RedFame

Journal of Education and Training Studies
Vol. 8, No. 8; August 2020
ISSN 2324-805X E-ISSN 2324-8068
Published by Redfame Publishing
URL: http://jets.redfame.com

Investigating the Reliability and Accuracy of Teachers' Judgement in Assessing Writing

Styliani Bill Xanthi

Correspondence: Styliani Bill Xanthi, Hellenic Open University, Greece.

Received: May 28, 2020 Accepted: June 30, 2020 Online Published: July 6, 2020

doi:10.11114/jets.v8i8.4878 URL: https://doi.org/10.11114/jets.v8i8.4878

Abstract

Based on two different types of researches, this article explores how teachers' judgements for the school performance of pupils in the last three grades of the elementary school, relate to their performance in written assessment tests. The aim was to examine the teachers' judgement accuracy and reliability, and investigate the individual factors that influence them. A series of tests was assigned to the pupils and a variety of writing and word processing skills was measured. The classroom teachers assessed the pupils based on their responses to the objectives of the curriculum in all subjects. They put a numerical score from 5 to 10. Also, the teachers assessed the pupils with Learning Disabilities in the language skills, as well as for the occurrence or not of specific behavioral problems, by completing a Likert 5-point questionnaire. In the first study, the results showed significant correlations, for more tests, at a high level. In the second study, they were low and medium. It seems that teachers evaluate the school performance of their pupils significantly based on their performance in writing tests. However, their judgments about school performance are broader and assess pupils' overall ability in a variety of subjects. The correlations between the two measurements appear to be influenced by several factors related to the class level or school, the pupils' abilities, the type of test, and the way of assessment.

Keywords: teacher' judgements, school performance of pupils, writing skills, language skills, reliability, accuracy

1. Introduction

The teachers' ability to measure pupils' performance accurately is considered an important aspect of their professional competence, as they are often the source of information about pupils' school performance. On an international level, the research has focused on the reliability and accuracy of teachers' judgements, mainly by examining the relationships between teachers' judgements and pupils' performance in a range of subjects. The research findings vary, others arguing that the teachers are good judges of their pupils' academic performance relatively, while others considered them inaccurate judges of pupils' abilities (Hoge & Coladarci, 1989; Ready & Wright, 2011; Südkamp et al., 2012, 2014). One possible explanation for the conflicting findings, is the different assessment methods and data analysis procedures used in the studies.

The role of the teacher in the process of evaluation and intervention has a long history that dates back to the first days of psychometry. A significant number of weighted tests, use the existence of high statistical relevance to teachers' assessments to prove their validity both internationally and in Greece (Coladarci, 1986, Triga, 2004 as cited in Triga-Mertica, 2010). Through continuous monitoring of their pupils and daily interaction with them, teachers are aware of the performance and skills of each pupil, as well as his behavior compared to his classmates, bearing in mind a wide range of different pupils' behaviors and cases. Especially for the detection of Learning Disabilities, it has been proved that compared to other detection tools, teachers are more effective in predicting pupils' school failure (Panteliadou & Sideridis, 2007). However, teachers' judgements also include the element of subjectivity, as they are often influenced by the expectations they form for each pupil's progress (Katsillis, 2005).

Teachers' judgements often come from a scale or a questionnaire that can be either specialized in a subject or related to the pupils' overall school performance. Numerous studies in various countries studied the possibility of specific tools in assessing both pupils' abilities and the occurrence or not of specific behavioral problems. Another critical application of such scales, is the ability to identify pupils with Learning Disabilities. A review of the international bibliography shows that these scales have been used in research in Norway (Salvesen & Undheim, 1994), Australia (Kenny & Chekaluk, 1993), USA (Gresman et al., 1997), Romania (Negovan et al., 2011), and Greece (Dimakos, 2007, 2008; Triga, 2004). The vast majority of the results of these studies present the teachers as accurate and reliable in their assessments.

The accuracy of teachers' views on school performance has been influenced by several individual factors (Südkamp et al., 2012). Research has shown a higher correlation score between teacher judgements and children's performance than that of the weighted tests, when the evaluation is based on the Curriculum-Based Measurement (CBM) (Eckert et al., 2006; Feinberg & Shapiro, 2003). Despite the number of studies that highlight the ability of teachers to function as evaluation tests, there is still a lack of confidence in their accuracy and reliability. Perhaps the best argument for the application and use of evaluative judgements by teachers comes from Gresham and colleagues who argued that "somewhere in the development of psycho-pedagogical decision, the teacher judgements have downgraded to conjecture, suspicion and, above all, suspicious data" (as cited in Triga-Mertica, 2010).

In the Greek educational reality, the available data concerning the study of the evaluative judgements of teachers about the evaluation of written skills of both pupils of typical development (Balasi, 2019; Dimakos, 2007, 2008) and pupils with Learning Disabilities (Machaira, 2019; Xanthi, 2014) is relatively few. Based on the review of previous studies, the current study explored how teachers' judgements for the school performance of pupils in the last three grades of the elementary school, relate to their performance in written assessment tests with two types of research. The aim was to examine the reliability and accuracy of these judgements through the evaluation of writing. Also, the study aimed to investigate the individual factors that influence the relationship between pupils' performance in written language tests and teachers' evaluative judgements about pupils' school performance (the performance type, learning difficulties, type of assessment, the gender of pupils, and teachers). It is considered that such an investigation would allow for useful conclusions to be drawn, both for the evaluation process, and for the teachers' ability to act as evaluation criteria. The focus of the first research was to examine the teachers' ability to predict pupils' school failure and detect pupils with Learning Disabilities. For this purpose, the teachers' judgements both for the school performance of pupils with Learning Disabilities, language skills, and behavioral problems are correlated, in order to identify convergences and deviations in their assessments. Numerous studies in various countries studied the ability of teachers in assessing both pupils' skills and the occurrence or not of specific behavioral problems. The vast majority of them, present the teachers as accurate and reliable in their assessments, especially in pupils with Learning Disabilities.

2. Literature Review

2.1 The School Performance and Its Assessment

The term "performance" is used in school to express the amount and quality of knowledge, skills, and abilities that a pupil seems to have in a particular course, at a specific period of time (quarter, semester, year). It is multidimensional and is related to all phases of human development: cognitive, emotional, social, and physical. In the context of school life, it is a product of specific behavior (diligent or not) that the student displays in all his age phases (Steinberger, 1993). Performance is not a matter of the pupil exclusively. Undoubtedly, it is influenced by him, but at the same time, it is influenced by other factors. The inherent and exogenous factors that affect the pupil performance are: the interest in the lesson, the pupils' study, the willingness to work, their motivations. Other significant factors are: the time he has, the way the teacher teaches, the quantity of material taught each time, school curricula, material and technical infrastructure in the school, home conditions, school communication with teachers, communication at home with parents (Athanasiou, 2003).

Assessing a pupils' school performance is a systematic process that determines the extent to which teaching goals and education, in general, have been achieved, and takes into account several elements. According to Constantinou (2004), the pupils' performance depends on a) his characteristics (biological, cognitive, psychokinetic), b) his family characteristics (educational and socio-economic level of parents, family relationships, and expectations), c) The social environment (peers, cultural and economic level of the region), d) the characteristics of the school (teacher-pupil relationship, classroom-authoritarian, competitive, or collaborative, teaching, and pedagogical means). In the Greek educational system, the teacher is called upon to take into account individual parameters of the cognitive field, the field of personal characteristics of the pupil, and the social field. Other criteria are also taken into consideration in order to formulate an overall evaluation, and to perform scores using a numerical scale or a score scale of symbols and verbal descriptions. The teachers have also taken into account other characteristics during the assessment of the pupil's performance in the Curriculum lessons, like the effort he makes, his interest, the initiatives he develops, his creativity, the collaboration with his classmates, and respect for the school's operating regulations (Government Gazette 4358/τ'Β΄/2017).

A recent study of data collection on the scores of all pupils in the country during their studies, in the 6th grade (2015-16), results that 10 is the predominant score in basic subjects, such as Language, Mathematics, Physics, ICT, and History (A.QA.P.S.E., 2019). Only for Language and Mathematics, the intermediate value is 9. The percentages of pupils with a grade less than 7, who have an increased chance of being functionally illiterate in the specific courses, were: in Language (7%), in Mathematics (8.7%), in Physics (5.5%), in History (9.4%), and in ICT (0.8%). In a relevant

study in Cyprus, the percentage of primary school pupils, who are likely to remain functionally illiterate, is just over 10% in Language and 7% in Mathematics (Petridou et al., 2009: 78-79). However, according to other researchers, in the pupil population, a percentage of 20-30% has school difficulties expressing difficulty or inability to attend the curriculum (Barbas, 2007).

The educational and scientific community is concerned by the phenomenon of low performance mainly due to its extensions. Low school performance can lead to school failure, which leads to school leakage, but it is also associated with difficulties in social and personal adjustment. Factors associated with low school performance can be either individual, such as intelligence, motivation, and self-perceptions, or social such as cultural and ethnic origin, gender, and family with socio-economic level, and the effects on the child's development. Some further difficulties arise in the evaluation of school performance to detect specific learning difficulties. The differentiation between groups of children with Learning Disabilities and low performance has always been problematic, and research findings on the differentiation of school skills between groups are contradictory (Gresham et al., 1996; Tur Kaspa & Bryan, 1995).

Empirical research has examined the extent to which grades reflect pupils' performance. In these studies, pupils' grades correlated with the results of standardized knowledge tests. The correlation between the grade and test scores in these studies ranges between 0.4 and 0.6 (Duckworth & Seligman, 2006; Pattison, Grodsky, & Muller, 2013; Woodruff & Ziomek, 2004). It is clear that there is a strong, but not absolute, overlap between the grade and pupils' performance. The relationship between grades, achievement, and gender has been repeatedly demonstrated – girls get better grades on average, although they do not outperform boys on achievement or IQ tests (Duckworth & Seligman, 2006).

2.2 The Accuracy of Teachers' Judgments About School Performance

The accuracy of the teachers' judgement is based on the correspondence between the teachers' assessments of the pupils' academic performance and the actual academic performance of the pupils measured by a standardized test. The correlation between them usually used a measure of comparison of this correspondence. However, other indicators can be used, such as the average difference between teachers' judgments and pupils' actual performance. According to the model of Südkamp et al. (2012), many individual factors influence the accuracy of the teachers' judgement. On the one hand, the pupils' performance on a test may depend on them, such as prior knowledge, motivation, and intelligence. But it can also depend on the characteristics of the test. That is the cognitive area of a particular test or the projects it includes, as well as the difficulty of these projects.

On the other hand, a teacher's judgement may depend on the teacher's special traits, such as professional expertise or stereotypes about pupils or the characteristics of the crisis. For example, if he is called upon to judge a particular pupil's ability, like reading, or to judge the general academic ability. Also, the correspondence between the characteristics of the crisis and the tests influence the accuracy of the teachers' judgment. The test may measure a specific academic ability, as numerical skills. But the teachers' crisis can be broader and give pupils an overall ability in mathematics that makes it more difficult for teachers to judge more accurately. Another relationship that may affect the accuracy of the teachers' judgement is the correspondence between the characteristics of the teachers and pupils, for example, gender, nationality (Südkamp et al., 2012). From the 1980s and more recently, researchers explored the level of achievement or ability of pupils as predictions for the accuracy of teachers' judgement. Most research found that teachers were more accurate in judging high performing pupils than low achieving ones (Begeny et al., 2008; Coladarci, 1986; Feinberg & Shapiro, 2009). Martin and Shapiro (2011) found that teachers increased accuracy in their ratings for low-achieving pupils compared to average-achieving pupils, but did not draw any conclusions for high-achieving pupils.

The correctness of the assessment of learning achievements systematically incorporated into two meta-analyses. Hoge and Coladarci (1989) found an average judgement accuracy in 16 studies r=0.66, while a recent meta-analysis of Südkamp et al. (2012) in 75 studies indicated an average accuracy r=0.63, but with a range correlation in individual studies (-0.03 to 0.92). These results are considered quite high but suggest that the teachers' judgements explain about 40% of the variation in learning achievement. Higher correlations were found between teacher judgements and performance measurements when teachers are informed of the extent to which was compared (Südkamp et al., 2012). Also, the correlations were higher when decisions and measures concerned the same field of knowledge or the same aspect of ability. The meta-analysis of Machts et al. (2016), which was based on 33 studies, reported a lower correlation (r=0.43) between the performance measured and the teachers' judgements about pupils' cognitive abilities. However, there is a significant variation in the size of the results in all studies.

Regarding the accuracy of the teachers' judgement about the evaluation of the writing from a recent study reported marginally higher correlations than 0.70, with a wide range in individual schools (-0.07 to 0.94) (Meissel et al., 2017). In Greece, both studies of Dimakos (2007, 2008) showed a teacher's small and medium ability for the reliable and valid diagnostic evaluation of their pupils (0.24 to 0.45). The high correlations refer to teachers' judgements about spelling and written expression. From the Balasi's study (2019), significant relationships statistically, are seen for all words, all

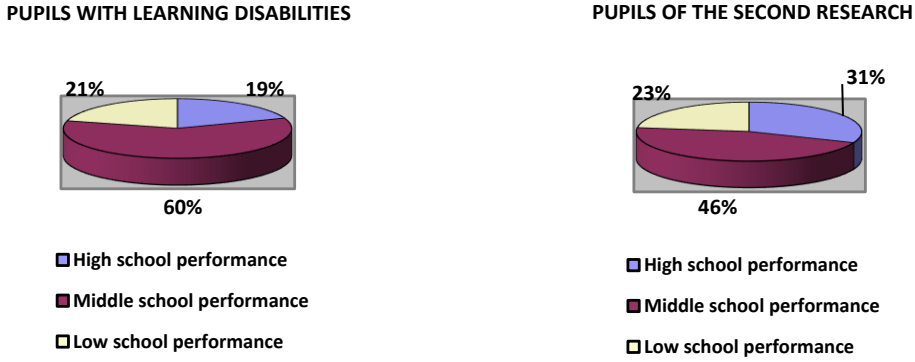
spelled words and spelling correctness with the evaluation of the written expression, of spelling, and the learning of language. A study of Xanthi (2014) showed moderate and low statistically significant relationships between pupils' performance and teachers' evaluative judgements on language skills (teachers of class 0.28 to 0.54 - teachers of integration departments 0.28 to 0.53). In the research of Machaira (2019) a small correlation was observed (0.18 to 0.28).

3. Method

3.1 Participants

The first study involved 27 class teachers, women (N=18), and men (N=9). A total of 89 primary school pupils were evaluated individually. Of these, 67 pupils, 27 girls and 40 boys, D' (N=20), E' (N=33), and F' (N=14) were pupils with Learning Disabilities. These pupils received supportive teaching intervention in integration departments, and they were diagnosed with Learning Disabilities by KESY or another medical and pedagogical service. The remaining 22, girls (N=14) and boys (N=8), who were in grades D' (N=5), E' (N=9), and F' (N=8) were pupils with high school performance. Based on the high response of children to the objectives of the curriculum in all subjects, a pupil with the best overall school performance was selected from each class where pupils with Learning Disabilities attended.

In the second study, 109 primary school pupils from two co-located public schools were evaluated as a group. The specific pupils, 60 girls and 49 boys, 4th grade (N=48), 5th grade (N=28), and 6th grade (N=33), divided into three groups with criterion the response to the objectives of the curriculum in all courses from their teachers. Based on pupils' overall grading, four women and two class teachers indicted 34 pupils with high school performance (9 to 10), 50 pupils with middle school performance (7 to 8), and 25 pupils with low school performance with a general grade of less than 7.



Graph 1. Pupils' percentage in each category based on the overall score

In the second study, the higher percentage was pupils with middle performance (Graph 1). The pupils with high performance were fewer, while the pupils with low performance were much fewer. As long as the pupils with Learning Disabilities is concerned, the higher percentage was pupils with middle performance. The high-performance percentage and the low-performance percentage were small. The first language of all pupils, in both studies, was Greek.

3.2 Instruments

The research tool chosen for this study was a test set that combines tests of psychometric criteria of language proficiency with language tests and free text-writing. The ecological validity of the tool was ensured since most informal tests are activities in the school books of the last three grades of the Primary School (M.N.E.R.A-P.I., 2007). Another feature that distinguishes the tool is the representative sampling of the dimensions of production and expression of the written word for evaluation. Different types of tests were included that covered a wide range of language skills and they were representative in terms of the peculiarities of the Greek language system, but balanced as well in terms of their representation. It was also distinguished by adequate coverage for each level of written language development. It included tests that adequately met different developmental stages of written language. It had a sufficient number of informal tests of escalating difficulty that can render the picture of pupils' language skills at any level adequately. Besides, it provided the teacher with the opportunity to assess emerging language skills in pupils with low performance and pupils with Learning Disabilities, including tests accessible to these groups. As a model, both theoretically and functionally, the Diagnostic Investigation Tool for Difficulties in the Writing of Pupils in C-F grade of Primary School were used (Porpodas et al., 2007) and L-a-T-o (Tzouriadou et al., 2008).

In the test1 the pupils' internalization and the application of the punctuation rules was evaluated to determine the sentences' limits and their types. Pupils had to check both the meaning of each sentence and the dialogue meaning and

put (18) eighteen punctuation marks.

The knowledge of both word syntax and structure of simple and compound sentences (active or passive) was evaluated by two tests. In the test2, pupils had to make two sentences with specific words. One has active syntax and the other passive. In the test3, pupils put together two pairs of sentences, main/subordinate, and they made compound sentences. There were chances for syntax errors in the subject, in the verb, in the object of simple sentences or determination. Also, there were chances for mistakes in the link of compound sentences.

The recognition of various morphological types of filling the gaps in the text was evaluated. The pupils had to find synonyms and opposite words. Also, they had to put the words in their correct declension. Furthermore, they had to use the production and synthesis relationships that exist between them. In the test4, pupils had to change the words in different grammatical types or put them in their correct declension contextually. In the test5, pupils had to fill a similar, opposite, or compound word with that of the parentheses. They had to understand the meaning of the text as well. There were the possibilities of inconsistent use of the time, the person, and the number verb. Also, there were the probabilities by inconsistent use in gender, the number, and the case of nouns, adjectives, and pronouns. Furthermore, pupils could make errors in the synonyms or the opposites.

The test6 tested the language experience, and the understanding of special meanings, like the knowledge of idioms, proverbs, and comparisons. The pupils were asked to complete the second part of two proverbs, of two simulations, and two metaphors. There were chances for mistakes in the type of metaphor or the meaning of the phrase, literal or metaphorical.

The knowledge of the word meaning and the ability to interpret multiple meanings of them were tested in test7. The pupils had to find two related words of the words: walking, healthy, space, height. It could have been a synonym, a characteristic, and a general or a proper metaphorical meaning, but also a relative grammar type. There was the possibility of using irrelevant words, either word with a constitutional relation or an opposite relation, and a different declension.

The understanding of word meaning with simple formal definitions was tested in test8. The pupils had to define the meaning of four words: hat, umbrella, horse, the thief (Georgas et al., 1997). There was a possibility that the pupils may fail partially using a vague or less relevant synonym, a secondary meaning without any further clarification, or even the successful but specific use of the word without any further clarification. But there was also the possibility of failing by saying a wrong answer. Furthermore, there was the possibility of failing by saying a common phrase that contained the word but did not show a real understanding of its meaning even after a clarifying question.

The writing ability, as well as the proficiency and fluency in writing, were evaluated by two tests. In the test9, the pupils had to write a narrative text based on unfinished phrases that could be organized freely but also appropriately. In the test10, the pupils were asked to write a narrative based on six images.

The ability to organize the sequence of events and the writing ability was tested in test11. The pupils had to identify the right pair of sentences, and the link which they were linked. Then, they had to put them in time order and to make a paragraph adding their words or their phrases.

The test12 tested the ability to correct the text content. The pupils were invited to elaborate on a text, which was written in telegraphic form, adding their details, describing better the news, and changing the noun phrases on the verb phrases. There was a possibility that the pupils focus their efforts on the superficial parts of the text without changing the noun phrases into verbal ones and without interfering with the meaning of the text.

3.3 Data Collection

The first study was carried out in 13 public schools in neighboring areas in Athens. The second study was carried out in 2 co-located public schools in Athens. A series of tests given to the pupils and a variety of written and word processing skills was measured. The testing process took place during school hours, with the consent of both the school principal and the class teachers and parents. In the first study, the pupils were tested individually, in a quiet place, outside their classroom. In the second study, the pupils were tested in groups, in a classroom, and only the completed tests were collected. In both studies, the teachers based on the pupils' responses to the objectives of the curriculum in all subjects, assessed and rated them.

For pupils with Learning Disabilities, in the first study, the teachers were asked to complete a Likert 5-point questionnaire (where 1=very low and 5=very high), in order to evaluate the overall performance in language development/oral expression, spelling, in grammar, vocabulary and written expression based on expected age and school performance in the classroom. They also asked to evaluate the occurrence or not of social problems/adjustment problems in school, attention and concentration problems, hyperactivity/impulsivity, emotional problems/anxiety, learning problems, behavioral/aggression problems with a 5-point scale (where 1=Never and 5=Very Often) based on

the behavior of pupils in the classroom (Protopapas et al., 2005).

Tests were provided based on a strict protocol with written instructions to the examiners recording the responses to individual assessment brochures. Administration performed in three phases, and it did on different days of the week. The first phase included the punctuation test, the syntax test, and the vocabulary test (40 minutes). The second phase involved the organizing test and the writing test with the help of phrases (40 minutes duration). The third phase included the content improvement test and the writing test with the help of pictures (40 minutes duration). Times respond to most pupils generally, but they have been adapted to the rhythms of each child so as not to affect their performance.

By combining the general impression, and the analytical correction method, in the writing tests and the improvement content tests evaluated individual parameters such as the content completeness, the text relevance, its effectiveness, the organizing, the vocabulary, the style, the syntax, the spelling-morphology, and the text clarity-readability. The total grade of each pupil, the maximum overall performance of 100, arose from the sum of the scores in the product control keys by the informal educational assessment: paragraph, content and productivity, spelling.

To ensure the credibility of the texts' grading assessement, the texts were coded and each one was corrected by two evaluators. One was the researcher, who is the author of the article, and the other was a teacher, who did not know the participants. At first, the agreement between the two evaluators was not high (59%), but after resolving certain disputes, with discussion, it reached (93%). The disagreements were about the divergence between the two gradings, mainly on tests of writing, organizing, and content improvement. To resolve the disagreement, the two evaluators reviewed the texts, and they agreed on a joint new grade.

The number of errors were counted in the tests with the one way answers. The scores of the pupils in the tests resulted from the number of correct answers. In the tests with diverse answers, the pupils' responses were appreciated, and they got specific units, depending on whether the answers were correct or partially correct based on the criteria mentioned above.

4. Findings

Table 1 presents the measures of central tendency and the dispersion measures for the variable of the pupils' grades in the two studies. It shows the level of all pupils but also each group separately. The data show that 8 is the predominant grade in both the school performance of all pupils in both studies and the school performance of pupils with Learning Disabilities and also in pupils with middle-performance. In high-performance pupils, the predominant grade of school performance is 10 in both studies. The highest grade is found in high-performance pupils, with a small predominance in pupils of the second study, and the lowest in low-performance pupils. The score in the middle is 8 for all pupils in both studies. Also, it is the same and for pupils with Learning Disabilities and pupils with middle-performance. For pupils with high-performance in both studies, it is 10, while for pupils with low-performance is 7. The highest variation of the score is observed in all pupils of the second study. On average, scores are 1.231 points away from the mean value (standard deviation). The smallest appears in pupils with middle-performance, with a standard deviation of 0.240. The range of scores is higher in all pupils in the second study.

Table .1 Measures of central tendency and dispersion measures in the two studies

Pupils' Croup	Number of Pupils	Mean	Median	Mode	Variance	Standard Deviation
First research						
All pupils	89	8.43	8	8	0.973	0.986
Pupils with high school performance	22	9.82	10	10	0.156	0.395
Pupils with Learning Disabilities	67	7.98	8	8	0.397	0.630
Second research						
All pupils	109	8.28	8	8	1.516	1.231
Pupils with high school performance	34	9.91	10	10	0.083	0.288
Pupils with middle school performance	50	7.94	8	8	0.058	0.240
Pupils with low school performance	25	6.76	7	7	0.273	0.523

In both studies, high-performance pupils performed much better in all tests with a statistically significant difference (Table 2). ANOVA analysis showed a statistically significant effect on the type of performance, $F(1,88)=19.267$, $p<0.001$, $n^2=0.75$ in the first study, and $F(1,108)=3.594$, $p<0.001$, $n^2=0.31$ in second. In the first research, except for the test concerning the synonyms, pronouns, and composition relations where the significance level was $p=0.001$. In the second study, a high effect was observed on the tests related to punctuation, the correct grammatical type of word, the production of text with phrases, the organizing, and text improvement ($p<0.001$). The effect was also significant in experiments involving sentence formation, concept definition, synonyms, pronouns, word production, and composition, and image text production ($p<0.05$). The effect was not significant on the two-sentence test and the metaphorical speech.

Table 2. The averages' performance of groups in each test

Tests / Groups	High school performance - first research (N=22)	High school performance - second research (N=34)	Middle school performance (N=50)	Low school performance (N=25)	Learning disabilities (N=67)
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
Test1	14.49 (1.740)	11.64 (2.840)	10.00 (4.128)	7.31 (4.298)	5.43 (3.536)
Test2	4.00 (0.000)	3.50 (0.492)	3.12 (0.817)	3.06 (0.905)	2.34 (1.317)
Test3	4.00 (0.000)	2.55 (1.341)	2.01 (1.426)	1.70 (1.479)	2.17 (1.295)
Test4	7.86 (0.467)	7.05 (1.594)	6.94 (1.300)	5.24 (2.420)	5.40 (2.181)
Test5	8.00 (0.000)	7.38 (0.652)	7.00 (0.968)	6.48 (1.357)	7.19 (1.104)
Test6	7.18 (1.006)	5.61 (1.938)	4.74 (2.211)	3.36 (2.325)	5.11 (2.002)
Test7	5.09 (0.629)	2.38 (1.348)	2.14 (1.702)	1.32 (1.651)	3.09 (1.495)
Test8	5.09 (0.811)	3.05 (1.475)	3.50 (1.581)	3.04 (1.968)	3.43 (1.221)
Test9	92.09 (17.171)	73.35 (27.534)	55.05 (23.581)	34.59 (25.739)	55.50 (16.433)
Test10	83.19 (19.220)	67.41 (27.013)	53.52 (23.310)	42.39 (22.455)	47.07 (20.462)
Test11	11.52 (1.085)	9.48 (2.843)	8.33 (2.975)	5.24 (3.657)	7.29 (3.708)
Test12	95.50 (12.340)	80.85 (25.002)	63.81 (27.877)	34.04 (23.388)	50.14 (22.049)

Testing for a correlation between the pupils' school performance in the first study and their performance in tests showed significance at a high level, 0.58 to 0.64, for punctuation, text production, and improvement (Table 3). The relations are significant at a moderate level, 0.40 to 0.48, for grammatical types, metaphorical speech, and syntax. The correlation is low, 0.32 to 0.35, in the organizing, and definition of words. Of individual controls, only one indicated for high-performance pupils. There are negative correlations in both groups. Furthermore, no correlation was observed in some tests concerning high performance pupils.

Table 3. Relevance indicators (Pearson's r) for all tests and grades of pupils in the first research

Tests / Groups	1st Research	
	Teachers (N=27)	
	All pupils - a general grade of 7 to 10 (N=89)	High school performance - a general grade of 9 to 10 (N=22)
Test1	0.638 (**)	-
Test2	0.466(**)	-
Test3	0.456(**)	-
Test4	0.415(**)	-0.141-
Test5	0.321(**)	-
Test6	0.397 (**)	-0.033
Test7	0.481(**)	0.453(*)
Test8	0.469 (**)	0.054
Test9	0.584(**)	0.151
Test10	0.586(**)	0.311
Test11	0.350(**)	-0.045
Test12	0.602(**)	0.140

(**) p <0.01, (*) p <0.05

In the second study, there is a statistically significant correlation between average level, 0.43 to 0.50, in the text production with phrases, in the text improvement, and organizing. In contrast, it decreases significantly for the remaining tests and is low, 0.22 to 0.37. It is noteworthy that not found for the metaphorical speech (0.02). Of individual controls for the three groups, only two significant correlations showed pupils with low-performance. There were negative correlations found in pupils with high and middle performance.

Table 4. Relevance indicators (Pearson's r) for all tests and grades of pupils in the second research

Tests / Groups	2nd Research			
	Teachers (N=6)			
	All pupils - a general grade of 6 to 10 (N=109)	High school performance - a general grade of 9 to 10 (N=34)	Midle school performance - a general grade of 7 to 8 (N=50)	Low school performance - a general grade of less than 7 (N=25)
Test1	0.352(**)	-0.151	-0.121	0.221
Test2	0.255(**)	-0.321	0.175	0.208
Test3	0.220(*)	-0.182	-	0.226
Test4	0.371(**)	-0.186	0.055	0.673(**)
Test5	0.349(**)	0.024	-	0.404(*)
Test6	0.358(**)	0.046	0.065	0.245
Test7	0.270(**)	-0.145	0.210	0.382
Test8	0.021	0.013	0.136	0.334
Test9	0.472(**)	-0.244	-0.033	0.367
Test10	0.362(**)	-0.120	0.045	0.322
Test11	0.433(**)	-0.131	0.120	0.347
Test12	0.500(**)	-0.154	0.028	0.332

(**) p <0.01, (*) p <0.05

Testing for a possible correlation between teachers' judgments for the school performance of pupils with Learning

Disabilities and their language skills showed a statistically significant correlation rating from 0.27 to 0.42 (Table 5). The highest correlation is for spelling and the lowest for grammar. Also, Pearson's r indicators turn out that there is no significance for any kind of problem. The correlation is positive only for adaptation to school. Besides, there is negative for attention, concentration, and behavioral problems, hyperactivity/impulsivity, and aggression.

Table 5. Relevance indicators (Pearson's r) of the general score of pupils with Learning Disabilities and teachers' assessments of pupils' abilities, difficulties, and problems

Pupils with Learning Disabilities (N=58)	
Type of possibility or difficulty	**General grade** (7 to 9)
Linguistic development/Oral expression	0.378(**)
Spelling	0.424(**)
Grammar	0.270(*)
Vocabulary	0.354(**)
Written expression	0.338(**)
Type of problem	
Social problems/adaptation to school	0.084
Attention and concentration problems	-0.174
Hyperactivity/Impulsivity	-0.088
Emotional problems/anxiety	0.176
Learning problems	0.304(*)
Execution Problems/Aggression	-0.173

(**) $p < 0.01$, (*) $p < 0.05$

Examination of the internal questionnaire consistency data with alpha Cronbach reliability factor showed that for each evaluation criterion as well as for the item-total, the rates are high for pupils' language skills (0.85 to 0.89), and satisfactory for the type of problem (0.52 to 0.67). From the control of correlations between pupils' performance in the tests and the teachers' judgments about their language skills, low and moderate correlations were observed (0.27 to 0.44) for the synonyms, pronouns, production and synthesis of words, production, text' improvement, correct use of grammatical types, syntax, and metaphorical speech. In contrast, the correlations between language skills were high (0.61 to 0.76) and moderate (0.44 to 0.56). The highest related to language development/oral expression with vocabulary, and the lowest to the same ability with spelling.

Testing using the Mann-Whitney U test showed statistically marginal significant results in gender only in the first study ($z=757.000$, $p=0.048$, $\eta^2=0.05$). However, it noted that the effect size found to be very small (Cohen, 1988). The median score was common to both sexes for the two studies (8). In the first study, the girls are more in grade 10. In the second study, the boys are more in the intervals of scores 10 and 8. On a scale of 6, the number is the same for both sexes, and there are very few pupils. Boys also range in grades 5. Mann-Whitney U non-parametric testing showed that there are statistically significant differences between men's and women's scores only in the second study ($z=5.977$, $p<0.001$). The magnitude of the effect is relatively small ($n^2=0.25$).

5. Discussion

This study investigated the correlations between teachers' judgments about their pupils' school performance and pupils' performance on informal writing tests based on pupils' curriculum. The results show that the correlations are significant for more tests, at a high level (average 0.48), in the first study. In contrast, in the second study, the correlations are low mainly, and medium and their average are 0.33. Lower mean crisis accuracy in both studies is found both in the meta-analysis of Hoge and Coladarci (1989), and in the meta-analysis of Südkamp et al. (2012), and the study of Meissel et al. (2017). However, the first study results are higher than Machts et al. (2016). Also, it is encouraging that the specific findings are consistent with similar studies in Greece (Balasi, 2019; Dimakos, 2007, 2008). In agreement with other empirical research, the correlations between the grade and test scores oscillate between 0.4 and 0.6 (Duckworth & Seligman, 2006; Pattison, Grodsky & Muller, 2013; Woodruff & Ziomek, 2004). It seems, therefore, that teachers evaluate the school performance of their pupils significantly based on their performance in writing tests. However, their judgments about school performance are broader and assess pupils' overall ability in a variety of

subjects.

It is an interest that there is a difference between the two studies in the correlation between school performance and pupils' performance in the organizing, which related to reasoning skills. Even it is at the medium size in the two studies, it ranges from the low level in the first study (0.35) to the higher in the second (0.43). Also, interest is the difference between the two studies in the correlation between school performance and performance in metaphorical speech. In the first study, it ranges at a high level (0.47), while, in the second, it is at a low level (0.02). Such differences were observed among other tests too, highlighting differences in the pupils' characteristics, but also the tests' characteristics and assessment, individual administration versus group. According to the model of Südkamp et al. (2012), such characteristics influence the teacher's accuracy.

5.1 The Pupils' General Scores and the Ability of Tests

The score distribution analysis of school performance shows a predominant score of 8 in all pupils at studies. This finding seems to be different from a recent data collection study on the grades of all pupils in the country, which shows a predominant score of 10 in basic courses of the curriculum in pupils 6th grade (A.QA.P.S.E., 2019).

Also, compared to data in Greece and Cyprus, in two types of research of the study the percentage of pupils with a performance of less than 7 was much higher, 16% in the first study and 23% in the second (A.QA.P.S.E., 2019; Petridou et al., 2009).

This differentiation of findings is due to factors related to the level of the classroom or school. Also, it is due to the abilities and age of the pupils. The first study attended by the majority of pupils with a diagnosis of Learning Disabilities, while, in both studies, pupils attended three last grades of primary school.

Regarding the ability of tests to reliably measure pupils' language skills, it appears that they meet specific criteria. On the one hand, they satisfied the need for pupils evaluated within the school, and on the other hand, they were based on the pupils' curriculum. Also, weighted tools used as models, both theoretical and functional (Porpodas et al., 2007; Tzouriadou et al., 2008), which combined psychometric language proficiency projects with informal language performance tests and free text writing projects.

5.2 Other Factors That Influence These Findings

In the first study, higher correlations were observed for more trials (testings). On the one hand, the assessment process, individual to the group, seems to have significantly affected pupils' performance. Research has shown that the assessment individually helps pupils focus more on their skills and perform better (Xanthi, 2019). On the other hand, the pupils' characteristics that deal with both their ability and level of achievement, as well as the group they belong to, pupils with low performance and Learning Disabilities, seem to influence the correlations between the two measurements. In agreement with Martin and Shapiro (2011), the teachers' accuracy increased in their judgements for low-achieving pupils than for average-achieving pupils. The gender of pupils and teachers does not seem to play a decisive role.

In agreement with Barba (2007), this study highlights a significant percentage of pupils who have difficulty or inability to attend the curriculum and perform poorly. The group of pupils with Learning Disabilities differs in terms of school performance, and only a small percentage show low performance. Both the school performance of pupils with Learning Disabilities, as a whole, and their performance in the tests show that they perform better than pupils with low performance. This finding contrasts with findings from other studies that show that children with Learning Disabilities lag behind low-performing children (Gresham et al., 1996; Tur Kaspa; 1995) and expands the discussion on the contradictions of the findings with the school skills of both groups of pupils.

The correspondence between the characteristics of crisis and the tests' characteristics shows that the pupils' school performance is related to writing elements and pupils' performance in specific tests. In both studies, the highest correlations were for text production and improvement. In these tests, the evaluation of pupil performance dealt with both general evaluation criteria such as the completeness of the content, the appropriateness of the text and its effectiveness, and individual parameters such as content, organizing, vocabulary, style, syntax, and spelling-morphology. In agreement with other research data, in which the correlations between narrative text production tests and teachers' judgements on specific language skills were investigated, and in the present study appear similar statistically significant correlation ratios (Dimakos, 2007, 2008).

Finally, it was found that teachers largely support their judgments about the overall performance of pupils with Learning Disabilities based on specific language skills, such as spelling and grammar. Also, they take into account parameters such as social problems and adaptation to school. In contrast to Machaira's (2019) research and agreement with other research data (Panteliadou & Antoniou, 2008; Triga, 2004), these findings are encouraging teachers' ability to assess their students' learning difficulties through school performance. The evaluation criteria of the teachers of the present

study were not limited to the individual examination of the pupils' performance in specific areas, but the components of their cognitive ability evaluated globally.

6. Conclusion

These findings added to a long list of research data that highlighted the adequacy of teachers' judgments. Teachers' assessments are valid and reliable because they are based on pupils' systematic observations daily. Despite any limitations, the data presented is encouraging. Without being rejected, not accepted either, teachers' judgements could be an additional criterion in pupils' assessing, especially those with Learning Disabilities and low performance. The two types of research in the study revealed some evidence about the factors that influence the evaluation of pupils' school performance through written speech performance. To draw safer and more conclusions, and to accurately identify these factors that affect the accuracy of teachers' judgements, other studies on the subject are suggested.

References

Athanasiou, L. (2003). *Assessment of student's performance in school and of the teaching work***.** Ioannina: self-publication.

Authority for Quality Assurance in Primary and Secondary Education (2019). Annual Report of A.Q.A.P.S.E. (Ed. H. Matsagouras). Retrieved 10 March 2020 from http://users.sch.gr//gbotsas/pdfs/grss/Amdeperigrafi.pdf

Balasi, K. (2019). *The use of CBM as a means of evaluation of written texts of students of primary school*. Postgraduate Diploma Thesis, University of Patras, Greece. Retrieved 20 March 2020 from https://nemertes.lis.upatras.gr/jspui/handle/10889/13272

Barbas, G. (2007). *School and learning (a deviant relationship)*. Thessaloniki: Prometheus.

Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgments and students' performance across a continuum of rating methods. *School Psychology Quarterly, 23*, 43. https://doi.org/10.1037/1045-3830.23.1.43

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coladarci, T. (1986). Accuracy of teacher judgments of students' responses to standardized test items. *Journal of Educational Psychology, 78*, 141-146. https://doi.org/10.1037/0022-0663.78.2.141

Constantinou, Ch. (2004). *Assessing student performance as pedagogical logic and school practice*. Athens: Gutenberg.

Dimakos, I. (2007). Alternative assessment of elementary school students' written skills. In: M. Vlasopoulou, A. Giannettopoulou, M. Diamanti, L. Kirpotin, E. Levadi, K. Lefteri, & G. Sakellariou (Eds.), *Language Difficulties and Writing in School Learning Framework* (pp.154-163). Athens: Grigoris.

Dimakos, I. (2008). Reliability and validity in the evaluation of written expression by teachers. *Psychology, 15*(3), 225-238.

Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology, 98*(1), 198-208. https://doi.org/10.1037/0022-0663.98.1.198

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools, 43*, 247-265. https://doi.org/10.1002/pits.20147

Feinberg, A. B., & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly, 18*, 52-65. https://doi.org/10.1521/scpq.18.1.52.20876

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *Journal of Educational Research, 102,* 453-462. https://doi.org/10.3200/JOER.102.6.453-462

Georgas, D. D., Paraskevopoulos, I. N., Bezevegis, H. G., & Giannitsas, N. D. (1997). Greek WISC-III: Wechsler Intelligence Scales for Children. Athens: Greek Letters.

Gresham, F. M., MacMillan, D. L., & Bocian, K. M. (1996). Learning disabilities, low achievement, and mild mental retardation: more alike than different? *Journal of Learning Disabilities, 29,* 570-581. https://doi.org/10.1177/002221949602900601

Gresman, F. M., MacMillan D. L., & Bocian, K. M. (1997). Teachers as "tests": Differential validity of teacher judgments in identifying students at-risk for learning disabilities. *School Psychology Review*, *26*, 47-60.

Hagley, F. (2002). Suffolk Reading Scale 2. Slough: NFER-Nelson.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*(3), 297-313. https://doi.org/10.3102/00346543059003297

Katsillis, M. I. (2005). Self-fulfilling prophecy, cultural capital and unequal opportunities in education. *Scientific Yearbook of Aretha, 3*, 131-138.

Kenny, D. T., & Chekaluk, E. (1993). Early reading performance: A comparison of teacher-based and test–based assessments. *Journal of Learning Disabilities, 26*, 227-236. https://doi.org/10.1177/002221949302600403

Machaira, M. (2019). *Teachers judgments' accuracy in the assessment of LD primary students' achievement*. Bachelor's thesis, Aristotle University of Thessaloniki, Greece. Retrieved 15 March 2020 from https://ikee.lib.auth.gr/record/302850/?ln=el

Machts, N., Kaiser, J., Schmidt, F. T. C., & Möller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: a meta-analysis. *Educational Research Review, 19*, 85-103. https://doi.org/10.1016/j.edurev.2016.06.003

Martin, S. D., & Shapiro, E. S. (2011). Examining the accuracy of teachers' judgments of DIBELS performance. *Psychology in the Schools, 48*, 343-356. https://doi.org/10.1002/pits.20558

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: exploring student characteristics that influence teacher judgments of student ability. *Teaching* and *Teacher Education, 65*, 48-60. https://doi.org/10.1016/j.tate.2017.02.021

Ministerial Decision Φ.7Α/ΦΜ/212191/Δ1/04-12-2017 (Government Gazette 4358/τ'Β΄/4.12.2017). Assessment of primary school students.

Ministry of National Education and Religious Affairs - Pedagogical Institute (2007). New School Textbooks - Language D, E & F Elementary. Athens: Pedagogical Institute - Organization for the Publishing of Textbooks. Retrieved 20 September 2017 from www.pi-schools.gr

Negovan, V., Bagana, E., & Dinca, S. (2011). Gender, age and academic standards of school differences in adolescents' self-discrepancy and self acceptance. *Procedia - Social and Behavioral Sciences*, *12*, 40-48. https://doi.org/10.1016/j.sbspro.2011.02.008

Panteliadou, S., & Antoniou, F. (2008). *Teaching approaches and practices for students with Learning disabilities.* Thessaloniki: Grafima.

Panteliadou, S., & Sideridis, G. (2007). *Detection of Learning Disabilities by Teachers*. Ministry of National Education and Religions EPEAEK. Retrieved 15 October 2013 from http://users.sch.gr//gbotsas/pdfs/grss/Amde_perigrafi.pdf

Pattison, E., Grodsky, E., & Muller, C. (2013). Is the sky falling? Grade inflation and the signaling power of grades. *Educational Researcher, 42*(5), 259–265. https://doi.org/10.3102/0013189X13481382

Petridou, A., Tsouris, Ch., Michailidou, A., & Kyriakidis, L. (2009, 6-7 November). The dimensions of functional illiteracy. In: *proceedings of the Conference Educational Research and Teacher Training*. Pedagogical Institute, Leukosia, Cyprus.

Porpodas, K., Diakogiorgi, K., Dimakos, I., Karantzi, I., Palaiothodoros, A., Yfanti, K., Tsaggari, G., & Karabetsou, M. (2007). Diagnostic Tool for Difficulties in the Writing of Pupils in C-F grade of Primary School. Ministry of Education - EPEAEK (2).

Protopapas, A., Mouzaki, A., Simos, P., Nikologianni, A., Spanou, E., & Xanthi, S. (2005, 13-17 July). Development of text comprehension in elementary grades: component processes. ISPA Colloquium, Athens, Greece.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal*, *48*(2), 335-360. https://doi.org/10.3102/0002831210374874

Salvesen, K., & Undheim, J. O. (1994). Screening for learning disabilities with teacher rating scales. *Journal of Learning Disabilities*, *27*, 60-66. https://doi.org/10.1177/002221949402700109

Steinberger, E. D. (1993). *Improving student achievement*. Virginia: American Association of School Administrators.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743-762. https://doi.org/10.1037/a0027627

Südkamp, A., Kaiser, J., & Möller, J. (2014). Teachers' judgments of students' academic achievement. In S.

Krolak-Schwerdt, S. Glock, & M. Böhmer (Eds.), *Teachers' professional development: Assessment, training, and learning* (pp. 5-25). Rotterdam, NL: Sense Publishers.

Triga, A. (2004). An analysis of teacher's rating scales as sources of evidence for a standardized Greek reading test. *Journal of Research in Reading, 27*, 311-320. https://doi.org/10.1111/j.1467-9817.2004.00234.x

Triga-Mertika, E. (2010). *Learning Disabilities - General and Special Learning Disabilities-Dyslexia*. Athens: Grigoris.

Tur-Kaspa, H., & Bryan, T. (1995). Teachers' ratings of the social competence and school adjustment of students with LD in elementary and junior-high-school. *Journal of Learning Disabilities, 28*, 44-52. https://doi.org/10.1177/002221949502800107

Tzouriadou, M., Sygolitou, E., Anagnostopoulou, E., & Bacola, I. (2008). Psychometric Criterion of Language Proficiency L-a-T-o. Thessaloniki.

Woodruff, D. J., & Ziomek, R. L. (2004). High school grade inflation from 1991 to 2003 (Research Report Series 2004-04). Iowa City, IA: ACT.

Xanthi, S. (2014). Evaluation of school performance and behavior problems in children with Learning Disabilities from teachers. *New Paidagogos, 2*, 83-92.

Xanthi, S. (2019). The Individual vs Group Assessment of the Writing: A Study in Pupils with High School Performance D-F Primary. *Education Journal, 8*(5), 232-238. https://doi.org/10.11648/j.edu.20190805.18