necessitate a more extended evaluation time and are geared towards specific goals and in-class evaluations rather than broader assessments. The use of small sample sizes, as seen in the examination of theses in relevant research, reflects these factors. The rubrics used in these studies were typically assigned levels of four, three, and five, with criteria often having equal weight and total scores being the predominant scoring method. The utilization of mean and median scores was limited.

The results of this research were summarized as follows:

- An analysis of theses utilizing rubrics as data collection tools showed that a majority of the publications were from educational sciences, science and mathematics education, and secondary and higher education. Master's theses made up the majority of the sample.
- The study found that 25% of the theses containing rubrics had misconceptions, and the rating scale was the most commonly used rubric type.
- The least number of misconceptions was found in educational sciences, science and mathematics education, and linguistics, while fine arts showed the highest number of misconceptions. Master's theses and dissertations had similar levels of misconceptions.
- The reliability and validity evidence of the theses with misconceptions were less than those without, and this difference was statistically significant.
- Validity evidence was reported more in theses without misconceptions, especially in theses in the field of Educational Sciences, compared to theses written in other fields.
- The most common validity evidence presented in theses without misconceptions is expert opinion, and the majority of these do not include statistics based on methods such as Lawshe/Davis.
- Percentage agreement was used as reliability evidence, and the use of methods such as Krippendorff's Alpha, generalizability and Rasch was very limited.
- The rubrics used in the theses mainly were equally weighted, analytical, and total score-based.

It should be noted that the results of this research are limited to theses published between 2005 and 2022 and do not encompass other forms of publication. Hence, the findings are restricted to the analysis of theses and may not be representative of the broader literature in the field.

The research highlights the need for increased training and education on rubric development, with a focus on their general features and reliability and validity evidence. It is suggested that experts with experience in scale development be included in thesis committees. It is recommended that, in order to mitigate the identified limitations and misconceptions in the use of rubrics in theses, thesis supervisors should encourage and recommend courses on scale development and adaptation for students working on projects involving measurement tools. The language barriers and resulting translation misconceptions can be addressed by establishing a common vocabulary or dictionary for concepts in the field of measurement and evaluation.

Declarations

Author Contribution: Fuat Elkonca: Methodology, analysis, discussion, writing & editing, visualization. Görkem Ceyhan: Analysis, discussion, writing & editing, visualization. Mehmet Şata: Introduction, discussion, writing & editing.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: The research data was obtained from theses scanned in the YÖK (Council of Higher Education) thesis system. Therefore, ethical approval is not required.

References

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Brookhart, S. M. (2018). Appropriate criteria: key to effective rubrics. *In Frontiers in Education, 3*(22), 1-12. https://doi.org/10.3389/feduc.2018.00022

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educ. Rev, 67*(3), 343–368. https://doi.org/10.1080/00131911.2014.929565

Corbin, J. & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed.). Thousand Oaks, CA: Sage.

Çolak-Ayyıldız, A. (2022). Alternatif eğitim konusunda yapılmış lisansüstü eğitim tezlerinin incelenmesi. *Gümüşhane Üniversitesi Sosyal Bilimler Dergisi, 13*(3), 877-886.

Dochy, F., Gijbels, D., & Segers, M. (2006). Learning and the emerging new assessment culture. In L. Verschaffel, F. Dochy, M. Boekaerts, & S. Vosniadou (Eds.), Instructional psychology: *Past, present and future trends*. Elsevier.

Forster, N. (1995). *The analysis of company documentation.* C. Cassell ve G. Symon (Eds.), Qualitative medhods in organizational research: A practical guide. Sage.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review, 2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Koyuncu, M. S., Şata, M. & Karakaya, İ. (2018). Eğitimde ölçme ve değerlendirme kongrelerinde sunulan bildirilerin doküman analizi yöntemi ile incelenmesi. *Journal of Measurement and Evaluation in Education and Psychology, 9*(2), 216-238. https://doi.org/10.21031/epod.334292

Lane, S., & Tierney S. T., (2008). Performance Assessment. Thomas L. G, (Ed), In *21st century education: A reference handbook* (Vol. 1), SAGE.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology, 28*(4), 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Morrison, G. R., & Ross, S. M. (1998). Evaluating technology-based processes and products. *New Directions for Teaching and Learning, 74*, 69-77. https://doi.org/10.1002/tl.7407

Moskal, B. M. (2000). Scoring rubrics: What, when and how*?. Practical Assessment, Research, and Evaluation, 7*(3), 1-5. https://doi.org/10.7275/a5vq-7q66

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, and Evaluation, 7*(10), 1-6. https://doi.org/10.7275/q7rm-gg74

Ocak, İ., & Yeter, F. (2018). Investigation of national theses and articles on "the nature of science" between 2006-2016 years. *Journal of Theoretical Educational Science*, *11*(3), 522-543. https://doi.org/10.30831/akukeg.344726

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational research review, 9*(1), 129-144. https://doi.org/10.1016/j.edurev.2013.01.002

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & evaluation in higher education, 35*(4), 435-448. https://doi.org/10.1080/02602930902862859

Reynolds-Keefer, L. (2010). Rubric-referenced assessment in teacher preparation: An opportunity to learn by using. *Practical Assessment, Research, and Evaluation, 15*(8), 1-9. https://doi.org/10.7275/psk5-mf68

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing writing*, *15*(1), 18-39. https://doi.org/10.1016/j.asw.2010.01.003

Wiggins, G. (1998). *Educative assessment*. Jossey-Bass.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development, 45*(3), 197-210. https://doi.org/10.1177/0748175612440286

Yenilmez, K., & Yaşa, E. (2008). İlköğretim öğrencilerinin geometrideki kavram yanılgıları. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi, 21*(2), 461-483.

Zembat, İ, Ö. (2010). Kavram yanılgısı nedir?. MF. Özmantar, E. Bingölbali & H. Akkoç (Eds.), *Matematiksel kavram yanılgıları ve çözüm önerileri içinde*. Pegem Akademi.